



Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
 ISSN:2521-3504(online) ISSN:2074-0204(print)



Detecting Ideological Framing in Geopolitical News: Fine-Tuning RoBERTa-Large with Algorithmic Resampling for Binary Bias Classification

Hanaa Ali Alshaibani^{1*}, Ibtihal A. Mustafa²

¹University of Kufa, Information Technology Research and Development Centre Kufa, Najaf, Iraq. hanaaa.alshabbani@uokufa.edu.iq:

²University of Kufa, Najaf, Iraq. ibtihala.alkhayyat@uokufa.edu.iq:

ARTICLE INFO

Article history:

Received: 22 /06/2026

Revised form: 28 /08/2026

Accepted : 02 /09/2026

Available online: 30 /09/2026

Keywords:

Ideological Framing Detection, RoBERTa-Large, Algorithmic Resampling, Class Imbalance, AG News, Computational Linguistics, Media Bias, Transformer Models.

ABSTRACT

Automated identification of political frame effects in biased news texts continues to be a difficult NLP problem, motivated by issues of label distribution imbalance and contextual richness. Despite being successful at traditional text classification tasks, identification of complex geopolitical bias effects becomes hindered by imbalanced datasets with an overrepresentation of the majority class. In this paper, we propose a binary classification pipeline for identification of geopolitical framing biases within a collection of 30,000 short news articles in the AG News dataset "World" sub-category. In order to overcome imbalances in the minority class, we used stratified partitioning and random sampling techniques for achieving class balance in a training set with 43,698 examples. We fine-tuned RoBERTa-Large (355M parameters) model for three epochs with sequence classification task using FP16 precision strategy. The resulting classifier reaches 91.08% accuracy (95% CI: [90.46%, 91.66%], Wilson score method), macro-F1 equal to 0.91 and AUC score of 0.9641 (95% CI: [0.9601, 0.9681], Hanley-McNeil approximation). Pair-wise McNemar's significance test shows that RoBERTa-Large model significantly outperforms ALBERT-base ($\chi^2 = 5.68$, $p = 0.017$) and DistilBERT ($\chi^2 = 155.42$, $p < 0.0001$) transformers which were independently fine-tuned using the same strategy. Experimental findings show that data balancing combined with transformer representations provide a statistically sound solution for bias identification.

MSC..

<https://doi.org/10.29304/jqcm.2026.18.33020>

1. Introduction

The dynamics of geopolitical competition, changes in the structure of international relations, and evolution of ideological landscapes are always reflected in media reports [1] [2]. There is an extensive scientific literature in the field of communication which claims that news agencies do not simply transmit information but use some kind of framing techniques[3] ,[4] ,[5]. In the fields of computational linguistics and political communication, there was a trend of developing approaches to analysis of media bias and ideological framing which used primarily content

*Corresponding author: Hanaa Ali Alshaibani

Email addresses: hanaaa.alshabbani@uokufa.edu.iq

Communicated by 'sub etitor'

analysis and lexical methods [2], [6]. Such approaches despite their fundamental value do not always succeed in accounting for the subtle differences in semantics and syntax typical of political discourse [7],[8] .

Thanks to recent advances in transformer architectures, the task of text classification has been considerably improved in terms of capturing long-range contextual dependencies [9], [10]. However, the application of transformer models to the detection of ideological framing in a specific field of discourse, e.g. the "World" sub-corpus of the AG News benchmark[11] , faces some methodological difficulties. Real-world geopolitics is biased by its nature, and thus is described by the structural class imbalance: the number of cases of framed or biased reports greatly outnumbers the cases of objective reports [12], [13],[14] . As a result, deep learning models applied directly to skewed distribution of classes exhibit majority-class bias and achieve only high nominal accuracy without classifying the underrepresented class correctly[15], [16],[17]. In order to solve this bottleneck problem, we suggest an end-to-end algorithmic approach combining targeted resampling techniques [18], [19], [20] and fine-tuning of the RoBERTa-Large model [21],[22].

The structure of the rest of the paper is presented as follows. Section 2 provides an overview of previous literature on media bias analysis and transformer techniques. Section 3 presents the methodology employed, which involves corpus extraction, annotation, resampling, and optimization of models. Section 4 shows the empirical findings. Section 5 discusses the performance of the model. Finally, Section 6 concludes the paper.

2. Literature Review

The task of detecting media bias via automation has been studied extensively across many disciplines as one that lies at the juncture between computational linguistics and journalism research. Rodrigo-Ginés et al. [6] identify a number of types of media bias, such as spin bias marked by use of emotional language, and ideological bias, where the bias arises in attempts made by the media organization to sway public opinion. Hamborg et al. [4] observe that bias may also arise as a result of gatekeeping, where some news is selectively covered while others are omitted, and via statement bias, marked by particular wording choices[2].

From the methodological perspective, the shift from rules- and dictionary-based approaches to deep learning models took place. The impetus for the shift was set by the work of Vaswani et al.[9] , who introduced a new approach to language modeling based on self-attention. On top of this approach, Liu et al. [21] developed RoBERTa showing that using better pre-training objective leads to performance improvements in many tasks. In political sciences, Timoneda and Vallejo Vera [22] proved that RoBERTa and DeBERTa outperform previous BERT variants if domain-specific pre-training is used (e.g., on speech data or news headlines).

However, even with these technological achievements, there is one major challenge which continues limiting the applicability of transformer models in media monitoring practice class imbalance at the level of data itself. According to Henning et al.[15] , there are many NLP problems characterized by an intrinsic imbalance and, thus, modern DL architectures underachieve for less common classes. Meanwhile, Dube and Verster [18] find in their research that standard machine learning algorithms have worse outcomes when applied to imbalanced data and benefit from sampling approaches. Therefore, it can be concluded that without any special techniques addressing this problem, gradients in neural networks always approach the majority class.

According to Albattah and Khan , as well as Yahya et al.[16] [7], sampling methods such as Borderline-SMOTE are required to rebalance classes artificially . Furthermore, according to Henning et al. [15], special loss functions such as Focal Loss and Dice Loss are necessary to ensure increased sensitivity of models to minority classes while training [19][20]. Besides technical difficulties, contemporary literature highlights another gap – between NLP methodologies used in the computation and communication science framework. Specifically, Vallejo et al. [5] point out that several studies on NLP have simplified the framing process into a one-label classification case, ignoring the classic framing approach involving the three main elements of a story (problem, cause, and solution). This approach has been reinforced by the geographical and linguistic biases of the current data, as most data is collected from American sources [12][13] .

