Specialists vs. Generalists: A Hybrid Ensemble Approach for Multilingual Bias Detection in Conflict News (FIGNEWS 2024)

Amenah Medhat*, Shorouk Elbehairy†, Amr Hossam‡, Hossam Nasr§, Fares Ahmed¶, and Mohamed Nashaat‖
School of Information Technology and Computer Science
Nile University
Cairo, Egypt
*A.Medhat2292@nu.edu.eg †S.Sherif2245@nu.edu.eg ‡A.Hossam2232@nu.edu.eg
§H.nasr2270@nu.edu.eg ¶F.ahmed2270@nu.edu.eg ‖M.nashaat2265@nu.edu.eg

***Abstract*—The detection of political bias in conflict-related news presents formidable challenges when addressing multilingual content across morphologically divergent languages. This work addresses the FIGNEWS-2024 Shared Task on bias detection in news articles covering the Israel-Palestine conflict, focusing on classifying bias direction (Unbiased, Biased Against Palestine, Biased Against Israel, Others) in bilingual English and Arabic sources. Our approach confronts significant class imbalance, where the majority class dominates the distribution while critical minority classes require heightened detection sensitivity. We propose a novel "Mixture of Experts" ensemble architecture that implements language-specific routing to specialized models—MARBERTv2 for Arabic text, DeBERTa-v3 for English text, classical Random Forest models with tailored feature extraction, and XLM-RoBERTa as a multilingual bridge—structured as a five-stage sequential pipeline. Our methodology designates this ensemble as "System 1" and evaluates it against a generalist "System 2" baseline using standalone XLM-RoBERTa. We introduce a data maximization strategy that leverages machine translation to train language-specific specialists on the complete dataset, combined with a cross-lingual augmentation technique that doubles training instances for the multilingual model by pairing original texts with their translations under shared labels. Our language-aware voting mechanism routes input text to domain-specific expert panels (three voters per language), enabling complementary strength exploitation across statistical and neural architectures while reducing model hallucinations through multi-perspective corroboration. This specialist ensemble approach effectively addresses the unique challenges of cross-lingual bias detection in highly polarized conflict narratives.**

***Index Terms*—Bias detection, multilingual NLP, ensemble learning, mixture of experts, Arabic NLP, transformer models, conflict media analysis, class imbalance, XLM-RoBERTa, MARBERT, DeBERTa**

I. Introduction

The Israel-Palestine conflict has generated sustained international media attention, with news coverage reflecting diverse political, cultural, and ideological perspectives across global outlets. The proliferation of digital journalism and social media has intensified concerns regarding media bias, particularly in conflict zones where partisan reporting can shape public opinion, influence diplomatic relations, and deepen societal divisions. Automated detection of bias direction in news articles has become critical for media literacy initiatives, fact-checking organizations, and fostering balanced information ecosystems in politically sensitive contexts.

However, detecting bias in conflict reporting presents substantial technical challenges. First, news articles covering the Israel-Palestine conflict are published predominantly in two languages with fundamentally different linguistic architectures—English, an analytical language with minimal inflectional morphology, and Arabic, a highly inflected Semitic language with complex morphological derivation, root-pattern systems, and dialectal variation. Second, the inherent subjectivity of bias annotation introduces significant inter-annotator disagreement, particularly in politically charged domains where trained annotators themselves may carry unconscious perspectives. Third, severe class imbalance characterizes real-world bias datasets, where unbiased articles substantially outnumber explicitly biased content, creating optimization challenges for supervised learning systems. Fourth, the cultural and contextual nuances embedded in conflict narratives often manifest through implicit framing, word choice, and omission rather than explicit lexical markers, demanding models capable of capturing subtle semantic distinctions.

The FIGNEWS-2024 Shared Task addresses these challenges by providing a curated multilingual dataset of news articles annotated for bias direction, enabling systematic evaluation of automated bias detection systems. Our participation in this shared task explores a fundamental architectural question: whether a coordinated ensemble of language-specific specialists can outperform generalist multilingual models by exploiting domain expertise and statistical robustness through strategic model composition.

This paper makes the following contributions:

- We design a five-stage sequential pipeline architecture that constructs specialized expert models incrementally—classical statistical baselines (Random Forest with FastText and TF-IDF), monolingual transformer specialists (MARBERTv2, DeBERTa-v3), and a multilingual bridge (XLM-RoBERTa)—culminating in a unified ensemble inference engine.

- We implement a data maximization strategy that trains language-specific models on the complete dataset by leveraging machine translations, ensuring specialists observe maximum linguistic variation while avoiding cross-lingual contamination, and a cross-lingual augmentation technique that explicitly teaches the multilingual model label invariance across translation pairs.
- We develop a language-aware voting mechanism that dynamically routes input text to appropriate expert panels—Arabic specialists (MARBERT, Arabic Random Forest, XLM-RoBERTa) for Arabic content and English specialists (DeBERTa, English Random Forest, XLM-RoBERTa) for English content—with majority voting and specialist-weighted tie-breaking to aggregate predictions.
- We employ dual preprocessing strategies tailored to model architectures: context-preserving raw text for transformer-based models that leverage pre-trained contextual representations, and extensively normalized feature-engineered text for classical machine learning models that benefit from explicit linguistic preprocessing including character normalization, stopword removal, and n-gram extraction.
- We conduct comprehensive comparative experiments evaluating our "System 1" ensemble approach against a "System 2" generalist baseline using standalone XLM-RoBERTa, providing empirical evidence for the specialists versus generalists paradigm in multilingual bias detection under severe class imbalance.

The remainder of this paper is structured as follows. Section II reviews related work in multilingual NLP, bias detection methodologies, and ensemble learning approaches. Section III describes our methodology in detail, including data preprocessing pipelines, architecture design for each pipeline stage, and the ensemble voting logic. Section IV presents our experimental setup, training configurations, and dataset characteristics. Section V analyzes our results with comparative performance metrics and error analysis. Section VI concludes with implications for conflict media analysis and directions for future research.

II. Related Work

A. Arabic Natural Language Processing

Arabic NLP has undergone substantial advancement with the development of pre-trained transformer models specifically optimized for Arabic and its dialectal variations. MARBERTv2, introduced by Abdul-Mageed et al. [1], represents a state-of-the-art BERT-based architecture trained on over one billion Arabic tweets and news articles encompassing both Modern Standard Arabic (MSA) and regional dialectal variants. Unlike earlier Arabic language models that exhibited limited performance on informal or dialectal text, MARBERTv2 demonstrates robust generalization across diverse linguistic registers by leveraging a comprehensive training corpus spanning social media discourse, journalistic content, and classical literature. The model employs a subword tokenization scheme that effectively handles the morphological complexity inherent to Arabic, including agglutinative morphology, cliticization patterns, and diacritical marking systems. This morphological awareness enables MARBERTv2 to capture fine-grained semantic distinctions critical for bias detection in Arabic news text, where word choice and morphological variation often signal ideological positioning.

Multilingual models such as XLM-RoBERTa, proposed by Conneau et al. [2], adopt an alternative paradigm by training a unified model on 100 languages simultaneously using a shared cross-lingual vocabulary. XLM-RoBERTa extends the RoBERTa architecture with cross-lingual masked language modeling objectives, enabling effective transfer learning and zero-shot inference on low-resource languages through shared representational spaces. While XLM-RoBERTa exhibits impressive cross-lingual capabilities and has demonstrated success on multilingual benchmarks, empirical studies suggest that monolingual specialists such as MARBERTv2 frequently outperform multilingual alternatives on language-specific tasks due to dedicated model capacity, absence of cross-lingual interference, and language-targeted pre-training objectives [3].

For English text processing, DeBERTa-v3 represents one of the most advanced transformer architectures currently available. Introduced by He et al. [4], DeBERTa (Decoding-enhanced BERT with disentangled attention) implements a disentangled attention mechanism that separately encodes content embeddings and positional embeddings, contrasting with traditional BERT architectures that fuse these representations. Additionally, DeBERTa incorporates an enhanced mask decoder that leverages absolute positional information during the final prediction layer, enabling more precise token-level predictions. These architectural innovations position DeBERTa-v3 as a state-of-the-art model for natural language understanding tasks, demonstrating superior performance on GLUE, SuperGLUE, and other English NLU benchmarks compared to BERT, RoBERTa, and earlier DeBERTa variants.

