Bias Detection in Media: Traditional Models vs. Transformers in Analyzing Social Media Coverage of the Israeli-Gaza Conflict

Marryam Yahya[1], Esraa Ismail[2], Mariam Nabil[3], Yomna Ashraf[4], Nada Radwan[5], Ziad Elshaer[6], Ensaf Mohamed[7]

School of Information Technology and Computer Science
Nile University, Giza, Egypt
Emails: {M.Yahya2163, E.Ismail2165, M.Nabil2184, Y.Ashraf2278, N.Ahmed2128, ZElshaer, EnMohamed}@nu.edu.eg

Abstract

Bias in news reporting significantly influences public perception, particularly in sensitive and polarized contexts like the Israel-Gaza conflict. Detecting bias in such cases presents unique challenges due to political, cultural, and ideological complexities, often amplifying disparities in reporting. While prior research has addressed media bias and dataset fairness, these approaches inadequately capture the nuanced dynamics of the Israel-Gaza conflict. To address this gap, we propose an NLP-based framework that leverages Nakba narratives as linguistic resources for bias detection in news coverage. Using a multilingual corpus focusing on Arabic texts, we apply rigorous data cleaning, pre-processing, and methods to mitigate imbalanced class distributions that could skew classification outcomes. Our study explores various approaches, including Machine Learning (ML), Deep Learning (DL), Transformer-based architectures, and generative models. The findings demonstrate promising advancements in automating bias detection, and enhancing fairness and accuracy in politically sensitive reporting.

Keywords: NLP, Text Classification, Bias-Detection, Nakba Narratives

1 Introduction

Bias detection in news reporting has become a crucial area of research, especially given its significant impact on public opinion and political polarization. In today's digital age, where information spreads rapidly through online platforms, news outlets are essential in shaping how people perceive events. However, media coverage often reflects underlying ideological or geopolitical biases, which can influence how audiences interpret the news. Detecting and understanding these biases is key to promoting ethical journalism and ensuring news reporting remains balanced and impartial. Recent advancements in Natural Language Processing (NLP) have introduced powerful new tools for identifying subtle biases in news articles. Machine learning models, in particular, have made significant strides in uncovering these hidden biases. Yet, these methods face unique challenges regarding sensitive and complex topics like the Israel-Gaza conflict. The language used in such coverage is heavily influenced by historical, cultural, and political factors, making it difficult for existing models to detect biases effectively. A more nuanced approach is needed to tackle this—one that goes beyond general bias detection methods and considers the conflict's specificities. While much of the existing research on bias detection in news has focused on more general forms of bias, such as political slant or ideological bias, the Israel-Gaza conflict presents a different set of challenges. Many studies have looked at these issues in languages like English, but often neglect the complexities of covering sensitive geopolitical topics. Additionally, the lack of annotated datasets focused on this conflict makes it even harder to develop effective bias detection tools. This paper addresses these gaps by creating a specialized NLP framework to detect and annotate biased coverage related to the Israel-Gaza conflict. Our work is based on the foundational project "BiasFignews" (SinaLab, 2024), which collected data on the Israeli-Gaza conflict. "BiasFignews" is a comprehensive multilingual corpus of 12,000 Facebook posts annotated for bias and propaganda. The corpus includes posts in Arabic, Hebrew, English, French, and Hindi, covering various events during the Israeli War on Gaza from October 7, 2023, to January 31, 2024.

Our main contributions include:

- Handling the imbalanced classes of the datasets to get the best performance for the models.
- Applying advanced linguistic and machine learning techniques to detect biases in news

content.

- Thoroughly evaluating the performance of these models.

By tackling the unique challenges of bias detection in conflict reporting, we hope to contribute to the development of more ethical journalism and improve the quality of media coverage in sensitive areas.