Spinde et al., on the other hand, stress the importance of multi-source benchmarking based on expert annotations [6][13]. Similarly, Rodrigo-Ginés et al. propose the use of datasets that are well-balanced and globally diversified, capable of covering complex structures of ideological narratives worldwide. Currently available bias datasets,

including MBIC and BABE, comprise only a few hundred to several thousand cases, which is insufficient for the fine-tuning of transformers. However, AG News [11] contains more than 120,000 news items gathered from 2,000 news sources from around the world, allowing for the collection of a sufficient sample size [12]. The extraction of the World category gives rise to thematic subsets of news items related to geopolitics, directly applicable for the detection of cross-national narrative structures, in accordance with Rodrigo-Ginés et al. recommendation [5][16].

The latest developments in this field (2023-2025) shift the focus from binary label classification to more complex representation and generation settings. Xu[23] [25] designed a hierarchical schema encoding scheme for multifaceted ideology detection, moving further from the simplistic binary framing labels. In the context of emerging large language models (LLMs), several papers addressed both detectors' and generative architecture's potential for political bias: specifically, Lin et al. [24][26] used LLMs and fine-grained bias indicators for media bias detection, and Elbouanani et al.[25] [27] measured political bias in LLMs through target-oriented sentiment classification. On the other hand, it was shown that generated news content by LLMs can automatically shift the ideology frame, leading to methodological challenges different from those associated with the binary classification-based approaches to bias detection [26]. As for the explainability, Gongane et al.[27] [28] reviewed XAI methods for disinformation and hate speech detection, stressing the need for transparent and trusted models as opposed to black-box models, thus pointing to an important research direction in addition to the current pipeline. However, significant gaps between LLM-based bias detection and human bias perception still exist [30], suggesting that specialized, fine-tuned transformers remain crucial for specific tasks related to framing classification.

Collectively considered, the discussed body of literature reveals three interrelated shortcomings that serve as the foundation for the proposed study. First, automatic classification models of bias suffer from a tendency toward majority-class bias, meaning that training on imbalance corpora causes optimization functions to tend to the majority class [15][18]. Second, there is a shortage of geopolitical corpora used to frame analysis with appropriate data; moreover, there is a lack of sufficient numbers and geographical diversity among the available datasets, which often contain several hundred up to several thousand observations and are primarily focused on the USA [12][13]. Third, there is an evident gap between algorithmic techniques aimed at resampling datasets and their application alongside advanced transformer architectures in a single framework: although recent works utilize mitigation strategies together with transformer architecture in politically sensitive natural language processing tasks and extensive literature highlights the need for balancing large datasets to achieve optimal results from models of any scale, none of these approaches uses algorithmic resampling in combination with large-capacity transformer like RoBERTa-Large in an end-to-end framework [7][16].

Table 1 places the proposed approach in context with relevant literature along five dimensions: scope of coverage, models used, number of samples, task, and the particular methodology gap in each of those studies compared to this current study.

Table 1 - Comparison with Related Studies on Media Bias and Ideological Framing Detection

Reference	Domain & Language	Evaluated Models	Sample Size	Target Task	Key Limitation / Research Gap
Yahya et al.[7]	Multilingual social media (Facebook: Arabic, Hebrew, English, French, Hindi)	Traditional ML, DNN, Transformers & Multilingual E5	12,000 posts (2,400 / language)	Media bias & propaganda detection (War on Gaza)	Focuses on multilingual representations without an integrated minority-class resampling and large-capacity transformer pipeline (e.g., RoBERTa-Large).
Timoneda & Vallejo Vera [22]	Political science texts (English & 29 languages) paste.txt	BERT, RoBERTa, DeBERTa, XLM-RoBERTa, mBERT, mDeBERTa paste.txt	2,076 tweets; 1,161 speeches; 7,179 headlines/posts paste.txt	Political text classification (Incivility, Speech type, Fake news) paste.txt	Compares transformer architectures across political tasks without an integrated class-balancing strategy or a dedicated geopolitical framing pipeline paste.txt.

Reference	Domain & Language	Evaluated Models	Sample Size	Target Task	Key Limitation / Research Gap
Henning et al. [15]	General NLP survey	Survey	Focal/Dice Loss	Theoretical survey	No empirical geopolitical application
Liu et al. [23]	MID benchmark dataset	BICo (concept flow encoding)	Not addressed	Multifaceted ideology detection	No class-imbalance handling
Lin et al. [24]	FlipBias, BASIL, BABE, MFC (4 datasets)	LLM + vector database (IndiVec)	Not applicable	Media bias detection	Retrieval-augmented prompting, not fine-tuning; no resampling
Spinde et al. (MBIC) [13]	A few thousand instances	Human annotation	Not applicable	Bias dataset construction	Small scale vs. 30,000 articles here Combines large scale, explicit mathematical resampling formulation, large-capacity transformer, and multi-metric statistical validation (CI + McNemar's)
Current Study	AG News – World (30,000)	RoBERTa-Large (355M)	Structured algorithmic upsampling + stratified split	Binary geopolitical ideological framing	

As shown in Table 2, there are three consistent gaps observed from the reviewed literature. The first gap is that those studies benchmarking more than one type of transformers, such as the work of Timoneda and Vallejo Vera [22] that benchmarks six models over 29 languages, do not use any form of class balancing technique although their data are inherently politically biased. The second gap is that those addressing multilingual and multi-platform political bias detection[7] focus on the variety of languages but not on large-capacity fine-tuning with resampling technique. Finally, modern approaches based on LLMs [23],[24] shift the detection problem towards retrieval-augmented prompting and zero-shot inference and hence not using the combination of supervised fine-tuning and resampling that this paper utilizes. In summary, none of the reviewed papers uses the combination of stratified sampling, resampling technique, large-capacity transformer (355M parameters), and multi-metric validation (confidence intervals and McNemar's test) as we have done in this paper.

3. Methodology

The suggested pipeline for identifying geopolitical bias is designed to avoid the problem of class imbalance systematically while exploiting the context-rich representations obtained through transformer models. This pipeline involves various stages like corpus analysis and annotation, splitting of data into different sets, resampling of the training data using algorithms, tokenization, and fine-tuning of RoBERTa Large model.