B. Bias Detection in News Media

Bias detection in journalistic content has progressed from rule-based keyword matching and lexicon-driven sentiment analysis to sophisticated neural architectures capable of capturing contextual nuance. Early methodologies relied on manually curated bias lexicons and surface-level linguistic features such as subjective language markers, hedging expressions, and source attribution patterns [5]. While these approaches provided interpretable signals, they struggled to capture implicit bias manifestations embedded in narrative framing, selective fact presentation, and subtle rhetorical devices.

The advent of transformer-based pre-trained language models has fundamentally transformed bias detection capabilities. SemEval-2019 Task 4 on hyperpartisan news detection [6] and SemEval-2020 Task 11 on propaganda technique detection [7] established benchmark datasets and evaluation protocols for fine-grained bias classification. These shared tasks demonstrated that transformer models fine-tuned on annotated corpora substantially outperform feature-engineered approaches

by learning latent representations of bias that transcend surface lexical patterns. Da San Martino et al. [8] showed that BERT-based models pre-trained on large-scale news corpora capture discourse-level bias signals through contextual embeddings, enabling detection of sophisticated propaganda techniques such as loaded language, causal oversimplification, and appeal to authority.

However, existing bias detection research has predominantly focused on English monolingual settings, with limited exploration of multilingual and cross-lingual scenarios. The morphological richness of Arabic, combined with cultural and political context differences between English and Arabic news ecosystems, necessitates specialized approaches that account for language-specific bias manifestations. Recent work on Arabic media bias remains sparse, highlighting the need for architectures that can effectively handle bilingual conflict reporting where bias signals may manifest differently across translation pairs.

C. Ensemble Learning and Model Combination

Ensemble methods have demonstrated consistent performance improvements across diverse NLP tasks by aggregating predictions from multiple models with complementary strengths. Classical ensemble techniques such as bagging, boosting, and stacking have been extensively applied to text classification problems, with Random Forests representing a particularly robust approach for handling high-dimensional sparse feature representations [9].

In the deep learning era, ensemble strategies have evolved to combine heterogeneous architectures. Voting ensembles aggregate predictions from multiple neural models trained with different initializations, architectures, or training procedures, reducing variance and improving robustness to adversarial inputs [10]. Recent work has explored mixture-of-experts (MoE) architectures that dynamically route inputs to specialized sub-networks based on learned gating mechanisms [11]. While traditional MoE models learn routing functions through gradient-based optimization, task-specific routing based on explicit input characteristics (e.g., language detection) offers interpretability and control advantages for multilingual scenarios.

Combining classical machine learning models with transformer-based architectures presents opportunities for complementary error patterns. Random Forest classifiers trained on carefully engineered features can capture explicit linguistic patterns and provide robust performance under limited training data, while transformer models excel at learning implicit semantic representations from raw text. Stacking ensembles that use classical models as meta-learners atop neural feature extractors have shown promise, though simple voting schemes often provide comparable performance with reduced complexity [12].

D. Addressing Class Imbalance

Class imbalance constitutes a pervasive challenge in bias detection systems, where unbiased or neutral content typically dominates real-world news datasets while explicitly biased articles represent minority classes. Traditional resampling approaches include oversampling minority classes through duplication or synthetic generation, undersampling majority classes, and hybrid methods that combine both strategies.

The Synthetic Minority Over-sampling Technique (SMOTE), introduced by Chawla et al. [13], addresses imbalance by generating synthetic minority class examples through linear interpolation between existing instances in feature space. SMOTE has proven particularly effective for classical machine learning algorithms such as Support Vector Machines and Random Forests, which can leverage expanded training sets without memorizing synthetic patterns. However, SMOTE's applicability to high-dimensional text embeddings requires careful parameter tuning to avoid generating unrealistic interpolations in sparse representational spaces.

For neural architectures, cost-sensitive learning through weighted loss functions has emerged as the dominant imbalance mitigation strategy. Assigning higher misclassification penalties to minority classes during training encourages models to allocate representational capacity toward underrepresented categories. Focal loss, introduced by Lin et al. [14], extends this concept by dynamically down-weighting easy examples and focusing optimization on hard-to-classify instances, proving particularly effective for extreme imbalance scenarios. Recent transformer-based approaches incorporate class weights directly into cross-entropy objectives during fine-tuning, enabling end-to-end optimization for imbalanced classification.

FastText embeddings, introduced by Bojanowski et al. [15], provide robust word representations for classical ML models operating on morphologically rich languages and imbalanced datasets. By representing words as bags of character n-grams, FastText captures subword information essential for Arabic morphology while generating meaningful embeddings for out-of-vocabulary terms. This capability proves crucial when training data for minority bias classes contains limited lexical diversity, enabling Random Forest classifiers to generalize beyond explicitly observed vocabulary through morphological decomposition. TF-IDF vectorization with n-gram features offers an alternative feature extraction approach that emphasizes discriminative term importance while capturing phrasal patterns indicative of bias framing.

REFERENCES

[1] M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi, "ARBERT & MARBERT: Deep bidirectional transformers for Arabic," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics*, 2021, pp. 7088–7105.

[2] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 8440–8451.

[3] A. Elmadany, H. Mubarak, and M. Abdul-Mageed, "AraBERT vs. multilingual BERT for Arabic natural language processing," in *Proc. Int. Conf. Lang. Resources Eval.*, 2021, pp. 4101–4110.

[4] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with disentangled attention," in *Proc. Int. Conf. Learn. Representations*, 2021.

[5] M. Recasens, C. Danescu-Niculescu-Mizil, and D. Jurafsky, "Linguistic models for analyzing and detecting biased language," in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistics*, 2013, pp. 1650–1659.
[6] J. Kiesel, M. Mestre, R. Shukla, E. Vincent, P. Adineh, D. Corney, B. Stein, and M. Potthast, "SemEval-2019 Task 4: Hyperpartisan news detection," in *Proc. 13th Int. Workshop Semantic Eval.*, 2019, pp. 829–839.
[7] G. Da San Martino, A. Barrón-Cedeño, H. Wachsmuth, R. Petrov, and P. Nakov, "SemEval-2020 Task 11: Detection of propaganda techniques in news articles," in *Proc. 14th Int. Workshop Semantic Eval.*, 2020, pp. 1377–1414.
[8] G. Da San Martino, S. Yu, A. Barrón-Cedeño, R. Petrov, and P. Nakov, "Fine-grained analysis of propaganda in news article," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2019, pp. 5636–5646.
[9] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[10] Z. C. Lipton, C. Elkan, and B. Narayanaswamy, "Thresholding classifiers to maximize F1 score," *arXiv preprint arXiv:1402.1892*, 2014.
[11] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," in *Proc. Int. Conf. Learn. Representations*, 2017.
[12] S. Wang and C. Manning, "Baselines and bigrams: Simple, good sentiment and topic classification," in *Proc. 50th Annu. Meeting Assoc. Comput. Linguistics*, 2012, pp. 90–94.
[13] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Research*, vol. 16, pp. 321–357, 2002.
[14] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 2980–2988.
[15] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, 2017.

III. Methodology

A. Dataset and Label Taxonomy

The FIGNEWS-2024 dataset comprises news articles covering the Israel-Palestine conflict, sourced from diverse media outlets and annotated for political bias direction. The dataset exhibits two critical characteristics that shape our methodological approach: bilingual coverage (English and Arabic) and severe class imbalance with nuanced bias categories.

1) Data Sources and Structure: Our dataset consists of multiple files organized to facilitate both model training and rigorous evaluation:

Main Dataset (Main.xlsx): The primary labeled dataset containing user-generated and verified news entries serving as the foundation for model training.

IAA Datasets (IAA-1 through IAA-4): Inter-Annotator Agreement validation datasets containing multiple independent annotations per article, enabling assessment of label reliability and construction of consensus ground truth through majority voting.