To achieve our goals, we use a comprehensive approach that includes data collection, cleaning, and pre-processing, followed by model development using various machine learning algorithms. Our first step is to create a multilingual annotated dataset scraped from social media platforms, focusing on news posts about the Israel-Gaza conflict. After addressing issues such as data imbalance, we apply advanced NLP techniques -such as transformer and sequential models like T5 and Bi-LSTM - and explore a variety of Machine Learning algorithms, including SVM, Random Forest, and XGBoost. Through experiments, we will benchmark these models and assess their performance in detecting bias, with a particular focus on how well they generalize across different languages and types of bias.

The rest of the paper is organized as follows: Section 2 will cover the Related Work, Section 3 will present our proposed Materials & Methods, Section 4 will present the Results & Discussion, Section 5 will Conclude the proposed work and discuss the recommendations of our future work and finally, Section 6 will represent the faced limitations in our work.

2 Related Work

This section reviews prominent studies on bias detection in NLP, focusing on their methodologies, challenges, and limitations. While existing work has explored media and language bias, few studies address the specific complexities of geopolitical conflicts like the Israel-Gaza conflict, especially in multilingual and culturally nuanced contexts.

Nadeem et al. (Nadeem and Raza, 2021) examine political bias in U.S. news articles, particularly content about former President Donald Trump. They apply a TensorFlow deep neural network (DNN) with Bag-of-Words (BoW) representation, TF-IDF weighting, and K-means clustering for pattern detection. The SimCSE framework outperforms these methods by effectively capturing subtle sentence-level biases.

Evans et al. (Evans et al., 2024) investigate how human biases influence NLP models, particularly in hate speech detection. Their work utilizes datasets to train the Emotion-Transformer model based on DistilBERT. While combining datasets improves bias detection for specific categories, they highlight persistent challenges in addressing imbalances in multi-target bias tasks.

Rodrigo-Gines et al. (Rodrigo-Ginés et al., 2024) conduct a systematic review categorizing types of media bias and distinguishing it from misinformation and disinformation. They emphasize the limitations of existing datasets and methods, calling for improved detection techniques to ensure accuracy and reliability in bias detection.

Khattak et al. (Donald et al., 2023) explore bias in customer interaction datasets, focusing on ethical data handling and fairness. They underscore the importance of mitigating bias during data preparation and advocate for enhanced methods to reduce biases in training datasets while ensuring compliance with GDPR.

Despite these advancements, existing studies lack a focus on media bias in the unique context of geopolitical conflicts, such as the Israel-Gaza conflict. The limited exploration of multilingual corpora, particularly in Arabic, and challenges with imbalanced data emphasize the need for specialized frameworks tailored to such sensitive and polarized scenarios.

3 Materials and Methods

This section provides a detailed overview of the dataset description and the proposed model pipeline, including data cleaning and preprocessing, handling imbalanced classes, the embedding models, and the classification models used.

3.1 Dataset Description

We employed a multilingual corpus annotated for bias and propaganda scraped from the Facebook platform (Duaibes et al., 2024) to implement our models. This corpus was constructed as a contribution to the FigNews 2024 Shared Task on News Media Narratives for framing the Israeli War on

Gaza. The dataset covers events during the war from October 7, 2023, to January 31, 2024. The corpus comprises 12,000 posts in five languages: Arabic, Hebrew, English, French, and Hindi, with 2,400 posts per language.

3.2 Methodology

The proposed model pipeline in Figure 1 consists of three phases: data cleaning and preprocessing, addressing class imbalance and ensuring class balance, and applying different models from various paradigms, including traditional machine learning models, transformer-based models, and generative models. These phases will be discussed in detail in the following subsections.

3.2.1 Data Cleaning & Pre-Processing

The initial phase of the pipeline involved cleaning and preparing the dataset to classify whether Arabic text is biased or not. Unnecessary columns such as Batch, Source Language, ID, Type, and others were removed, retaining only the "Arabic MT" and "Bias" columns. Null and duplicate fields were dropped, reducing the dataset to 10,800 rows.