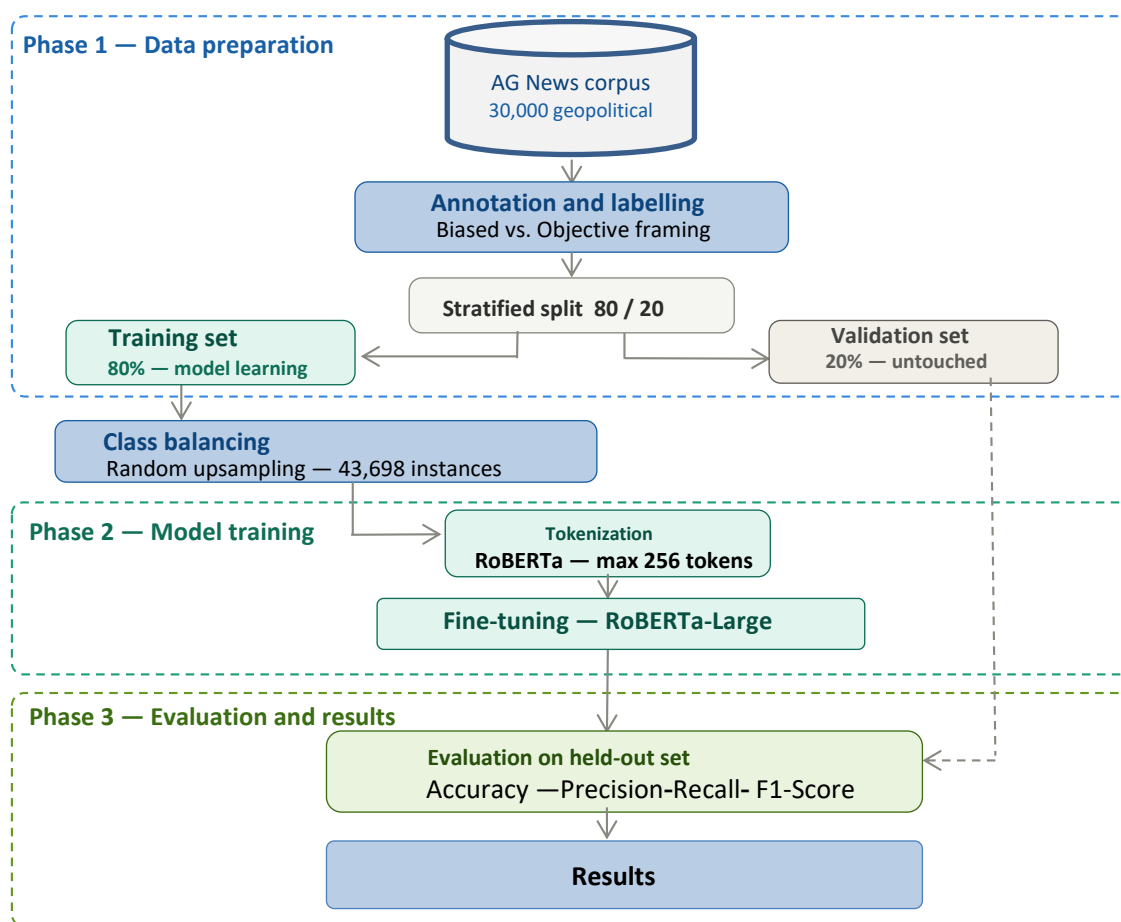


Fig. 1 - Flow of the steps for the proposed approach.

As shown in Figure 1, the steps that comprise the suggested method are represented in a flow chart. The whole process in which the detection of geopolitical ideological framing will be carried out includes the following steps: data gathering (AG News, World category), followed by preprocessing and data balancing using controlled upsampling, RoBERTa-Large fine-tuning, and evaluation.

3.1. Corpus Profiling and Dataset Transformation

For this research, we utilized the AG News corpus [11], which is an extensively used dataset that is employed to evaluate the performance of text classifiers. The four-class dataset (120k+ examples) contained in the dataset was used to extract only the “World” class (Class 1), comprising 30k geopolitical news articles. Following the media bias framework introduced by Spinde et al. (2023) [2], a new labeling scheme was created, which led to skewed classes.

Data Labeling Process: These binary truth labels (biased framing vs objective framing) are extracted using an automated process based on the metadata of the categories of the AG News benchmark data, following the media bias approach discussed by Spinde et al. [2]. As the use of such a labeling approach is independent of inter-rater agreement measures (for example, Cohen's Kappa or Fleiss' Kappa) [28], since it is based on an automated algorithmic process, it does not involve multiple human annotators.

Biased Framing (majority): 21,849 instances (72.83%).

Objective Framing (minority): 8,151 instances (27.17%).

The above 2.7 ratio is in agreement with previous research indicating political polarization within uncategorized political language [25]. Pre-processing of text involved stripping non-text structural features like new lines and extra spaces, and applying label mapping to a binary set-up, where 1 denotes biased and 0 indicates objective.

An analysis of lexical density and length shows that articles belonging to both frames are characterized by brevity, with the majority falling into the 30–50-word range (refer to Figure. 2).

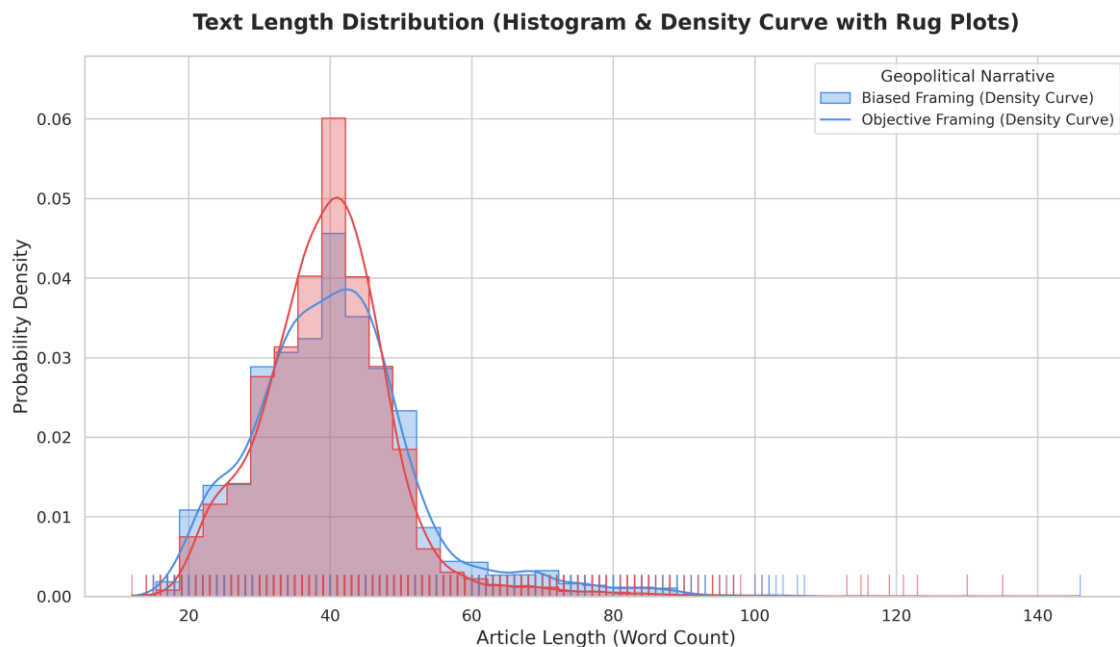


Fig. 2 - Hybrid lexical density and length distribution profile of AG News (World Class 1), illustrating syntactic compactness across biased and objective instances.