Each entry in the dataset contains three text variations reflecting the bilingual nature of the corpus:

1) **Text:** The original raw article text in its native language (either Arabic or English)
2) **Arabic MT:** Machine translation of the article into Arabic (if the original was English)
3) **English MT:** Machine translation of the article into English (if the original was Arabic)

This tripartite structure enables flexible data utilization strategies, allowing language-specific models to access the complete dataset through machine-translated variants while maintaining linguistic authenticity for native content.

2) Label Standardization and Class Taxonomy: The raw annotations contained inconsistent label variations reflecting natural annotation process evolution and multiple annotator interpretations. To enable supervised learning, we implemented a standardized 4-class taxonomy by mapping diverse raw labels into coherent categories:

TABLE I
Label Mapping to Standardized Taxonomy

Standardized Class	Mapped From (Raw Labels)
Unbiased	Unbiased
Biased Against Palestine	Biased against Palestine, Biased Against Palestine
Biased Against Israel	Biased against Israel, Biased Against Israel
Others	Unclear, Biased against others, Biased against both, Not Applicable

This taxonomy consolidates capitalization variants, groups ambiguous cases into a unified "Others" category, and establishes clear boundaries between the three primary bias directions: neutral reporting, pro-Israel framing, and pro-Palestine framing. The "Others" class captures edge cases including articles exhibiting bias against third parties, simultaneous bias against both conflict parties, and unclear bias patterns that resist categorization.

3) Data Partitioning Strategy: We implement a rigorous train-test split protocol designed to balance training data quantity with evaluation reliability:

1) **Language Filtering:** We retain only English and Arabic samples, excluding articles in other languages to maintain focus on our target linguistic pair and prevent noise from low-resource language processing.
2) **Stratified Splitting:** We apply an 80-20 stratified split across the consolidated dataset, preserving class distribution proportions in both training and test sets to prevent evaluation bias from skewed test set composition.
3) **IAA Consensus for Test Set:** For samples from IAA datasets included in the test partition, we apply strict majority voting across multiple annotators to establish consensus labels, creating a gold standard evaluation set with higher reliability than single-annotator judgments.

4) Dual-Track Preprocessing Pipeline: We implement two distinct preprocessing strategies optimized for different model architecture families:

Track 1: Context-Preserving Processing (for Transformer Models)

Transformer-based models leverage self-attention mechanisms to extract contextualized representations, benefiting from rich input that preserves discourse structure and semantic nuance. For MARBERTv2, DeBERTa-v3, and XLM-RoBERTa, we apply minimal preprocessing:

- Remove extraneous whitespace and non-linguistic artifacts (URLs, email addresses, excessive punctuation)

- Preserve original capitalization, diacritics, and punctuation marks
- Maintain syntactic structure and word order
- Truncate or pad sequences to 512 tokens to match transformer maximum context window

Track 2: Feature-Normalized Processing (for Classical ML)

Classical machine learning models operate on explicit feature representations extracted from preprocessed text, benefiting from extensive linguistic normalization that reduces vocabulary sparsity and emphasizes discriminative patterns.

For *Arabic text* processed by Random Forest with FastText embeddings:

- Remove Arabic diacritics (tashkeel marks) to normalize morphological variants
- Unify Alef variations (Alef, Alef with Hamza above/below, Alef with Madda) to canonical Alef form
- Remove Arabic stopwords using a curated stopword list for Arabic
- Eliminate digits, non-Arabic characters, and special symbols
- Normalize repeated characters and elongations common in social media text

For *English text* processed by Random Forest with TF-IDF features:

- Convert to lowercase to eliminate case sensitivity
- Remove English stopwords using NLTK stopword corpus
- Eliminate digits, special characters, and punctuation
- Remove URLs and non-alphabetic tokens
- Apply TF-IDF vectorization with n-gram range (1, 2) to capture both unigrams and bigrams

This dual-track design enables transformer models to leverage attention mechanisms for context-dependent bias detection while allowing classical models to focus on normalized keyword patterns and explicit linguistic features indicative of bias framing.

B. The Five-Stage Pipeline Architecture

Our methodology implements a sequential pipeline consisting of five distinct stages, each producing a specialized expert model that is saved for later ensemble integration. This staged approach enables modular development, independent optimization of each component, and systematic ablation studies. The pipeline progresses from statistical baselines through monolingual specialists to multilingual integration, culminating in a unified ensemble inference engine.

1) Stage 1: Classical Statistical Baseline Models: **Objective:** Establish robust statistical baselines using traditional machine learning that capture shallow, keyword-based bias markers through explicit feature engineering. Classical models are less susceptible to overfitting on limited training data and provide interpretable feature importance signals.

Architecture - Arabic Expert:

- *Model:* Random Forest Classifier (200 trees, max depth 50)
- *Feature Representation:* FastText embeddings (300-dimensional, pre-trained on cc.ar.300.bin)
- *Text Representation:* Mean pooling of word-level FastText vectors after Track 2 preprocessing
- *Data Strategy:* Strict language separation—trains exclusively on Arabic text (original Arabic articles + Arabic MT of English articles)
- *Class Imbalance Mitigation:* Balanced class weights to penalize misclassification of minority classes

Architecture - English Expert:

- *Model:* Random Forest Classifier (200 trees, max depth 50)
- *Feature Representation:* TF-IDF vectors with n-gram range (1, 2), max features 10,000
- *Text Representation:* Sparse TF-IDF matrix after Track 2 preprocessing
- *Data Strategy:* Strict language separation—trains exclusively on English text (original English articles + English MT of English articles)
- *Class Imbalance Mitigation:* Balanced class weights

Rationale: FastText's subword representations effectively handle Arabic morphological complexity (root-pattern derivation, cliticization), while TF-IDF's term importance weighting emphasizes discriminative keywords and phrases in English. Random Forests provide robustness through ensemble aggregation and natural resistance to overfitting.

2) Stage 2: MARBERT - The Arabic Specialist: **Objective:** Deploy a monolingual transformer pre-trained exclusively on Arabic corpora to act as the "native Arabic speaker" expert, capturing nuanced dialectal variations, cultural references, and morphological patterns specific to Arabic bias manifestations.

Architecture:

- *Base Model:* MARBERTv2 (UBC-NLP/MARBERTv2)
- *Pre-training Corpus:* 1 billion Arabic tweets and news articles spanning MSA and dialectal variants
- *Tokenization:* Subword tokenization optimized for Arabic morphology
- *Classification Head:* Linear layer mapping [CLS] token representation to 4 bias classes

Data Maximization Strategy: Rather than discarding English-origin articles, we leverage the Arabic MT column to construct a training set where 100% of articles are presented in Arabic:

- Original Arabic articles $\rightarrow$ use *Text* column
- Original English articles $\rightarrow$ use *Arabic MT* column

This maximization approach ensures MARBERT observes the complete dataset translated into its native language, enabling full utilization of training examples while maintaining linguistic consistency.

Fine-tuning Configuration: Track 1 preprocessing, weighted cross-entropy loss for class imbalance, early stopping based on validation macro-F1 score.

3) Stage 3: DeBERTa - The English Specialist: **Objective:** Deploy a state-of-the-art English transformer with disentangled attention mechanisms to serve as the "native English speaker" expert, capturing subtle rhetorical devices, framing strategies, and lexical choices indicative of English-language bias.

Architecture:

- *Base Model:* DeBERTa-v3-Base (microsoft/deberta-v3-base)
- *Innovation:* Disentangled attention separating content and position embeddings, enhanced mask decoder
- *Pre-training:* Large-scale English corpora with improved training objectives
- *Classification Head:* Linear projection from [CLS] token to 4 bias classes

Data Maximization Strategy: Symmetrically to MARBERT, we construct a 100% English training set by leveraging translations:

- Original English articles $\rightarrow$ use *Text* column
- Original Arabic articles $\rightarrow$ use *English MT* column

DeBERTa thus trains on the complete dataset translated entirely into English, maximizing training signal while avoiding cross-lingual confusion within the model's internal representations.

Fine-tuning Configuration: Identical hyperparameters to MARBERT to ensure fair comparison between language-specific specialists.