Subsequently, text pre-processing was applied, including the removal of hashtags, URLs, emails, emojis, Arabic diacritics, and Tatweel. Arabic text normalization was performed by unifying Alif variants, replacing Taa Marbuta with Haa, and Alef Maqsura with Ya, as well as removing repeated characters. The dataset was then checked for class balance, as imbalanced data can lead to biased models favoring majority classes.

3.2.2 Handling Imbalanced Classes

To address the class imbalance, we employed Borderline-SMOTE (Han et al., 2005), which focuses on generating synthetic samples for minority class instances near the decision boundary. Unlike traditional SMOTE, this method emphasizes borderline samples likely to be misclassified due to proximity to majority class instances, enhancing model performance for minority classes.

We applied Borderline-SMOTE1, which generates synthetic samples exclusively from borderline minority samples. This approach improved decision boundary learning and classification performance. A comparison of label distributions before and after applying Borderline-SMOTE is shown in Figure 2.

3.2.3 Embeddings Model

To generate numerical representations of text data, we utilized the Multilingual E5 model (Wang et al., 2024), a large language model pre-trained on diverse languages and tasks. This model encodes text into high-dimensional vectors that capture semantic meaning. Using Hugging Face Transformers, the model tokenizes input text, encodes it via its encoder, and applies mean pooling to produce fixed-size embeddings. These embeddings map semantically similar words or phrases to vectors close to each other, enabling effective clustering, classification, and bias analysis.

In our research, we combined advanced large language models (LLMs), sequential, and transformer-based models to ensure robust and nuanced text representations for further bias detection.

3.2.4 Generative and Transformer-Based Models

- **Silma LLM:** A 9-billion-parameter generative model optimized for Arabic text tasks. It was used to detect bias in news articles by employing prompt engineering, which guides the model to classify text accurately and suggest neutralizing strategies for biased language.

- **T5 Encoder-Decoder Model:** The T5 model (Raffel et al., 2019) treats all tasks as text-to-text transformations, leveraging its pre-trained architecture to generate embeddings. This model captures complex semantic relationships in the dataset, enabling detailed and meaningful analysis for bias detection.

- **AraBERT Model:** AraBERT (Antoun et al., 2020), a BERT-based model tailored for Arabic, was fine-tuned using weighted sampling and Focal Loss to handle class imbalance. Despite its strong performance in general tasks, it struggled with minority class predictions in bias detection.

3.2.5 Deep Learning Models

Deep learning models are powerful tools for extracting complex patterns and representations from data. These models excel in analyzing text by capturing nuanced relationships and dependencies, making them essential for tasks like text classification, bias detection, and sentiment analysis.

- **LSTM:** Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997)