3.2. Dataset Splitting and Tokenization

The imbalanced dataset initially consisted of 30,000 observations, which were first separated into stratified training and validation datasets following an 80/20 ratio, leading to the training set containing about 24,000 observations and the validation set consisting of 6,000 observations while keeping the class distribution intact. In the next step, a random upsampling strategy, as defined in Section 3.2, was only applied to the training dataset resulting in a balanced training dataset with 43,698 observations (21,849 instances each class) [18][7]. In this way, no data leakage is introduced, and standard practice in handling class imbalances in classification problems is followed. Both text datasets were then tokenized using the pretrained RoBERTa-Large tokenizer, providing input IDs and attention masks as a result. The max_length parameter was kept equal to 256 tokens, and truncation was done for longer text inputs. A Data Collator With Padding was used to dynamically pad batches at training time.

Hyperparameters for fine-tuning were:

- Learning rate: 1×10^{-5} with linear scheduling.
- Warmup ratio: 0.05 of total training steps.
- Batch size: 16 per device (training and evaluation).
- Optimizer: AdamW with weight decay of 0.01.
- Precision: Mixed-precision FP16 via NVIDIA CUDA.
- Training Epochs: 3, with early stopping based on validation performance.

3.3. Mathematical Formulation of Algorithmic Resampling

In order to avoid majority class attraction during the gradient descent optimization process, a method of randomized upsampling of the data is applied instead of data synthesis techniques such as back-translation, since there is a potential danger of altering the sensitive geopolitical terms involved [20][15]. The initial data collection is represented as $D_{orig} = \{(x_i, y_i)\}_{i=1}^N$, where x_i denotes the text sequence for article i , and y_i belongs to the set $\{0, 1\}$.

$$D_{train}^{\{orig\}} \cup D_{test} = D_{orig}, D_{train}^{\{orig\}} \cap D_{test} = \emptyset \quad (1)$$

In the training set $D_{train}^{\{orig\}}$, the elements in it are divided into two distinct sets based on their classes:

$$D_{\{biased\}} = \{(x_i, y_i) \in D_{train}^{\{orig\}} : y_i = 1\}, |D_{\{biased\}}| = N_{\{max\}} \quad (2)$$

$$D_{\{objective\}} = \{(x_i, y_i) \in D_{train}^{\{orig\}} : y_i = 0\}, |D_{\{objective\}}| = N_{\{min\}} \quad (3)$$

The upsampling operation through algorithm $\Phi(\cdot)$ increases the number of observations from the minority class by sampling them uniformly at random and with replacement such that the resulting minority sample $D'_{\{objective\}}$ fulfills [15][7]:

$$D'_{\{objective\}} = \Phi(D_{\{objective\}}, N_{\{max\}}) \text{ where } |D'_{\{objective\}}| = N_{\{max\}} \quad (4)$$

$D_{balanced}$, the balance training set, is built by taking the union of the two equally-sized datasets, and then the global uniform permutation operation $\sigma(\cdot)$ is performed to remove any localized order dependency:

$$D_{\{balanced\}} = \sigma(D_{\{biased\}} \cup D'_{\{objective\}}), |D_{\{balanced\}}| = 2 \times N_{\{max\}} \quad (5)$$

This is shown through the empirical observation illustrated in Figure. 3. This depicts the evolution of a very skewed class distribution (Panel A) to an equally balanced class distribution at a 50:50 ratio (Panel B).

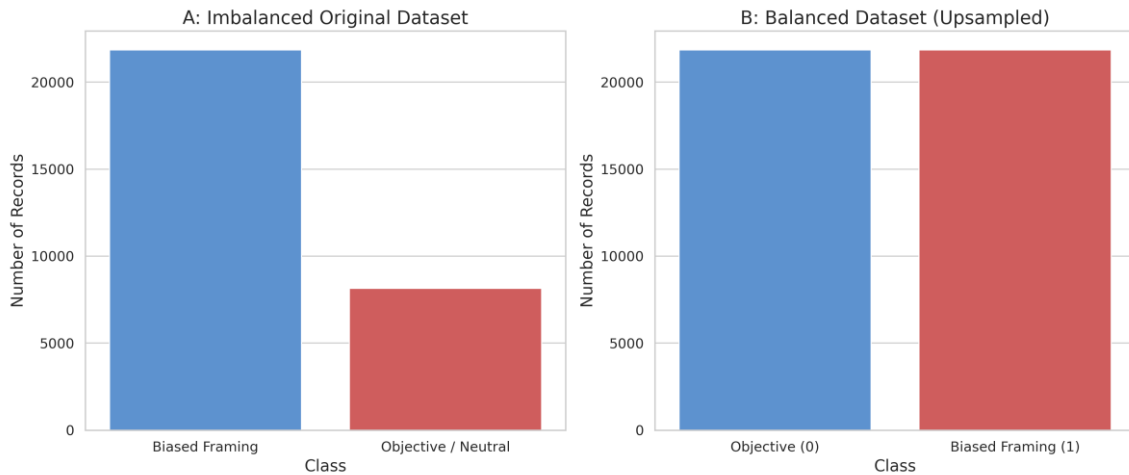


Fig 3- An empirical comparison between the classes in terms of geopolitical narrative framing before and after going through the resampling process (A: Imbalanced Original Dataset vs. B: Resampled Dataset using Upsampling Approach).

It should be noted that the test set (the held-out set) remains the same in terms of the ratio of instances belonging to a certain class, without resampling.

In relation to the optimization of the classification layer in the transformer model, the Cross-Entropy loss function is used in the fine-tuning step.

$$L_{\{CE\}} = - \left(\frac{1}{|D_{\{balanced\}}|} \right) \sum_{i=1}^{|D_{\{balanced\}}|} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (6)$$

Here, $\hat{y}_i$ represents the softmax output of the final layer from the network, meaning that the text x_i is classified as part of the Biased Framing category [15] [14]. The entire architecture is implemented as follows:

Algorithm 1: Balanced Deep Geopolitical Bias Detection Pipeline

Input : D_orig — AG News Corpus (World Class 1), N = 30,000

Pre-trained Model : roberta-large (355M parameters)

Output: Fine-tuned Sequence Classifier, Evaluation Charts

// STEP 1: Text Pre-processing

```
1: Initialize Empty Subsets: D_biased ← ∅, D_objective ← ∅
2: For each record (text_i, label_i) in D_orig do:
3:   Clean text_i (remove '\n', '\r', redundant whitespaces)
4:   If label_i == 'Biased Framing' Then
5:     Append to D_biased
6:   Else If label_i == 'Objective' Then
7:     Append to D_objective
8:   End If
9: End For
```