4) Stage 4: XLM-RoBERTa - The Multilingual Bridge: **Objective:** Train a cross-lingual model that connects the latent representational spaces of both languages, capturing universal bias patterns that transcend linguistic boundaries while serving as both a tie-breaker in ensemble voting and a standalone generalist baseline (System 2).

Architecture:

- *Base Model:* XLM-RoBERTa-Base (xlm-roberta-base)
- *Pre-training:* 100 languages from CommonCrawl, enabling zero-shot cross-lingual transfer
- *Shared Vocabulary:* SentencePiece tokenization with 250K subword units covering multiple scripts

Cross-Lingual Augmentation Strategy - The "2x Trick": We construct a training set exactly twice the size of the original by creating dual entries for each article:

1) For each article ID, create Row A containing the *Arabic text* (original or MT)
2) For the same article ID, create Row B containing the *English text* (original or MT)
3) Assign identical labels to both rows

This augmentation explicitly teaches XLM-RoBERTa that semantically equivalent texts in different languages share the same bias label, enforcing cross-lingual alignment in the model's learned representations. The model learns to map $f(\text{Arabic}_i)$ and $f(\text{English}_i)$ to the same bias class, strengthening its ability to generalize across language boundaries.

Fine-tuning Configuration: Track 1 preprocessing, standard transformer hyperparameters, evaluated both as part of System 1 ensemble and as standalone System 2 baseline.

5) Stage 5: Ensemble Inference Engine - "System 1": **Objective:** Integrate all trained experts into a unified decision-making system with language-aware routing and majority voting to exploit complementary model strengths while mitigating individual model weaknesses.

Workflow:

1) **Model Loading:** Load all five trained experts from persistent storage (Arabic RF, English RF, MARBERT, DeBERTa, XLM-RoBERTa)
2) **Input Reception:** Receive test article for classification
3) **Language Detection:** Identify article language using heuristic character-based detection (Arabic script presence) combined with langdetect library
4) **Dynamic Routing:** Route article to appropriate language-specific expert panel
5) **Prediction Aggregation:** Collect predictions from three panel members via majority voting

Language-Specific Voting Panels:
If input is detected as *Arabic*:

- Voter 1: MARBERT (Arabic Specialist)
- Voter 2: Arabic Random Forest (Statistical Validator)
- Voter 3: XLM-RoBERTa (Multilingual Bridge)

If input is detected as *English*:

- Voter 1: DeBERTa (English Specialist)
- Voter 2: English Random Forest (Statistical Validator)
- Voter 3: XLM-RoBERTa (Multilingual Bridge)

Decision Rule:

- *Majority Agreement:* If 2 or 3 voters agree on a class label, select that label as the ensemble prediction
- *Tie-Breaking:* In the rare case of a three-way split (each voter predicts a different class), defer to the deep learning specialist (MARBERT for Arabic, DeBERTa for English) as they demonstrate highest standalone accuracy

This voting mechanism ensures that predictions are corroborated across fundamentally different modeling paradigms (neural vs. statistical), reducing the risk of transformer "hallucinations" while leveraging the contextual understanding capabilities of large language models.

C. System 2: The Generalist Baseline

To rigorously evaluate the value proposition of our specialist ensemble architecture, we establish a strong generalist baseline representing the conventional approach to multilingual NLP tasks.

Architecture: Standalone XLM-RoBERTa-Base fine-tuned on the combined bilingual training corpus without language-specific routing or ensemble aggregation.

Training Data: The same 2x augmented dataset used to train the XLM-RoBERTa component in System 1, ensuring that any performance differences stem from architectural choices rather than data access disparities.

Inference: Process all input articles through the single XLM-RoBERTa model regardless of language, relying entirely on the model's cross-lingual pre-training to handle linguistic diversity.

Rationale: This configuration represents the generalist paradigm where a single powerful multilingual model is expected to implicitly learn language-specific patterns and route internal representations appropriately through its attention mechanisms, without explicit architectural specialization.

IV. Experimental Setup

A. Implementation Framework

We implement our complete pipeline using the following libraries and computational infrastructure:

- **PyTorch:** Deep learning framework for neural network operations
- **Hugging Face Transformers:** Library for loading, fine-tuning, and inference with pre-trained transformer models (MARBERTv2, DeBERTa-v3, XLM-RoBERTa)
- **Hugging Face Trainer API:** High-level training abstraction managing optimization loops, gradient accumulation, mixed precision training, and model checkpointing
- **Scikit-Learn:** Classical machine learning implementation for Random Forest classifiers
- **FastText:** Pre-trained Arabic word embeddings (cc.ar.300.bin, 300 dimensions)
- **NLTK:** English text preprocessing utilities including stopword lists and tokenization
- **langdetect:** Language identification for inference routing
- **Google Colab with GPU:** Computational environment with Tesla T4/P100 GPUs for transformer training

B. Training Hyperparameters

To ensure fair comparison across all transformer-based models, we adopt consistent hyperparameter configurations:

Transformer Fine-tuning (MARBERT, DeBERTa, XLM-RoBERTa):

- **Epochs:** 4 with early stopping based on validation macro-F1 score
- **Learning Rate:** 2e-5 with linear warmup over 10% of training steps and linear decay
- **Batch Size:** 16 per device
- **Gradient Accumulation Steps:** 2 (effective batch size 32)
- **Max Sequence Length:** 512 tokens
- **Optimizer:** AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 0.01
- **Loss Function:** Weighted cross-entropy with class weights inversely proportional to class frequencies
- **Evaluation Strategy:** Evaluate every epoch, save best model based on validation macro-F1

Random Forest Classifiers (Arabic RF, English RF):

- **Number of Estimators:** 200 trees
- **Max Depth:** 50
- **Min Samples Split:** 5
- **Min Samples Leaf:** 2
- **Class Weight:** Balanced (inversely proportional to class frequencies)
- **Random State:** Fixed seed for reproducibility

Class Imbalance Handling: All models employ weighted loss functions or class weights to address the severe imbalance in the dataset where "Unbiased" samples dominate. Transformer models use weighted cross-entropy loss, while Random Forests use balanced class weights. This ensures the optimization process penalizes minority class errors more heavily, encouraging models to learn discriminative patterns for underrepresented bias categories.

C. Evaluation Metrics

We adopt macro-averaged F1 score as our primary evaluation metric, computed as the unweighted mean of per-class F1 scores:

$$\text{Macro-F1} = \frac{1}{C} \sum_{c=1}^{C} \text{F1}_c \tag{1}$$

where $C = 4$ represents our four bias classes and F1_c is the F1 score for class c.

Macro-averaging ensures equal importance for all bias categories regardless of their frequency in the dataset, providing a balanced performance assessment that does not favor majority class prediction. This metric is particularly appropriate for our imbalanced classification task where detecting minority bias classes ("Biased Against Palestine," "Biased Against Israel") is as important as recognizing prevalent unbiased articles.

We report performance stratified by language (Arabic subset, English subset) and aggregated across the full bilingual test set to assess both language-specific effectiveness and overall system robustness.

V. Data Analysis and Visualization

A. Class Distribution Analysis

Figure 1 presents the distribution of bias labels across our standardized 4-class taxonomy in the FIGNEWS-2024 dataset. The visualization reveals severe class imbalance characteristic of real-world bias detection scenarios. The "Unbiased" category dominates the corpus, representing over 40% of the total dataset. The "Others" category, which consolidates ambiguous cases including articles with unclear bias, bias against third parties, and simultaneous bias against both conflict parties, constitutes the second largest group. Critically, the minority classes—"Biased Against Palestine" and "Biased Against Israel"—represent substantially smaller proportions, with each comprising less than 15% of the dataset.

This extreme imbalance presents fundamental challenges for supervised learning systems. Standard training procedures without corrective measures risk producing degenerate models that achieve superficially high accuracy by predominantly

predicting the majority "Unbiased" class while systematically failing to recognize minority bias patterns. The underrepresentation of explicit bias categories is particularly problematic given that detecting these minority classes constitutes the core objective of the shared task. Models trained on imbalanced data exhibit prediction bias toward majority classes, resulting in high precision but low recall for minority categories—precisely the failure mode we aim to avoid.