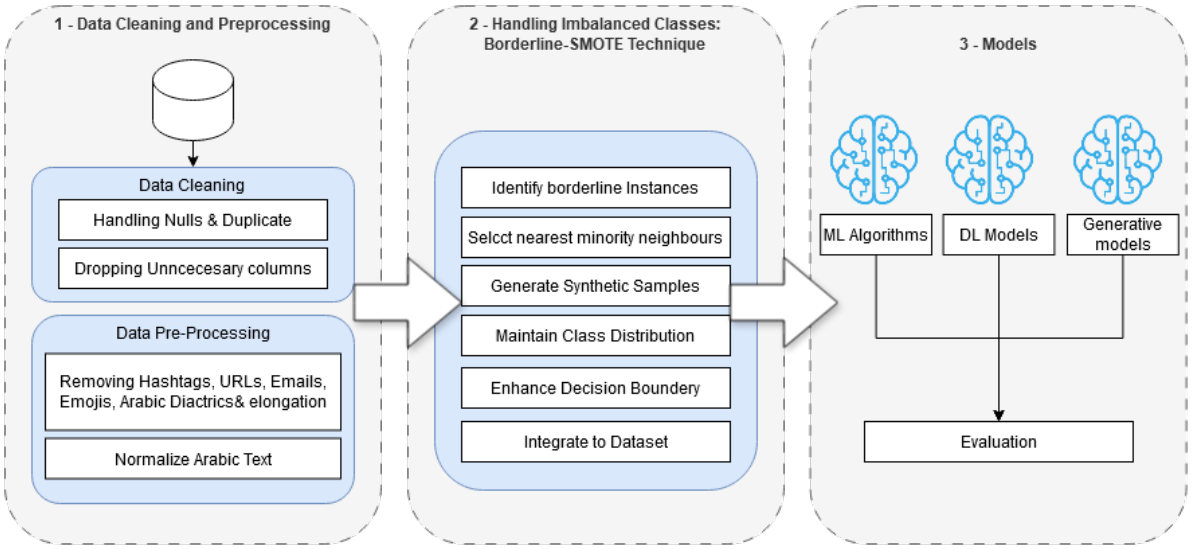

Figure 1: Model Pipeline

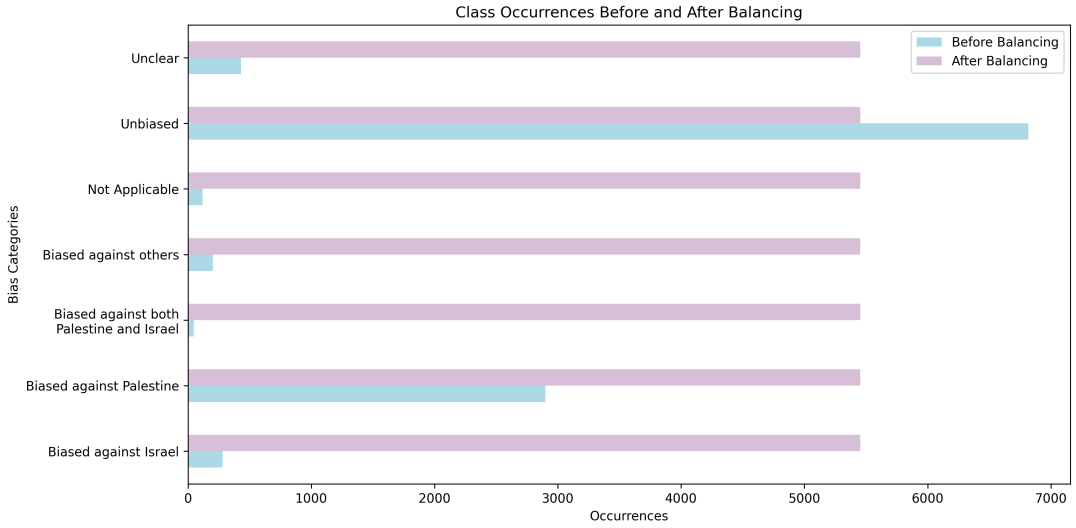

Figure 2: Bias Class Distribution

capture long-term dependencies in sequential data. In our research, LSTMs generated embeddings by preserving context over sequences, aiding in comprehensive text representation.

- **Bi-LSTM:** Bidirectional LSTM (Huang et al., 2015) extends LSTMs by capturing context from both past and future sequences. This bidirectional capability enhanced the quality of embeddings for deeper textual analysis.

- **Bi-GRU with Attention:** Combining Bidirectional GRUs (Wang et al., 2017) and attention mechanisms, this model highlighted important text features. Its computational efficiency and focus on relevant input parts improved the embeddings for bias detection and information retrieval tasks.

3.2.6 Machine Learning Algorithms

We utilized several machine learning algorithms to classify biased text effectively. Below is a concise summary of the models implemented:

- **SVM:** Support Vector Machine (SVM) identifies the optimal hyperplane that separates classes. Using kernel functions (e.g., RBF), it handles non-linear separations efficiently. In our implementation, SVM demonstrated robust performance for binary classification tasks by leveraging its mathematical rigor.