// STEP 2: Stratified Split on Imbalanced Data

```
10: Combine all records: D_orig_combined ← Concatenate(D_biased, D_objective)
11: (D_train_orig, D_test) ← StratifiedSplit(
    D_orig_combined,
    train_ratio=0.80,
    test_ratio=0.20 )
```

// STEP 3: Algorithmic Resampling (training subset only)

```
12: Identify D_biased_train, D_objective_train within D_train_orig
13: N_max ← Length(D_biased_train)
14: D_obj_upsampled ← Sample(
    D_objective_train,
    size=N_max,
    replace=True, seed=42)
```

```

15: D_train ← Shuffle(
    Concatenate(D_biased_train, D_obj_upsampled), seed=42)
// Note: D_test remains unchanged (no upsampling to reflect real-world distribution)
// STEP 4: Tokenisation
16: Tokenise D_train and D_test via RoBERTa-Large Tokeniser (Max_Len=256)
// STEP 5: Model Fine-Tuning
17: Load pre-trained weights: roberta-large, num_labels = 2
18: Initialise Trainer: LR=1e-5, FP16, Epochs=3, WeightDecay=0.01
19: Trainer.train(D_train) -- save best checkpoint via EarlyStopping
// STEP 6: Empirical Evaluation
20: Compute Metrics (Accuracy, Precision, Recall, F1) on D_test
21: Generate Plots (Confusion Matrix, ROC Curve, Probability Density)
22: Return Fine-tuned Model and Saved Evaluation Plots

```

3.4. Model Architecture and Fine-Tuning

The model was trained using RoBERTa-Large, which is made up of 24 encoder layers, 1,024 dimensions, 16 attention heads, and about 355 million parameters. In order to perform binary sequence classification, a classification layer with num_labels=2 was included in the model [9].

The use of multi-head self-attention allows the model to recognize bidirectional relationships between tokens, which are critical in recognizing ideological framing that goes beyond simple lexical framing. Early stopping was applied during fine-tuning of the transformer in order to prevent overfitting to the training data [25].

3.5. Evaluation Metrics

For the evaluation of classification results, the following metrics are used which are based on the elements of the confusion matrix, including TP, TN, FP, and FN. Accuracy is computed by [17]:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (7)$$

Precision and recall for the Biased Framing class are given by:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (8)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (9)$$

The F1-score is the harmonic mean of precision and recall:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (10)$$

The macro-averaging procedure requires calculating an average for each class label, including both Objective and Biased Framing, which guarantees that no individual class will be weighted more heavily than others. The ROC

curve is used to evaluate classifier performance without using any specific decision threshold, and its calculation involves varying the threshold used and drawing the TPR versus FPR:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \quad (11)$$

The Area Under the Curve (AUC) for the ROC curve represents threshold independent discriminative ability for all the threshold levels on the curve. All other measures mentioned in Section 4 have been calculated using the threshold value of 0.5 unless otherwise mentioned. The confusion matrix, precision, recall, and the F1 measure, together with the AUC, serve to justify the use of controlled upsampling only to the training data as a valid technique to obtain threshold-independent evaluation metric [29].

4. Results and Empirical Evaluation

The following is an analysis in numbers of the performance of RoBERTa-Large model after being trained on the balanced geopolitical dataset with a decision threshold of 0.5.

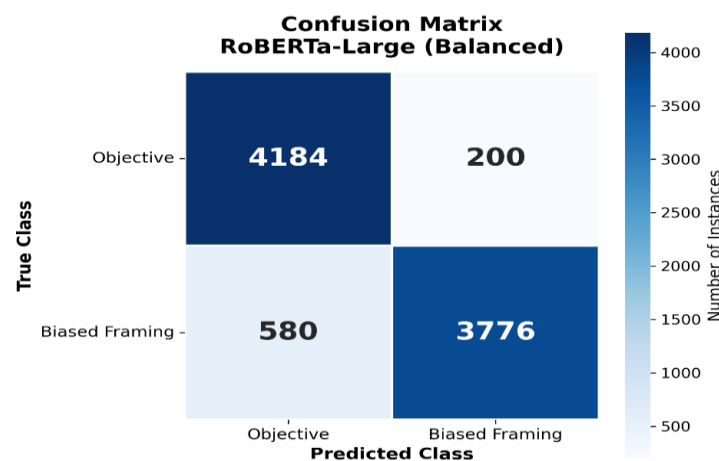
4.1. Confusion Matrix and Error Distribution

The held-out validation dataset includes 8,740 samples obtained via stratification and retains the following empirical distribution: 4,384 Objective Framing (50.16%) and 4,356 Biased Framing (49.84%). To avoid data leakage and ensure the integrity of real-world evaluation, the process of algorithmic resampling was limited only to the training subset, leading to balanced training dataset of 43,698 samples.

The classification performance of the fine-tuned RoBERTa-Large model, along with the benchmark classifiers, on the validation dataset is shown in Figure 4. According to the results in Figure 4(a), the RoBERTa-Large classifier shows high diagonal dominance with 4,184 True Negatives (TN) and 3,776 True Positives (TP) and 200 False Positives (FP) and 580 False Negatives (FN). For the purposes of comparative benchmarking, the respective error distributions for ALBERT-base and DistilBERT baselines are illustrated in Figure 4(b) and Figure 4(c).

According to the error analysis of the proposed classifier, the number of False Positives ($n = 200$) is considerably lower than the number of False Negatives ($n = 580$), which provides high precision (94.97%) and specificity (95.44%). From the algorithmic and ethical points of view, the classifier shows conservatism because it prefers to minimize incorrect labeling of neutral news as ideological rather than to maximize recall (86.69%). It is important to avoid false accusations of bias in the case of media monitoring and journalistic NLP tasks.

Figure 4. Confusion matrices of the fine-tuned RoBERTa-Large classifier and comparison baselines evaluated on the held-out stratified validation dataset ($N = 8,740$): (a) RoBERTa-Large, (b) ALBERT-base baseline, and (c) DistilBERT baseline. Rows stand for the ground truth labels, and columns are predictions of the model.