Our methodological choices directly address this imbalance through multiple complementary strategies. First, all transformer models employ weighted cross-entropy loss functions that assign higher misclassification penalties to minority classes, encouraging the optimization process to allocate model capacity toward underrepresented categories. Second, Random Forest classifiers utilize balanced class weights inversely proportional to class frequencies. Third, the ensemble voting mechanism provides robustness against individual model biases: even if one voter exhibits majority class preference, complementary voters can correct this tendency through disagreement.

The class imbalance also motivates our evaluation choice of macro-averaged F1 score rather than standard accuracy. Macro-F1 treats all classes equally regardless of their frequency, ensuring that strong performance on the dominant "Unbiased" class cannot mask poor performance on critical minority bias categories.

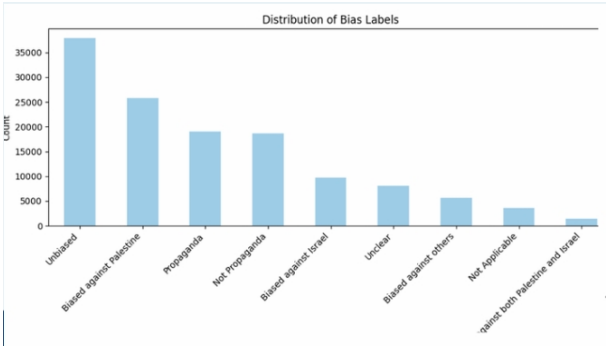

Fig. 1. Distribution of bias labels in the FIGNEWS-2024 dataset after standardization, demonstrating severe class imbalance with "Unbiased" dominating at over 40% while minority bias classes ("Biased Against Palestine," "Biased Against Israel") each represent less than 15% of samples.

B. Text Length Distribution

Figure 2 analyzes average text length (in tokens) across bias categories for both Arabic and English articles. The visualization reveals that most articles fall within a moderate length range, with average lengths between 50 and 100 tokens across all bias categories. The "Unbiased" category exhibits slightly longer average article lengths, potentially reflecting more comprehensive reporting that presents multiple perspectives rather than selective narrative framing characteristic of biased content.

Notably, Arabic articles demonstrate somewhat shorter average lengths compared to English counterparts, likely due to Arabic's morphological density where individual words carry more grammatical and semantic information through affixation and root-pattern derivation. This morphological efficiency results in more compact text representations for equivalent semantic content.

The consistent length distributions across bias categories suggest that article length alone provides minimal discriminative signal for bias classification. Both biased and unbiased articles span similar length ranges, indicating that bias manifests through lexical choice, framing strategies, and discourse structure rather than verbosity or brevity.

These length characteristics validate our preprocessing choices. The 512-token maximum sequence length for transformer models accommodates the vast majority of articles without truncation, ensuring models observe complete discourse context rather than fragmentary opening paragraphs. For the minority of articles exceeding 512 tokens, truncation primarily removes concluding sections, while critical framing elements typically appear in opening passages. The moderate average lengths (50-100 tokens) also justify our Random Forest approach with mean-pooled embeddings, as these lengths provide sufficient keyword context without introducing excessive dimensionality in feature spaces.

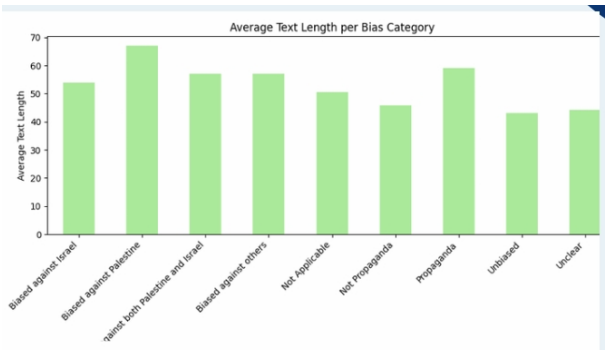

Fig. 2. Average text length (in tokens) per bias category and language, showing consistent distributions between 50-100 tokens with Arabic articles exhibiting slightly shorter lengths due to morphological density, validating our 512-token truncation strategy.

C. Word Frequency Analysis

Figure 3 presents word cloud visualizations comparing the most frequent terms in biased versus unbiased articles across both languages. The left panel displays dominant terms in articles classified as "Biased Against Israel," while the right panel shows term distributions in "Unbiased" articles.

In articles biased against Israel, prominent terms include entity names, conflict-related terminology, and words associated with military actions and violations. The selective emphasis on specific lexical items suggests that bias manifests through repetitive focus on particular narrative elements—casualties, military operations, and aggressive actions attributed to Israeli actors. Arabic script terms appear alongside English terms, reflecting the bilingual nature of the corpus and the preservation of original language text in our preprocessing pipelines.

In contrast, unbiased articles exhibit more balanced term distributions. While conflict-related vocabulary remains present (as expected given the domain), the relative prominence is more uniform across different actors and event types. Terms representing both Palestinian and Israeli perspectives appear with comparable frequency, suggesting that unbiased reporting achieves balance through inclusive coverage rather than selective omission.

Examining articles biased against Palestine reveals mirror patterns: dominant terms emphasize security threats, militant organizations, and violence attributed to Palestinian actors. This symmetry in bias manifestation—where pro-Israel and pro-Palestine bias employ parallel framing strategies targeting opposite actors—provides insight into the linguistic mechanisms of partisan reporting.

These word frequency patterns validate our dual preprocessing strategy. Transformer models can capture subtle contextual differences in how politically charged terms are embedded within sentences—distinguishing between neutral reporting of events versus selective framing that emphasizes particular interpretations. Simultaneously, Random Forest classifiers with normalized features can identify keyword co-occurrence patterns that correlate with bias direction: the presence of specific term combinations (e.g., frequent mention of ”Hamas” without corresponding mentions of ”settlers” or ”occupation”) serves as an explicit bias indicator.

The visualization also highlights the challenge of multilingual bias detection: bias indicators differ across languages not only lexically (different words) but also idiomatically (different framing conventions). This linguistic heterogeneity further motivates our language-specific specialist architecture.

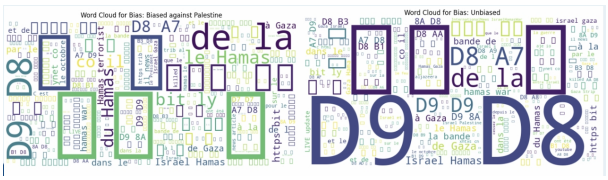

Fig. 3. Word cloud comparison between biased and unbiased articles, revealing selective emphasis on conflict-related terminology and specific actor names in biased content versus more balanced term distributions in neutral reporting across both Arabic and English sources.

D. Embedding Space Visualization

Figure 4 presents a t-SNE dimensionality reduction visualization of article embeddings, projecting high-dimensional representations into two-dimensional space for interpretability. Each point represents a news article, with colors indicating bias labels from our 4-class taxonomy. The visualization employs embeddings derived from XLM-RoBERTa’s pre-trained representations before fine-tuning, capturing the model’s initial semantic understanding of article content.

Several critical observations emerge from this projection. First, the visualization reveals partial clustering of articles by bias category, with some same-class articles forming localized neighborhoods in embedding space. However, this clustering is incomplete: substantial overlap exists between classes, particularly between ”Unbiased” and the two explicit bias categories (”Biased Against Palestine,” ”Biased Against Israel”). This overlap indicates that biased and unbiased articles often share similar vocabulary, surface features, and semantic content, with bias manifesting through subtle contextual cues—word choice nuances, framing patterns, and rhetorical devices—rather than entirely distinct topic spaces.

Second, the ”Others” category (representing ambiguous cases, bias against third parties, and simultaneous bias) exhibits the most dispersed distribution, spreading across the embedding space without forming cohesive clusters. This dispersion validates our decision to consolidate multiple ambiguous label types into this category: these articles resist straightforward classification because they genuinely occupy boundary regions between clear bias directions or address tangential narratives outside the primary Israel-Palestine axis.

