FIGNEWS-2024: Detecting Bias in Israel-Palestine Conflict News

A Multilingual Mixture of Experts Approach

Amr Hossam	221000832
Hossam Nasr	221000770
Fares Ahmed	221000570
Shorouk Sherif	221000645
Amenah Medhat	221001792
Mohamed Nashaat	221001565

The Problem

Detecting Bias in Conflict News Coverage

- News articles about Israel-Palestine conflict often contain political bias
- Articles published in Arabic and English (morphologically different languages)
- Manual detection is subjective and time-consuming

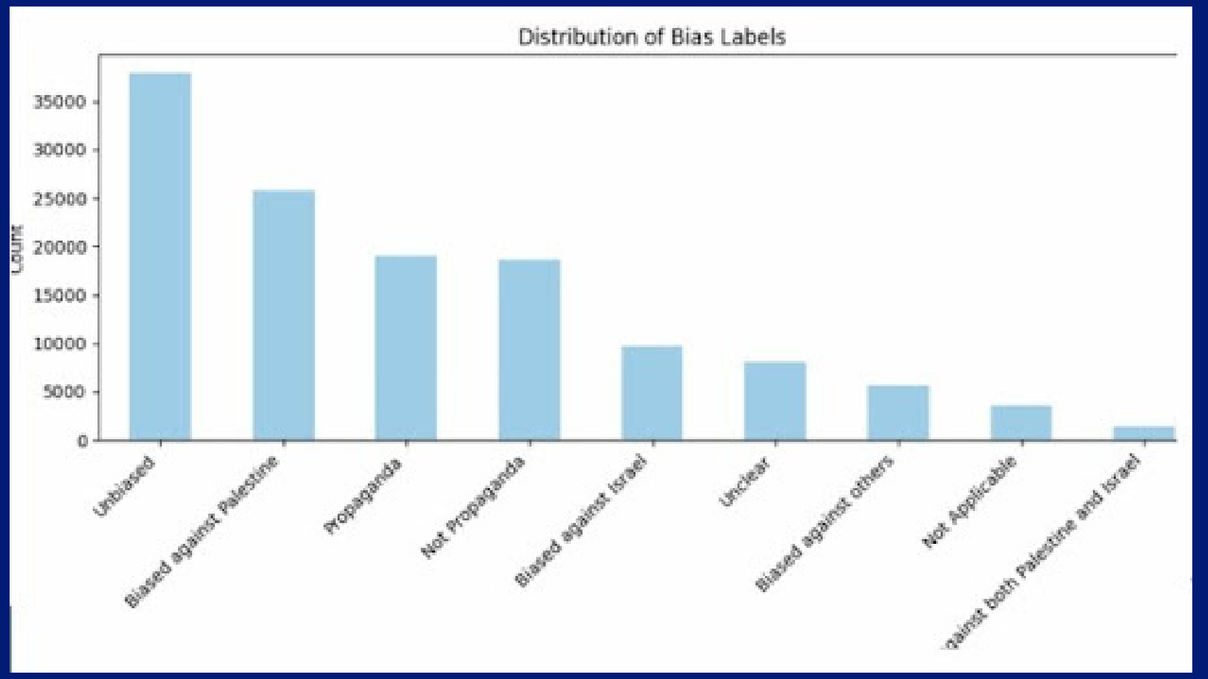
Our Goal: Automatic Classification

- Unbiased
- Biased Against Palestine
- Biased Against Israel
- Others (unclear/mixed bias)

Challanges

Three Major Challenges:

1. Linguistic Diversity • Arabic: complex morphology, root-pattern system • English: analytical structure, minimal inflection
2. Severe Class Imbalance • Unbiased articles: >40% of dataset • Biased articles: <15% each
3. Subtle Bias Manifestations • Not just keywords, but framing and context • Cultural and contextual nuances



Our Approach - Two Systems

System 1: "Mixture of Experts" Ensemble

- Language-specific specialists (Arabic + English experts)
- 5 different AI models working as a team
- Democratic voting mechanism
- Each expert trained on what it does best

System 2: Generalist Baseline

- Single multilingual model (XLM-RoBERTa)
- Handles both languages simultaneously
- Relies on cross-lingual pre-training

Research Question: Specialists vs. Generalist?