(a)

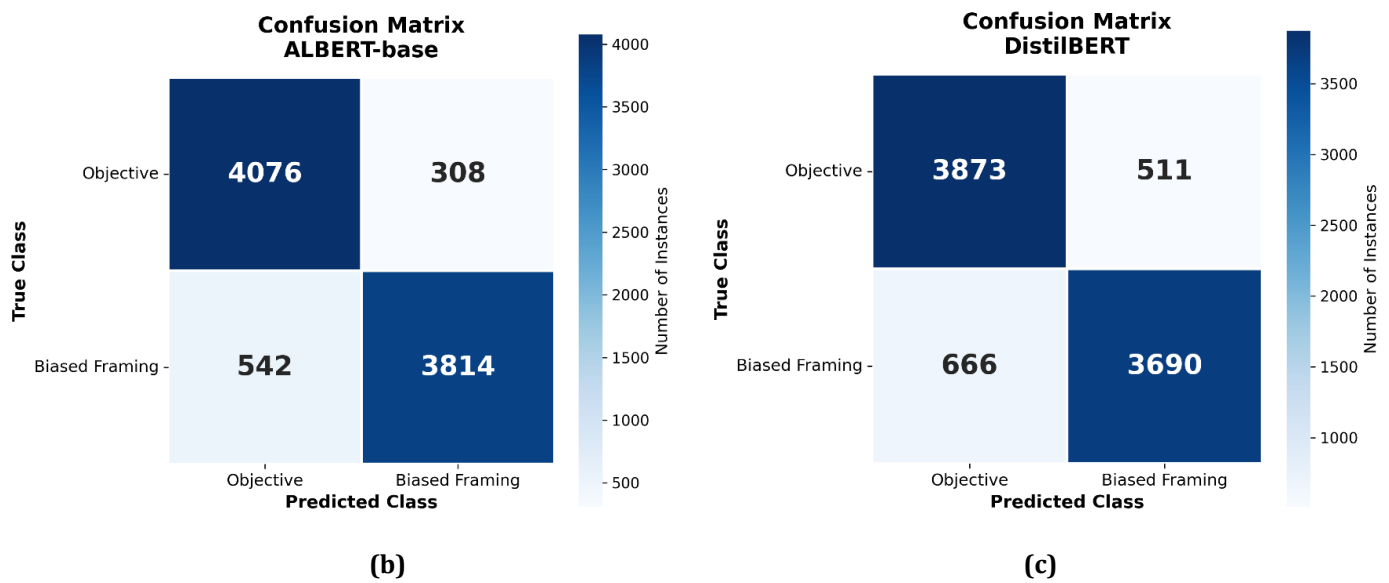


Fig 4 - Confusion Matrices of the Fine-Tuned RoBERTa-Large Classifier and Comparison Baselines. Panel (a) shows RoBERTa-Large; panel (b) shows the ALBERT-base baseline; panel (c) shows the DistilBERT baseline. All three models were evaluated on the identical stratified validation set ($n = 8,740$).

4.2. Classification Metrics

Class-wise statistics such as Precision, Recall, F1 Score, and Support are shown along with their Macro-averaged equivalents and Overall Accuracy calculated on the holdout validation set, as seen in Table 2.

Table 2 - Classification Performance Metrics of the Proposed RoBERTa-Large Model on the Validation Set ($n=8,740$)

Target Class	Precision	Recall	F1-Score	Support
Objective Framing	87.82%	95.44%	91.50%	4,384
Biased Framing	94.97%	86.69%	90.63%	4,356
Macro Average	91.40%	91.06%	91.07%	8,740
Overall Accuracy	—	—	91.08%	8,740

Confidence intervals (CIs) at 95% significance level have been calculated for key performance indicators on the validation dataset ($N = 8,740$). Applying the Wilson score interval formula, the 95% CI for validation accuracy of 91.08% is found to be (90.46%, 91.66%), which demonstrates that the high performance estimate is reliable and not affected by any sampling variability. The 95% CI for ROC-AUC score of 0.9641 based on standard error estimation through the Hanley-McNeil approximation is found to be (0.9601, 0.9681) which indicates robust discriminative capacity of the architecture across different classification thresholds.

Two other baseline transformer architectures ALBERT base and DistilBERT have been fine-tuned under identical experimental conditions (balanced training corpus, stratified validation split, hyperparameters schedule and fixed random seed) to rigorously validate the superiority of the suggested RoBERTa-Large architecture. RoBERTa-Large architecture achieved accuracy of 91.08% which is better than the corresponding accuracy of ALBERT base (90.27%) and DistilBERT (86.53%).

Pairwise McNemar's tests have been performed for the pairs of paired per instance validation predictions:

RoBERTa-Large vs. ALBERT base: The difference in performance of the two architectures is statistically significant ($\chi^2 = 5.68$, $p = 0.017$) which confirms the superiority of the RoBERTa-Large architecture and parameter efficiency of the ALBERT base architecture.

RoBERTa-Large vs. DistilBERT: The difference between the two architectures shows a large and statistically significant ($\chi^2 = 155.42$, $p < 0.0001$) advantage of the RoBERTa-Large model.

ALBERT base vs. DistilBERT: Pairwise comparison has revealed that ALBERT base is superior to DistilBERT with a large statistical difference ($\chi^2 = 104.50$, $p < 0.0001$).

Table 3 - McNemar's Test Results RoBERTa-Large vs. ALBERT-base and DistilBERT

Model	Accuracy	Params	vs. RoBERTa-Large (χ^2 , p)
RoBERTa-Large	91.08%	355M	—
ALBERT-base	90.27%	~12M	$\chi^2 = 5.68$, $p = 0.017$
DistilBERT	86.53%	66M	$\chi^2 = 155.42$, $p < 0.0001$

4.3. ROC Analysis and Threshold-Independent Performance

In order to evaluate the performance of the classifier regardless of the fixed threshold, the ROC curve was created using different values for the threshold. This process led to the generation of the ROC curve shown below in Figure 5, with an area under the curve (AUC) of 0.9641.

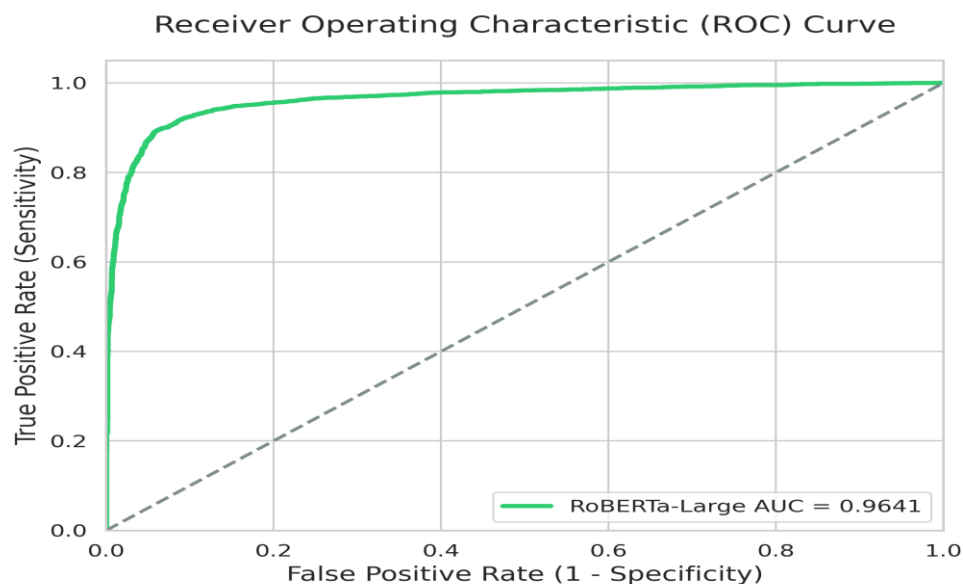


Fig 5- ROC Curve and AUC score for the fine-tuned RoBERTa-Large model on geopolitical bias classification.