- **Random Forest:** An ensemble method that combines multiple decision trees, leveraging bagging to avoid overfitting. Each tree trains on a random subset of data, and predictions are made via majority voting. We used 100 trees with a random state of 42 to ensure consistent results.

- **XGBoost:** A boosting algorithm that sequentially builds trees to minimize residual errors. Configured with 100 estimators, a learning rate of 0.1, and a maximum depth of six, XGBoost provided high accuracy by optimizing for performance with hyperparameters like subsample and column sampling.

- **Decision Tree:** This interpretable model splits data into subsets based on feature values but risks overfitting without proper pruning. Using the Gini impurity criterion, we trained the model with a random state of 42 to ensure reproducibility.

- **CatBoost:** A gradient boosting model optimized for categorical features. By using ordered boosting and innovative handling of categorical data, CatBoost provided high accuracy. Parameters like 150 iterations, a learning rate of 0.1, and depth 6 were used for optimization.

- **Logistic Regression:** A statistical model for binary classification, Logistic Regression assumes linear separability of classes. Configured with a maximum of 1000 iterations and a random state of 42, the model offered simplicity and interpretability.

- **Gaussian Naive Bayes:** A probabilistic model leveraging Bayes' theorem with Gaussian distributions to handle continuous features. It proved effective for text classification, with its simplicity making it ideal for high-dimensional data.

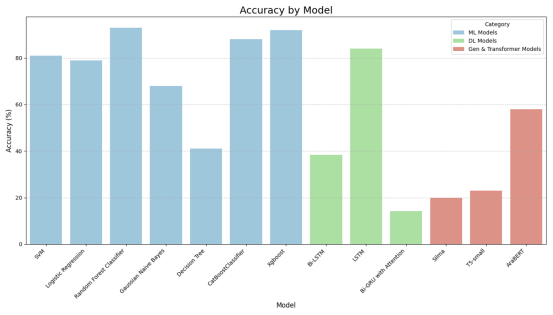

Figure 3: Accuracy

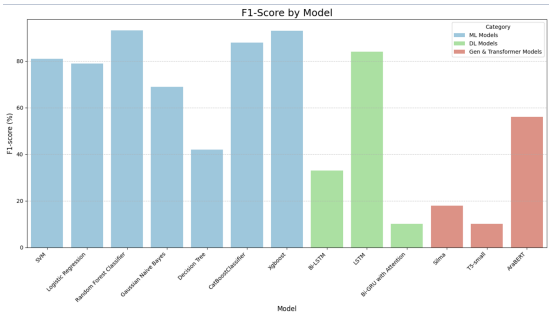

Figure 4: F1-Score

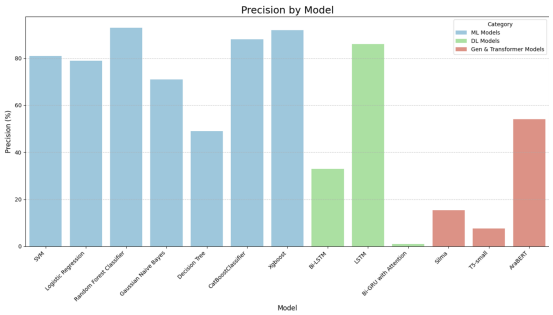

Figure 5: Precision

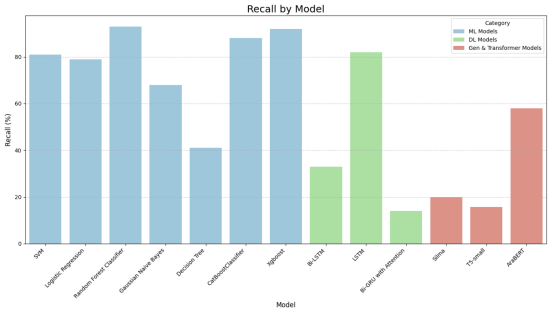

Figure 6: Recall

Figure 7: Performance Metrics Visualization

4 Results & Discussion

The models exhibit varying performance, with the accuracy and F1-scores for each summarized in table 1.