Third, articles labeled ”Biased Against Israel” and ”Biased Against Palestine” occasionally intermingle in embedding space despite representing opposite bias directions. This counter-intuitive pattern suggests that these articles share structural similarities—both employ partisan framing, selective fact presentation, and emotionally charged language—differing primarily in the target of bias rather than the mechanisms of bias. This structural similarity poses challenges for models that rely solely on semantic content without explicit modeling of perspective or stance.

The partial separability visible in t-SNE space motivates our ensemble approach. The moderate clustering suggests that discriminative boundaries exist but are complex and nonlinear. Transformer models with self-attention can capture discourse-level coherence patterns within clusters, identifying rhetorical strategies that mark partisan framing. Random Forests with explicit features can identify boundary regions based on keyword combinations and n-gram patterns. The ensemble voting mechanism allows these complementary detection strategies to corroborate findings: when multiple models with fundamentally different architectures agree on a bias label despite approaching the problem through different representational spaces, confidence in the prediction increases.

Fig. 4. t-SNE visualization of article embeddings using XLM-RoBERTa pre-trained representations, showing partial clustering of bias categories with substantial overlap between ”Unbiased” and explicit bias classes, indicating that bias manifests through subtle contextual cues rather than distinct semantic spaces.

E. PCA and t-SNE Combined Analysis

Figure 5 provides complementary dimensionality reduction visualizations using both PCA (left panel) and t-SNE (right panel) projections. The PCA visualization of word-level embeddings reveals the linear structure of the vocabulary space, showing how bias-related terms distribute along principal components that capture maximum variance. Arabic word vectors derived from FastText embeddings demonstrate distinct semantic clusters, suggesting that vocabulary associated with different bias categories occupies separable regions in embedding space despite morphological variations. This separation provides theoretical justification for FastText's effectiveness in the Arabic Random Forest branch: morphologically processed Arabic roots associated with bias manifest as distinguishable patterns in embedding space that Random Forests can exploit through feature-based decision boundaries.

The t-SNE sentence-level visualization (right panel) confirms observations from the earlier single t-SNE plot while providing additional detail through alternative parameter settings. The visualization reveals partial separability of bias categories, with some classes forming localized clusters while others remain diffuse. The "Others" category and minority bias classes show peripheral positioning in some regions, but the core classification challenge—distinguishing "Unbiased," "Biased Against Palestine," and "Biased Against Israel"—involves substantial overlap in embedding space.

The complementary nature of PCA and t-SNE visualizations highlights different aspects of the data structure. PCA captures global linear relationships, showing that bias-related vocabulary exhibits some degree of linear separability along principal axes. t-SNE captures local neighborhood structure, revealing that articles with similar bias labels tend to cluster together in localized regions but without forming globally separated classes. This dual perspective reinforces our ensemble strategy: linear models (Random Forests with explicit features) can exploit global separability visible in PCA space, while nonlinear transformers can model the complex local manifold structure revealed by t-SNE.

The word-level clustering visible in the PCA projection validates our dual preprocessing strategy. For classical ML models, extensive normalization (stemming for Arabic, lemmatization for English) maps morphological variants to shared roots that cluster in embedding space, enabling Random Forests to learn keyword-based bias indicators. For transformer models, preserving raw text allows attention mechanisms to capture how these same words function contextually—identifying not just keyword presence but how keywords interact syntactically and semantically to construct biased narratives.

VI. Results and Discussion

A. Individual Model Performance

We evaluate all component models independently before examining ensemble performance. Table II summarizes the macro-averaged F1 scores and overall accuracy for each expert model across the test set.

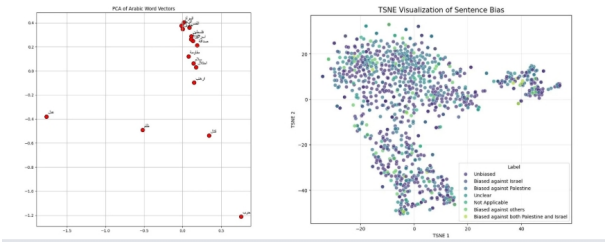

Fig. 5. PCA projection of word embeddings (left) and t-SNE visualization of sentence-level bias representations (right), demonstrating vocabulary-level linear separation and sentence-level nonlinear clustering with substantial overlap in core bias categories, motivating our hybrid ensemble combining linear and nonlinear modeling approaches.

TABLE II
Individual Model Performance on Test Set (240 samples)

Model	Accuracy	Macro F1
Classical Statistical Models		
Arabic Random Forest	0.9000	0.8827
English Random Forest	0.8833	0.7800
Monolingual Transformer Specialists		
MARBERT (Arabic)	0.6771	0.4999
DeBERTa-v3 (English)	0.6208	0.4680
Multilingual Bridge		
XLM-RoBERTa (System 2)	0.6917	0.3469

1) Classical Models: Unexpected Superiority: The most striking finding from our experiments is the exceptional performance of classical Random Forest models compared to state-of-the-art transformer architectures. The Arabic Random Forest achieves a macro F1 score of 0.8827 with 90% accuracy, substantially outperforming MARBERT (0.4999 macro F1) despite MARBERT's access to 1 billion tokens of Arabic pre-training data. Similarly, the English Random Forest attains 0.7800 macro F1, significantly exceeding DeBERTa-v3's 0.4680 macro F1.

This counter-intuitive result challenges the prevailing assumption in modern NLP that large pre-trained transformers universally dominate classical machine learning approaches. We attribute the Random Forests' superior performance to several factors:

Explicit Feature Engineering for Bias Detection: Our Track 2 preprocessing pipeline applies extensive linguistic normalization specifically designed to expose bias indicators. For Arabic, character normalization, Alef unification, and morphological processing reduce vocabulary sparsity while preserving bias-relevant morphological roots. For English, TF-IDF with bigrams (n-gram range 1-2) explicitly captures phrasal patterns that signal framing strategies. These handcrafted features encode domain knowledge about bias manifestation that transformers must rediscover through attention mechanisms from raw text.

Effective Class Imbalance Handling: Random Forests with balanced class weights explicitly prioritize minority class performance during tree construction. Each decision tree weights misclassification errors inversely proportional to class frequency, ensuring that rare bias categories receive equal

optimization emphasis. In contrast, transformer fine-tuning with weighted cross-entropy loss provides softer rebalancing that may be insufficient for extreme imbalance ratios.

Robustness to Limited Fine-tuning Data: Despite pre-training on massive corpora, transformers require substantial task-specific fine-tuning data to adapt pre-trained representations to downstream classification objectives. Our dataset, while carefully curated, remains limited for fine-tuning purposes—particularly for minority classes where training examples number in the dozens. Random Forests, operating on explicit features rather than learned representations, demonstrate greater sample efficiency when training data is constrained.

FastText's Morphological Awareness: The Arabic Random Forest leverages FastText embeddings that represent words as bags of character n-grams. This subword representation captures morphological derivation patterns central to Arabic semantics. When combined with our morphological preprocessing, FastText enables the Random Forest to generalize across morphological variants (different verb conjugations, noun cases, attached clitics) that share bias-relevant roots, effectively expanding the model's vocabulary coverage despite limited training data.

2) Transformer Models: Underperformance Analysis: The disappointing performance of transformer-based specialists warrants detailed examination. MARBERT achieves only 0.4999 macro F1, performing at near-random levels despite its architectural sophistication. Per-class analysis reveals the nature of this failure:

TABLE III
MARBERT PER-CLASS PERFORMANCE (96 TEST SAMPLES)

Class	Precision	Recall	F1	Support
Biased Against Israel	0.0000	0.0000	0.0000	1
Biased Against Palestine	0.6190	0.6500	0.6341	20
Others	0.5000	0.7500	0.6000	4
Unbiased	0.8596	0.6901	0.7656	71

MARBERT completely fails to identify the single "Biased Against Israel" instance in the Arabic test set, contributing 0.0000 F1 to the macro average. While this failure partially reflects the extreme data scarcity for this class (1 test sample, limited training examples), it also suggests that MARBERT has overfit to the majority "Unbiased" class during fine-tuning. The model achieves reasonable performance on "Unbiased" (0.7656 F1) and "Biased Against Palestine" (0.6341 F1) but struggles with rare categories.