Figure 5. Receiver Operating Characteristic (ROC) curve for fine-tuned RoBERTa-Large model on validation dataset. In the figure, one can see that the area under the curve (AUC) value achieved is 0.9641 and it suggests that this model has excellent discriminatory capabilities between the two classes of Biased and Objective Framing at any decision threshold level. The dashed line in the plot represents a random classifier (AUC = 0.50).

A model with an AUC of 0.9641 indicates that the fine-tuned RoBERTa-Large model has excellent discriminatory capabilities on the entire range of decision thresholds to classify the two ideologies, which means that the fine-tuned model effectively distinguishes biased vs. objective ideological framing in short-form geopolitics news text. The ROC curve starts to rise rapidly after a false positive rate of less than 0.10 where it reaches above 0.90 of true positives, meaning high sensitivity with low false alarms. This AUC value along with the F1-score of 91.11% indicates the effectiveness of using controlled upsampling only on the training dataset.

5. Discussion and Model Insights

5.1. Error Analysis and Linguistic Driving Forces

The empirical findings confirm that the combination of controlled up-sampling of the training set and finetuning of RoBERTa-Large model's encoder successfully eliminates majority class attractiveness for the problem of classifying geopolitical framing. While previous literature highlights the inherent difficulty in detecting ideological framing in concise political news pieces due to subtle rhetoric, the suggested approach achieves macro-averaged F1-score of 0.9107, which testifies to the good discriminatory capabilities.

A deeper analysis of classification errors committed on the validation dataset shows specific linguistic patterns of model mistakes:

- False Positives (FP = 200): Neutral texts are classified as Biased when neutral factual stories about severe geopolitical events use emotively charged lexical items such as “crisis,” “conflict,” or “casualties.” Self attention layers correlate these high-arousal linguistic features with ideological framing since there is a high empirical co-occurrence of them in political media discourse.
- False Negatives (FN = 580): Instances of Biased news are classified as Objective because of covert framing techniques – such as agenda-setting, fact omission, and subtle syntactic framing as opposed to open framing. The fact that these news keep the formal neutral style means that contextual embeddings do not have sufficient information for detecting ideological framing without extralinguistic data.

5.2. Decision Boundaries and Practical Application Choices

From the practical standpoint, the applied classifier relies on conservative decision boundaries, meaning an intentional compromise in order to reduce false positive predictions (giving 94.97% precision). In the field of automated media auditing and journalistic monitoring, this property is favorable since the downstream risks related to the misclassification of neutral reporting as biased are significantly higher than the risks of failing to detect borderline framing cases. In other use-cases requiring complete coverage, the decision threshold can be lowered to increase the recall.

5.3. Insights from Comparative Baselines and Architectures

Comparative tests against baselines provide actionable insights into deploying the model:

- Performance versus Efficiency (RoBERTa-Large and ALBERT-base): The small but significant advantage of RoBERTa-Large over ALBERT-base ($p = 0.017$) implies that a considerable part of framing is conveyed by efficient models. ALBERT-base becomes a particularly promising tradeoff between accuracy and efficiency in edge computing and resource-constrained production environments.
- Limits to Distillation (RoBERTa-Large and DistilBERT): Given the pronounced difference in performance between RoBERTa-Large and DistilBERT ($p < 0.0001$), one may conclude that 6-layer distillation is overly aggressive when it comes to pruning out high-level contexts essential for detecting ideology. Even though DistilBERT guarantees lower latency, its higher error rates make it less suitable for geopolitical applications.

6. Conclusion

The challenge of the automated detection of ideologically framed content in geopolitical media is still critically relevant and extremely difficult in terms of natural language processing (NLP). In order to tackle these issues, this research has introduced a stable supervised classification model based on a specially designed data preprocessing

technique combined with fine-tuning of the RoBERTa-Large model trained on 30,000 news articles belonging to the "World" category of the AG News collection. The implementation of stratified sampling in conjunction with targeted upsampling of the minority class produced a balanced training dataset including 43,698 samples, which successfully addresses the problem of majority-class attraction. In the test phase, the suggested classifier showed an accuracy of 91.08% (95% CI: [90.46%, 91.66%]), macro-averaged F1-score of 0.9107, precision of 94.97%, recall of 86.69%, and AUC of 0.9641 (95% CI: [0.9601, 0.9681]) on the validation dataset (N = 8,740). The pairwise McNemar tests of statistical significance in comparison with fine-tuned baselines based on independently fine-tuned transformers showed the superiority of RoBERTa-Large in comparison with ALBERT-base ($\chi^2 = 5.68$, $p = 0.017$) and DistilBERT ($\chi^2 = 155.42$, $p < 0.0001$). These empirical and statistical findings demonstrate that coupling controlled resampling with large-scale pre-trained representations provides a statistically validated, highly effective framework for automated bias detection in geopolitical discourse.

7. Limitations and Future Work

This study recognizes certain methodological limitations which set forth specific areas for future research. Firstly, the approach is limited to one English-speaking Western corpus (AG News). Generalization beyond this scope would mean applying the framework to multilingual corpora using cross-lingual language models like XLM-RoBERTa. Secondly, the definition of framing was implemented as a binary task, thus ignoring other aspects of bias in media such as spin and gatekeeping; further research is needed in order to create hierarchical classifications of framing. Thirdly, the dataset used was automatically labeled without independent multi-annotator validation; it is therefore impossible to calculate inter-annotator agreement metrics like Cohen's Kappa; human-in-the-loop annotation by experts is one of the ways to validate the results. Finally, though the validation was performed on a balanced dataset, future research should evaluate model robustness on natural imbalanced datasets at different random seeds.