The table shows that in machine learning algorithms, the Random Forest Classifier has the highest performance with an accuracy of 93%, an F1-score of 93.23%, a precision of 93%, and a recall of 93%. XGBoost followed closely with an accuracy of 92%, an F1-score of 93%, a precision of 92%, and a recall of 92%, indicating strong performances along all metrics. Similarly, CatBoostClassifier achieved good performance with an accuracy of 88%, an F1-score of 88%, a precision of 88%, and a recall of 88Where for the Deep Learning

ML Algorithms				
Model	**Accuracy (%)**	**F1-score (%)**	**Precision (%)**	**Recall (%)**
SVM	81	81	81	81
Logistic Regression	79	79	79	79
Random Forest Classifier	**93**	**93.23**	**93**	**93**
Gaussian Naive Bayes	68	69	71	68
Decision Tree	41	42	49	41
CatBoostClassifier	88	88	88	88
XGBoost	92	93	92	92
Deep Learning Models				
Model	**Accuracy (%)**	**F1-score (%)**	**Precision (%)**	**Recall (%)**
Bi-LSTM	38.4	33	33	33
LSTM	**84**	**84**	**86**	**82**
Bi-GRU with Attention	14.18	10.2	8.9	14.18
Generative Models & Transformer Based Models				
Model	**Accuracy (%)**	**F1-score (%)**	**Precision (%)**	**Recall (%)**
Silma	20	18	15.32	20
T5-small	15.65	10.10	7.62	15.65
AraBERT	**58**	**56**	**54**	**58**

Table 1: Model Performance Summary by Type with Precision and Recall

models, the LSTM model achieved an accuracy of 84%, an F1-score of 84%, a precision of 86%, and a recall of 82%. Bi-LSTM and Bi-GRU with Attention achieved lower results: Bi-LSTM with an accuracy of 38.4%, an F1-score of 33%, a precision of 33%, and a recall of 33%, and Bi-GRU with Attention with an accuracy of 14.18%, an F1-score of 10.2%, a precision of 8.9%, and a recall of 14.18%. On the other hand, the generative models T5-small also had a low accuracy of 15.65%, an F1-score of 10.10%, a precision of 7.62%, and a recall of 15.65%. The generative model Silma, based on prompt engineering, performed with an accuracy of 20%, an F1-score of 18%, a precision of 15.32%, and a recall of 20%.

Among the traditional models, SVM achieved an accuracy of 81%, an F1-score of 81%, a precision of 81%, and a recall of 81%, which were reasonable but not high as compared to the ensemble methods. Logistic Regression also achieved a similar performance with an accuracy of 79%, an F1-score of 79%, a precision of 79%, and a recall of 79%. Gaussian Naive Bayes showed an accuracy of 68%, an F1-score of 69%, a precision of 71%, and a recall of 68%, while Decision Tree had a moderate performance with an accuracy of 41%, an F1-score of 42%, a precision of 49%, and a recall of 41%. AraBERT, a pre-trained language model specific to the Arabic language, achieved an accuracy of 58%, an F1-score of 56%, a precision of 54%, and a recall of 58%. While its performance outperformed the Decision Tree model and some of the Deep Learning models. It is shown that Generative and Transformer-based models such as SILMA and T5 performed worse than traditional machine learning (ML) models. This is due to many reasons, including that traditional ML models often benefit from feature engineering, where manually selecting and transforming relevant features can lead to better performance. Additionally, traditional ML models are designed for specialized tasks like classification, making them more effective for these specific problems compared to generative models optimized for generating new data. Moreover, traditional ML models have built-in inductive biases that make them well-suited for certain tasks, such as Random Forest being particularly adept at constructing multiple decision trees during training and outputting the mode of the classes, whereas transformers may require more data and computational resources to achieve similar results. Figures 3,4,5,6 illustrate the results of the Machine learning, Deep Learning, Generative and Transformer-based models for this work.