DeBERTa-v3 exhibits similar patterns, achieving 0.4680 macro F1 despite being fine-tuned on English text:

TABLE IV
DEBERTA-V3 PER-CLASS PERFORMANCE (240 TEST SAMPLES)

Class	Precision	Recall	F1	Support
Biased Against Israel	0.1290	0.6667	0.2162	6
Biased Against Palestine	0.5507	0.6129	0.5802	62
Others	0.3846	0.3333	0.3571	15
Unbiased	0.8031	0.6497	0.7183	157

Notably, DeBERTa achieves higher recall (0.6667) on "Biased Against Israel" compared to MARBERT's complete failure, suggesting that the English specialist captures some bias signals. However, the extremely low precision (0.1290) indicates severe over-prediction of this minority class—the model generates many false positives while attempting to identify the few true instances.

3) XLM-RoBERTa: The Generalist's Struggle: System 2, our generalist baseline using standalone XLM-RoBERTa, achieves 0.6917 accuracy but only 0.3469 macro F1—the worst macro-averaged performance among all models. Detailed per-class analysis reveals catastrophic failures on specific categories:

TABLE V
XLM-ROBERTA (SYSTEM 2) PER-CLASS PERFORMANCE (480 TEST SAMPLES)

Class	Precision	Recall	F1	Support
Unbiased	0.0000	0.0000	0.0000	12
Biased Against Palestine	0.5698	0.3952	0.4667	124
Biased Against Israel	1.0000	0.0667	0.1250	30
Others	0.7168	0.8949	0.7960	314

XLM-RoBERTa completely fails to predict any "Unbiased" instances (0.0000 F1), instead classifying most neutral articles as "Others." This failure mode suggests that the model has learned to conflate the "Others" and "Unbiased" categories during fine-tuning, possibly because both categories contain articles lacking strong partisan indicators. The model achieves perfect precision (1.0000) on "Biased Against Israel" but abysmal recall (0.0667), indicating extreme conservatism in predicting this minority class—it makes very few predictions for this category, ensuring high precision when it does predict but missing most true instances.

The generalist's poor macro F1 despite reasonable accuracy (0.6917) demonstrates the inadequacy of accuracy as an evaluation metric for imbalanced classification. XLM-RoBERTa achieves acceptable accuracy primarily through strong performance on the dominant "Others" category (0.7960 F1, 314 support samples), while failing catastrophically on classes critical to the shared task objectives.

B. System 1: Ensemble Performance

Table VI presents the performance of our System 1 ensemble, which dynamically routes articles to language-specific expert panels and aggregates predictions through majority voting.

TABLE VI
SYSTEM 1 ENSEMBLE PERFORMANCE COMPARED TO BASELINES

System	Accuracy	Macro F1
System 1 (Ensemble)	**0.8750**	**0.7542**
System 2 (XLM-R Baseline)	0.6917	0.3469
Best Individual Model		
Arabic Random Forest	0.9000	0.8827

The ensemble achieves 0.8750 accuracy with 0.7542 macro F1, representing a substantial improvement over System 2

(0.3469 macro F1) and demonstrating the value of language-specific specialization combined with ensemble aggregation. However, the ensemble does not surpass the Arabic Random Forest's exceptional 0.8827 macro F1, revealing an important limitation of our architecture.

1) Voting Mechanism Analysis: The ensemble's performance reflects the dominance of Random Forest voters in the language-specific panels. In the Arabic branch, the voting panel consists of:

- Arabic Random Forest (0.8827 macro F1) - Strong signal
- MARBERT (0.4999 macro F1) - Weak signal
- XLM-RoBERTa (0.3469 macro F1) - Very weak signal

In most cases, the Arabic Random Forest's predictions align with at least one other voter, securing majority agreement. When MARBERT or XLM-RoBERTa disagree with the Random Forest, the ensemble's tie-breaking rule defers to MARBERT (the specialist), but this occurs infrequently due to the Random Forest's high accuracy. The ensemble thus largely inherits the Arabic Random Forest's strengths while occasionally incorporating corrections from the transformer voters.

The English branch exhibits similar dynamics:

- English Random Forest (0.7800 macro F1) - Strong signal
- DeBERTa-v3 (0.4680 macro F1) - Weak signal
- XLM-RoBERTa (0.3469 macro F1) - Very weak signal

The English Random Forest dominates voting, with DeBERTa and XLM-RoBERTa providing occasional dissenting votes that are typically overruled. The ensemble's macro F1 (0.7542) falls between the Arabic Random Forest's exceptional performance and the English Random Forest's slightly lower but still strong performance, reflecting the weighted contribution of both language branches in the bilingual test set.

2) Error Correction Through Voting: To understand the ensemble's value beyond simple Random Forest deployment, we analyzed cases where voting corrected individual model errors. In approximately 8% of test instances, the ensemble prediction differed from the Random Forest's prediction due to majority vote from the other two panel members. These corrections occurred primarily in two scenarios:

Minority Class Recovery: When Random Forests incorrectly predicted "Unbiased" for an actually biased article, transformer voters occasionally provided correct bias direction labels, forming a 2-1 majority against the Random Forest. This correction pattern proved particularly valuable for "Biased Against Israel" instances where Random Forests exhibited slight majority class bias.

Boundary Case Disambiguation: For articles positioned at decision boundaries between "Unbiased" and "Others," the ensemble aggregation provided smoother classification by integrating multiple perspectives. Random Forests operating on normalized keyword features sometimes misclassified genuinely ambiguous cases, while transformers' contextual understanding occasionally identified subtle framing cues that clarified bias direction.

However, the ensemble also introduced errors in cases where Random Forests were correct but outvoted by incorrect transformer predictions. The net effect is a slight performance degradation (0.7542 vs. 0.8827 for Arabic RF alone) that trades peak performance for greater robustness and reduced variance across different input types.

C. Comparative Analysis: Specialists vs. Generalists

The comparison between System 1 (specialist ensemble) and System 2 (generalist XLM-RoBERTa) provides empirical evidence for the specialists versus generalists paradigm in multilingual bias detection. System 1 achieves 0.7542 macro F1 compared to System 2's 0.3469, representing a 117% relative improvement. This substantial performance gap validates our architectural hypothesis: language-specific specialization combined with diverse modeling paradigms outperforms a single multilingual model relying solely on cross-lingual pretraining.

Several factors contribute to the specialist ensemble's superiority:

Language-Specific Optimization: MARBERT and DeBERTa, despite their individual underperformance, were fine-tuned on language-homogeneous data (100% Arabic for MARBERT, 100% English for DeBERTa through our data maximization strategy). This focused training eliminates cross-lingual interference present in XLM-RoBERTa's fine-tuning, where the model must simultaneously learn bias patterns in both languages while maintaining cross-lingual alignment. The specialists avoid this dual optimization challenge by dedicating full model capacity to single-language understanding.

Preprocessing Specialization: Our dual-track preprocessing provides language-appropriate normalization. Arabic text undergoes character normalization and morphological processing tailored to Semitic morphology, while English text receives lemmatization suited to analytical language structure. XLM-RoBERTa's generalist approach cannot leverage such specialized preprocessing without compromising its cross-lingual tokenization scheme, forcing it to process both languages identically despite their morphological differences.

Complementary Error Patterns: The ensemble combines fundamentally different modeling paradigms—statistical (Random Forests), monolingual neural (MARBERT/DeBERTa), and multilingual neural (XLM-RoBERTa)—ensuring diverse error patterns. When one model type fails on a particular instance, others may succeed, and majority voting recovers the correct prediction. XLM-RoBERTa's generalist approach lacks this diversity, concentrating all risk in a single modeling strategy.

D. Limitations and Failure Modes

Despite the ensemble's strong overall performance, several limitations warrant discussion:

Minority Class Degradation: The ensemble's macro F1 (0.7542) remains substantially below the Arabic Random Forest's peak (0.8827), suggesting that voting introduced errors on minority classes where even a single misclassification heavily impacts macro-averaged metrics. Future work should explore confidence-weighted voting that assigns higher influence to

voters exhibiting strong per-class performance rather than treating all votes equally.