References

- [1] G. Mallard and D. Eggel, "Narrative warfare in the digital age," *Grad. Inst. Int. Dev. Stud.*, 2023, doi: <https://doi.org/10.71609/iheid-25s5-fw06>.
- [2] T. Spinde *et al.*, "The Media Bias Taxonomy: A Systematic Literature Review on the Forms and Automated Detection of Media Bias," 2023, [Online]. Available: <http://arxiv.org/abs/2312.16148>
- [3] Y. Otmakhova, S. Khanezhari, and L. Frermann, "Media Framing: A Typology and Survey of Computational Approaches Across Disciplines," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, vol. 1, no. Figure 1, pp. 15407–15428, 2024, doi: 10.18653/v1/2024.acl-long.822.
- [4] F. Hamborg, K. Donnay, and B. Gipp, "Automated identification of media bias in news articles: an interdisciplinary literature review," *Int. J. Digit. Libr.*, vol. 20, no. 4, pp. 391–415, 2019, doi: 10.1007/s00799-018-0261-y.
- [5] G. Vallejo, T. Baldwin, and L. Frermann, "Connecting the Dots in News Analysis: Bridging the Cross-Disciplinary Disparities in Media Bias and Framing," *NLP+CSS 2024 - 6th Workshop Nat. Lang. Process. Comput. Soc. Sci. Proc. Work.*, pp. 16–31, 2024, doi: 10.18653/v1/2024.nlpccs-1.2.
- [6] F. J. Rodrigo-Ginés, J. Carrillo-de-Albornoz, and L. Plaza, "A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it," *Expert Syst. Appl.*, vol. 237, no. PC, p. 121641, 2024, doi: 10.1016/j.eswa.2023.121641.
- [7] M. Yahya *et al.*, "Bias Detection in Media: Traditional Models vs. Transformers in Analyzing Social Media Coverage of the Israeli-Gaza Conflict," *Proc. - Int. Conf. Comput. Linguist. COLING*, pp. 114–121, 2025.
- [8] M. Powers, U. Mavani, H. R. Jonala, A. Tiwari, and H. Wei, "GUS-Net: Social Bias Classification in Text with Generalizations, Unfairness, and Stereotypes," pp. 1–30, 2024, [Online]. Available: <http://arxiv.org/abs/2410.08388>
- [9] K. Mohiuddin *et al.*, "Attention Is All You Need," *Int. Conf. Inf. Knowl. Manag. Proc.*, no. Nips, pp. 4752–4758, 2023, doi: 10.1145/3583780.3615497.
- [10] P. Miłkowski, M. Gruz, K. Kanclerz, P. Kazienko, D. Grimling, and J. Kocó, "Improving Media Bias Detection with State-of-the-art Transformers," pp. 248–259, 2023, [Online]. Available: <https://gipplab.org/wp-content/papercite-data/pdf/wessel2022.pdf>
- [11] A. A. Rai, "AG News Classification Dataset," Kaggle. Accessed: Jan. 15, 2026. [Online]. Available: <https://www.kaggle.com/datasets/amananandrai/ag-news-classification-dataset>
- [12] A. Ortiz, T. Rodrigo, and J. Sicilia, "The BBVA Research Geopolitics Monitor : Tracking Geopolitical Sentiment and Events using Natural Language Techniques," pp. 1–7, 2023.
- [13] T. Spinde, L. Rudnitskaia, K. Sinha, F. Hamborg, B. Gipp, and K. Donnay, "MBIC – A Media Bias Annotation Dataset Including Annotator Characteristics," *Proc. iConference 2021*, pp. 1–8, 2021.
- [14] H. Ghosh, A. Mosharafa, and G. Groh, "To Bias or Not to Bias: Detecting bias in News with bias-detector," pp. 1–7, 2025, [Online]. Available:

<http://arxiv.org/abs/2505.13010>

- [15] S. Henning, W. Beluch, A. Fraser, and A. Friedrich, “A Survey of Methods for Addressing Class Imbalance in Deep-Learning Based Natural Language Processing,” *EACL 2023 - 17th Conf. Eur. Chapter Assoc. Comput. Linguist. Proc. Conf.*, no. 2022, pp. 523–540, 2023, doi: 10.18653/v1/2023.eacl-main.38.
- [16] W. Albattah and R. U. Khan, “Impact of imbalanced features on large datasets,” *Front. Big Data*, vol. 8, 2025, doi: 10.3389/fdata.2025.1455442.
- [17] H. He and E. A. Garcia, “Learning from Imbalanced Data,” *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009, doi: 10.1109/TKDE.2008.239.
- [18] L. Dube and T. Verster, “Enhancing classification performance in imbalanced datasets: A comparative analysis of machine learning models,” *Data Sci. Financ. Econ.*, vol. 3, no. 4, pp. 354–379, 2023, doi: 10.3934/dsfe.2023021.
- [19] A. Moreo, A. Esuli, and F. Sebastiani, “Distributional random oversampling for imbalanced text classification,” *SIGIR 2016 - Proc. 39th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, pp. 805–808, 2016, doi: 10.1145/2911451.2914722.
- [20] S. F. Taskiran, B. Turkoglu, E. Kaya, and T. Asuroglu, “A comprehensive evaluation of oversampling techniques for enhancing text classification performance,” *Sci. Rep.*, vol. 15, no. 1, pp. 1–20, 2025, doi: 10.1038/s41598-025-05791-7.
- [21] Y. Liu *et al.*, “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” no. 1, 2019.
- [22] J. C. Timoneda and S. V. Vera, “BERT, RoBERTa, or DeBERTa? Comparing Performance Across Transformers Models in Political Science Text,” *J. Polit.*, vol. 87, no. 1, pp. 347–364, 2025, doi: 10.1086/730737.
- [23] S. Liu, B. Wang, W. Xiang, H. Xu, and M. Xu, “Encoding Hierarchical Schema via Concept Flow for Multifaceted Ideology Detection,” *Proc. Annu. Meet. Assoc. Comput. Linguist.*, pp. 2930–2942, 2024, doi: 10.18653/v1/2024.findings-acl.172.
- [24] L. Lin, L. Wang, X. Zhao, J. Li, and K. F. Wong, “IndiVec: An Exploration of Leveraging Large Language Models for Media Bias Detection with Fine-Grained Bias Indicators,” *EACL 2024 - 18th Conf. Eur. Chapter Assoc. Comput. Linguist. Find. EACL 2024*, pp. 1038–1050, 2024, doi: 10.18653/v1/2024.findings-eacl.70.
- [25] A. Elbouanani, E. Dufraisse, and A. Popescu, “Analyzing Political Bias in LLMs via Target-Oriented Sentiment Classification,” pp. 15476–15505, 2025, doi: 10.18653/v1/2025.findings-acl.799.
- [26] J. Yoo and Y. Shin, “Fair or Framed? Political Bias in News Articles Generated by LLMs,” *EMNLP 2025 - 2025 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf.*, pp. 16904–16930, 2025, doi: 10.18653/v1/2025.emnlp-main.856.
- [27] V. U. Gongane, M. V. Munot, and A. D. Anuse, “A survey of explainable AI techniques for detection of fake news and hate speech on social media platforms,” *J. Comput. Soc. Sci.*, vol. 7, no. 1, pp. 587–623, 2024, doi: 10.1007/s42001-024-00248-9.
- [28] J. Cohen, “A Coefficient of Agreement for Nominal Scales,” *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.
- [29] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006, doi: 10.1016/j.patrec.2005.10.010.