5 Conclusion

In this paper, we deal with the critical task of detecting bias in news reporting on the conflict between

Israel and Gaza. Applying an Arabic corpus of texts, advanced preprocessing methods, and several machine learning models, we have arrived at a robust framework for the detection of bias applicable to the nuanced and politically charged context of news reports on conflict situations. Among the tried methods, ensemble methods such as Random Forest and XGBoost showed better performance; thus, they are more suitable for this challenging classification. Our results indicate that although there are inherent difficulties arising from data imbalance, language-specific challenges, and subtle bias indicators, a good combination of data augmentation strategies such as Borderline-SMOTE along with state-of-the-art machine learning techniques can improve the detection of bias considerably. It is part of the larger aim of ensuring ethical journalism and offers a scalable methodology for media coverage analysis in sensitive geopolitical situations.

6 Limitations

This study has several limitations that are important to highlight. While the dataset is multilingual and extensive, it is limited to Facebook posts from a specific period, which makes it harder to generalize the findings to other platforms, time frames, or contexts. Annotating bias and propaganda is inherently subjective, especially in politically sensitive topics like the Israel-Gaza conflict, which could affect the quality of model training and evaluation. The transformer-based models we used, though effective, rely heavily on the training data and often struggle to identify subtle or context-specific biases shaped by historical and cultural factors. Similarly, addressing class imbalance with Borderline-SMOTE might oversimplify the complexity of real-world data, risking overfitting for minority classes and missing nuances in bias detection. Working with Arabic texts brought its own set of challenges, such as the language's rich morphology, diverse dialects, informal variations, and frequent code-switching, all of which made preprocessing more difficult and may have caused some loss of linguistic subtleties. Moreover, the lack of standardized resources for Arabic and domain-specific tools limited our ability to fully capture the complexity of biased reporting. Moving forward, we plan to address these limitations by expanding the dataset to include posts from other platforms and time frames, fine-tuning transformer models with domain-specific adaptations, and exploring hybrid approaches that combine linguistic insights with advanced deep learning techniques to better detect bias, particularly in Arabic texts.

References

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Andy Donald, Apostolos Galanopoulos, Edward Curry, Emir Muñoz, Ihsan Ullah, M. A. Waskow, Maciej Dabrowski, and Manan Kalra. 2023. Bias detection for customer interaction data: A survey on datasets, methods, and tools. *IEEE Access*, 11:53703–53715.

Lina Duaibes, Areej Jaber, Mustafa Jarrar, Ahmad Qadi, and Mais Qandeel. 2024. Sina at fignews 2024: Multilingual datasets annotated with bias and propaganda. *Preprint*, arXiv:2407.09327.

Ana Sofia Evans, Helena Moniz, and Luísa Coheur. 2024. A study on bias detection and classification in natural language processing. *Preprint*, arXiv:2408.07479.

Hui Han, Wenyuan Wang, and Binghuan Mao. 2005. Borderline-smote: A new over-sampling method in imbalanced data sets learning. In *International Conference on Intelligent Computing*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

MU Nadeem and S Raza. 2021. Detecting bias in news articles using nlp models.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Francisco-Javier Rodrigo-Ginés, Jorge Carrillo de Albornoz, and Laura Plaza. 2024. A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it. *Expert Systems with Applications*, 237:121641.

SinaLab. 2024. Biasfignews: A multilingual corpus of facebook posts annotated for bias and propaganda. GitHub.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Nan Wang, Jin Wang, and Xuejie Zhang. 2017. Ynu-hpcc at ijcnlp-2017 task 4: Attention-based bi-directional GRU model for customer feedback analysis task of English. In *Proceedings of the IJCNLP 2017, Shared Tasks*, pages 174–179, Taipei, Taiwan. Asian Federation of Natural Language Processing.