The 5-Stage Pipeline

Sequential Expert Construction

Stage 1: Classical Baselines

- Random Forest + FastText (Arabic)
- Random Forest + TF-IDF (English)

Stage 2: MARBERT Fine-tuning

- Arabic specialist trained on 1B Arabic tweets

Stage 3: DeBERTa Fine-tuning

English specialist with disentangled attention

Stage 4: XLM-RoBERTa Fine-tuning

Multilingual bridge connecting both languages

Stage 5: Ensemble Integration

Load all experts → Route → Vote

Data Maximization Strategy

Problem: Language-specific models miss data in other languages

Our Solution: Use Machine Translation

- **Arabic Expert sees 100% of dataset:**
  - Original Arabic articles → use as-is
  - Original English articles → use Arabic MT

- **English Expert sees 100% of dataset:**
  - Original English articles → use as-is
  - Original Arabic articles → use English MT

- **Result: No specialist misses training examples!**

"2x Augmentation" for XLM-RoBERTa:

- **Pair original + translation with same label**
- **Dataset size doubles → teaches label invariance**

Results - The Surprise!

Performance Comparison (Macro F1 Score)

Model	Macro F1	Winner
Arabic Random Forest	**0.88**	🏆 Best
English Random Forest	0.78	⭐ Strong
System 1 (Ensemble)	**0.75**	✅ Good
MARBERT (Arabic)	0.50	⚠️ Weak
DeBERTa (English)	0.47	⚠️ Weak
System 2 (XLM-R)	**0.35**	❌ Poor

Key Finding: Classical ML outperforms state-of-the-art transformers!

Why Simple Models Won

- **Explicit Feature Engineering**
- **Balanced Class Weights**
- **Sample Efficiency**
- **FastText for Arabic**

	precision	recall	f1-score	support
Biased Against Israel	1.0000	1.0000	1.0000	1
Biased Against Palestine	0.8889	0.8533	0.8707	75
Others	0.7500	0.7059	0.7273	17
Unbiased	0.9205	0.9456	0.9329	147
accuracy			0.9000	240
macro avg	0.8899	0.8762	0.8827	240
weighted avg	0.8989	0.9000	0.8992	240

	precision	recall	f1-score	support
Biased Against Israel	0.5000	1.0000	0.6667	1
Biased Against Palestine	0.8889	0.8533	0.8707	75
Others	0.6875	0.6471	0.6667	17
Unbiased	0.9067	0.9252	0.9158	147
accuracy			0.8833	240
macro avg	0.7458	0.8564	0.7800	240
weighted avg	0.8839	0.8833	0.8831	240

Lesson: Domain knowledge + simple ML > raw model scale

Specialists vs. Generalist

System 1 (Specialist Ensemble): 0.75

- Language-specific optimization (no cross-lingual interference)
- Tailored preprocessing per language
- Diverse models correct each other's errors
- +117% improvement over System 2

System 2 (Generalist): 0.35

- Cross-lingual confusion during fine-tuning
- Cannot use language-specific preprocessing
- All risk concentrated in single model
- Failed completely on "Unbiased" class (0.0 F1)

Verdict: Specialists win decisively

Key Contributions

- 5-Stage Sequential Pipeline
- Modular architecture for reproducibility
- Language-Aware Voting System
- Dynamic routing based on language detection
- Dual Preprocessing Strategy
- Specialists (0.75) >> Generalist (0.35)
- Classical ML (0.88) >> Transformers (0.47)

Limitations & Future Work

Current Limitations:

- Ensemble (0.75) below best individual model (0.88)
- Transformers underperformed → underutilized in voting
- Minority classes still challenging (<15% of data)
- Translation quality affects specialist training

Future Directions:

- Better transformer fine-tuning (few-shot learning, prompting)
- Confidence-weighted voting (trust stronger voters more)
- Advanced imbalance techniques (focal loss, SMOTE)
- Larger datasets for minority bias classes
- Interpretability analysis (what features drive decisions?)

Conclusions

- **Specialist ensemble outperforms generalist by 117%**
  - System 1 (0.75) vs System 2 (0.35)
- **Classical ML surprised us by beating transformers**
  - Random Forest (0.88) vs MARBERT (0.50)
- **Language-specific routing matters**
  - Avoid cross-lingual interference
- **Domain knowledge is powerful**
  - Feature engineering beats raw model scale

Thank You