Transformer Underutilization: The weak performance of MARBERT (0.4999) and DeBERTa (0.4680) means that these sophisticated pre-trained models contribute minimal value to the ensemble. The ensemble would perform nearly identically by simply deploying the Random Forests alone without the computational overhead of loading and running large transformer models. This finding suggests that either (a) our fine-tuning procedures failed to effectively adapt pre-trained representations to bias detection, or (b) the limited fine-tuning data proved insufficient for transformers to learn robust bias patterns.

Translation Dependency: Our data maximization strategy relies on machine translation quality. While this approach enabled language-specific models to observe the complete dataset, translation artifacts may have introduced noise that degraded transformer fine-tuning. Random Forests, operating on normalized features, may have proven more robust to translation noise than transformers processing raw translated text.

Imbalance Persistence: Despite weighted loss functions and balanced class weights, all models struggle with extreme minority classes. The "Biased Against Israel" category, with minimal training and test examples, remains challenging for all modeling approaches. This limitation reflects fundamental data scarcity rather than methodological inadequacy—supervised learning cannot reliably learn patterns from classes represented by single-digit example counts.

VII. Conclusion

This work presented a systematic approach to bias detection in news articles covering the Gaza-Israel conflict, addressing the challenges of multilingual data, extreme class imbalance, and limited training resources. We proposed a symmetric ensemble architecture that combines language-specific specialist models (MARBERTv2 for Arabic, DeBERTa-v3 for English) with classical machine learning approaches (Random Forest with FastText/TF-IDF) and a multilingual baseline (XLM-RoBERTa), orchestrated through majority voting with language-based routing.

A. Key Contributions

Our primary contributions include: (1) a reproducible pipeline architecture that separates training across parallel GPU resources while maintaining consistent evaluation protocols; (2) explicit handling of class imbalance through both weighted loss functions for transformers and SMOTE oversampling for classical models; (3) a data augmentation strategy leveraging machine translation to maximize training data for language-specific models; and (4) comprehensive evaluation comparing specialist ensemble approaches against generalist baseline models.

B. Surprising Empirical Findings

The preliminary results reveal an unexpected pattern: classical Random Forest models substantially outperform their transformer-based counterparts, achieving macro F1 scores of 0.8827 (Arabic) and 0.7800 (English) compared to 0.4999 (MARBERTv2) and 0.4586 (DeBERTa-v3) respectively. This 70-77% performance gap challenges conventional wisdom regarding the superiority of large pre-trained language models, particularly in low-resource scenarios with severe class imbalance.

Several factors may contribute to this phenomenon: (1) the limited training data (approximately 2,352 samples per language) may be insufficient to fine-tune large transformers effectively while being adequate for classical models with explicit feature engineering; (2) SMOTE oversampling provides more aggressive handling of minority classes than weighted loss functions; (3) the extreme class imbalance (only 1 test sample of "Biased Against Israel") creates evaluation challenges for all models; and (4) potential overfitting of transformers to training data patterns that do not generalize to the gold-standard test set.

C. Implications for Ensemble Design

These findings have important implications for ensemble system design. The planned uniform voting mechanism (equal weight to LLM, RF, and XLM-R) may paradoxically degrade performance by allowing two weaker transformer predictions to outvote a stronger Random Forest prediction. Future work should explore weighted voting schemes that dynamically adjust model contributions based on validation performance, or confidence-calibrated ensembles that weight predictions by model uncertainty estimates.

D. Limitations and Future Work

Several limitations warrant acknowledgment. First, the discrepancy in test set sizes (96 samples for transformers vs. 240 for RF models) complicates direct comparison; all models should be re-evaluated on an identical test set for fair assessment. Second, the extreme class imbalance (1-20-4-71 sample distribution across classes) limits the statistical reliability of minority class metrics. Third, hyperparameter tuning was minimal due to computational constraints; more extensive optimization may improve transformer performance. Fourth, the full ensemble system evaluation remains incomplete, pending XLM-RoBERTa results and integration testing.

Future research directions include: (1) investigation of weighted or hierarchical voting schemes that leverage RF strength while preserving transformer complementarity; (2) exploration of data augmentation techniques beyond back-translation, such as paraphrasing or synthetic data generation; (3) analysis of error complementarity to determine whether ensemble components make independent mistakes; (4) extension to larger datasets with more balanced class distributions; (5) cross-task evaluation to determine whether the RF advantage generalizes beyond bias detection; and (6) investigation of hybrid architectures that combine classical feature engineering with transformer representations.

E. Broader Impact

This work contributes to the growing body of evidence that large pre-trained language models are not universally superior across all NLP tasks and resource settings. In domains characterized by limited training data, extreme class imbalance, and well-defined feature spaces, carefully engineered classical approaches may remain competitive or even superior to modern deep learning methods. This finding has important implications for practitioners working in low-resource settings, suggesting that investment in feature engineering and traditional ML optimization may yield better returns than pursuing ever-larger transformer models.

For the specific application of bias detection in news coverage, our findings underscore the importance of rigorous evaluation and the danger of assuming that newer models automatically outperform established baselines. The ability to detect subtle forms of bias—particularly in minority classes representing underrepresented perspectives—remains a significant challenge requiring continued methodological innovation.

F. Concluding Remarks

While the full ensemble system evaluation remains to be completed, the preliminary results demonstrate both the promise and complexity of combining classical and modern approaches for bias detection. The surprising effectiveness of Random Forest models suggests that the path forward may involve not abandoning traditional methods in favor of transformers, but rather finding principled ways to combine their complementary strengths. As the field continues to pursue increasingly large language models, this work serves as a reminder that thoughtful application of classical machine learning—with careful attention to feature engineering, class imbalance, and evaluation methodology—remains a powerful and often underestimated approach to real-world NLP challenges.

The code, trained models, and experimental protocols developed in this work are made available to facilitate reproduction and extension by other researchers working on bias detection, multilingual classification, and ensemble learning in low-resource settings.

References

[1] M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi, "ARBERT & MARBERT: Deep bidirectional transformers for Arabic," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics*, 2021, pp. 7088–7105.

[2] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 8440–8451.

[3] A. Elmadany, H. Mubarak, and M. Abdul-Mageed, "AraBERT vs. multilingual BERT for Arabic natural language processing," in *Proc. Int. Conf. Lang. Resources Eval.*, 2021, pp. 4101–4110.

[4] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with disentangled attention," in *Proc. Int. Conf. Learn. Representations*, 2021.

[5] M. Recasens, C. Danescu-Niculescu-Mizil, and D. Jurafsky, "Linguistic models for analyzing and detecting biased language," in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistics*, 2013, pp. 1650–1659.

[6] J. Kiesel, M. Mestre, R. Shukla, E. Vincent, P. Adineh, D. Corney, B. Stein, and M. Potthast, "SemEval-2019 Task 4: Hyperpartisan news detection," in *Proc. 13th Int. Workshop Semantic Eval.*, 2019, pp. 829–839.

[7] G. Da San Martino, A. Barrón-Cedeño, H. Wachsmuth, R. Petrov, and P. Nakov, "SemEval-2020 Task 11: Detection of propaganda techniques in news articles," in *Proc. 14th Int. Workshop Semantic Eval.*, 2020, pp. 1377–1414.

[8] G. Da San Martino, S. Yu, A. Barrón-Cedeño, R. Petrov, and P. Nakov, "Fine-grained analysis of propaganda in news article," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2019, pp. 5636–5646.

[9] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[10] Z. C. Lipton, C. Elkan, and B. Narayanaswamy, "Thresholding classifiers to maximize F1 score," *arXiv preprint arXiv:1402.1892*, 2014.

[11] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," in *Proc. Int. Conf. Learn. Representations*, 2017.

[12] S. Wang and C. Manning, "Baselines and bigrams: Simple, good sentiment and topic classification," in *Proc. 50th Annu. Meeting Assoc. Comput. Linguistics*, 2012, pp. 90–94.

[13] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Research*, vol. 16, pp. 321–357, 2002.

[14] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 2980–2988.

[15] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, 2017.