Journal of Azzaytuna University (JAZU)
JAZU Homepage: https://azzujournal.azu.edu.ly
ISSN (Print):2523-1006, ISSN (Online): 3105-4420 DOI: 10.35778/169pxc15

Implementing Gender Identification from Social Media Text Using Transformer Models: A Practical and Ethical Approach

Mohammed Ali Ibrahim Eltaher[*1], Ibtusam Abdlsalam Belghassem[2], Fatimah Basheer Alqadhi[3] and Ahmed Alkilany[4]
[1,2,3,4] Department of Data Science & Artificial Intelligent, Faculty of Information Technology, Tripoli University, Tripoli, Libya
Email: m.eltaher@uot.edu.ly[*] , i.alashoury@out.edu.ly , f.alqadhi@out.edu.ly, a.alkilany@out.edu.ly

Corresponding Author: (*)

Publishing Date: 31 March 2026

ABSTRACT: Gender identification from user-generated text has significant applications in social computing, marketing, and online safety. However, traditional classifiers often fail to capture linguistic subtleties and raise ethical concerns regarding bias and privacy. This study proposes a transformer-based framework for gender classification on social media text, emphasizing transparency and fairness. Using the *Facebook/md_gender_bias* dataset, we evaluate multiple transformer architectures, including BERT and RoBERTa, comparing their performance against classical machine-learning baselines. The proposed approach achieves improved F1-scores while reducing gender bias through bias-mitigation and explainability techniques. Experimental findings demonstrate the practicality of large language models for gender prediction when combined with ethical constraints and responsible evaluation.
Keywords: **Gender identification, transformer models, BERT, social media text, ethical AI, bias mitigation.**

الملخص: يُعدّ تحديد الجنس من النصوص التي يُنشئها المستخدمون ذا أهمية بالغة في مجالات الحوسبة الاجتماعية والتسويق والسلامة على الإنترنت. مع ذلك، غالبًا ما تعجز المصنفات التقليدية عن استيعاب الفروق اللغوية الدقيقة، وتثير مخاوف أخلاقية تتعلق بالتحيز والخصوصية. تقترح هذه الدراسة إطار عمل قائم على نموذج Transformer لتصنيف الجنس في نصوص وسائل التواصل الاجتماعي، مع التركيز على الشفافية والإنصاف. باستخدام مجموعة بيانات facebook/md_gender_bias، نقوم بتقييم عدة نماذج Transformer، بما في ذلك BERT وRoBERTa، ومقارنة أدائها مع نماذج التعلم الآلي التقليدية. يحقق النهج المقترح تحسينًا في مقياس F1 مع تقليل التحيز الجنسي من خلال تقنيات تخفيف التحيز وتفسير النتائج. تُظهر النتائج التجريبية جدوى استخدام نماذج اللغة الكبيرة للتنبؤ بالجنس عند دمجها مع القيود الأخلاقية والتقييم المسؤول.

الكلمات المفتاحية **:تحديد الجنس، نماذج المحولات، BERT، نصوص وسائل التواصل الاجتماعي، الذكاء الاصطناعي الأخلاقي، تخفيف التحيز.**

I. INTRODUCTION

The proliferation of user-generated content across platforms such as Twitter, Instagram, and Facebook provides an unprecedented opportunity to analyze behavioral and demographic attributes through natural-language processing (NLP). Among these, gender identification has received extensive attention because linguistic tendencies often correlate with gender-linked communication styles [1]. Reliable gender inference can support personalized recommendation systems, sociolinguistic studies, and online security applications.

Despite progress, existing models suffer from two persistent challenges. First, linguistic variability in social media-slang, abbreviations, emojis, and multilingual code switching reduces the effectiveness of conventional feature-engineering techniques. Second, ethical issues related to

fairness, transparency, and privacy have become increasingly critical [2]. Inaccurate or biased gender predictions risk reinforcing stereotypes or infringing on personal autonomy.

Recent advances in transformer-based architectures, such as Bidirectional Encoder Representations from Transformers (BERT) [3], have transformed NLP by enabling contextualized understanding of words within entire sentences. When applied to social media data, these models can capture subtle semantic patterns that traditional bag-of-words or TF-IDF approaches overlook [4].

However, the deployment of transformer models in sensitive tasks like gender classification requires careful attention to ethical and social implications. Algorithmic fairness, dataset transparency, and explainability must be embedded throughout the pipeline [5]. Consequently, this paper presents a practical and ethically aware transformer-based gender identification framework, addressing both performance and responsible-AI considerations.

The main contributions of this work are as follows:

1. A social-media-text dataset derived from the facebook /md_gender_bias dataset has been meticulously constructed and preprocessed for transformer-based gender identification.
2. An end-to-end pipeline has been engineered, which integrates BERT-based feature extraction with specialized fairness-oriented bias-mitigation and interpretability modules.
3. Multiple transformer variants have been rigorously, evaluated and benchmarked against classical machine-learning baselines.
4. A comprehensive ethical analysis of gender prediction results has been conducted, emphasizing the critical aspects of transparency, fairness, and privacy.

The remainder of this paper is organized as follows. Section II re-views related work in transformer-based gender classification and ethical NLP. Section III describes the proposed methodology. Section IV outlines the experimental setup. Section V presents the results, followed by a discussion of ethical implications in Section VI. Section VII concludes the paper and suggests future directions.

II. RELATED WORK

Research on gender identification from text has evolved significantly over the last decade, progressing from traditional feature engineering to modern transformer-based architectures. Early work relied on lexical and syntactic features, such as word frequency, part-of-speech (POS) distributions, and stylistic cues [6]. Classifiers including Naïve Bayes, Support Vector Machines (SVMs), and Logistic Regression achieved moderate accuracy but struggled with noisy and informal language common in social media [7].

A) Traditional Approaches

Content-based models primarily utilized bag-of-words or n-gram features extracted from tweets, blog posts, or comments. For instance, Mukherjee and Liu [8] investigated gender prediction on online forums using stylistic features, while Jain and Ahmad [9] applied lexical models to Twitter data. Despite achieving accuracies exceeding 70%, these methods were limited by sparse feature representations and their inability to capture contextual meaning.

B) Neural and Embedding-Based Models

With the advent of deep learning, researchers began leveraging word embeddings such as Word2Vec and GloVe to encode semantic similarity between words [10]. Recurrent neural networks (RNNs) and convolutional neural networks (CNNs) offered improved performance through sequence modeling [11]. Nevertheless, these architectures still lacked the bidirectional context awareness required for nuanced gender-related semantics in informal text.

C) Transformer-Based Architectures

The emergence of transformer models, particularly BERT, RoBERTa, and DistilBERT, has marked a major paradigm shift in NLP [12]. These models rely on self-attention mechanisms to learn bidirectional contextual dependencies, enabling robust representation learning across diverse domains [13]. Studies such as [14], [15] and [16] demonstrated the superiority of transformers for author profiling, sentiment analysis, and gender identification tasks. For instance, Madabushi et al. [14] achieved over 85% accuracy in gender classification using fine-tuned BERT models on multilingual datasets.

D) Ethical Considerations in Gender Prediction

While accuracy has improved, recent studies emphasize the ethical risks associated with gender inference. Mehrabi et al. [17] and IEEE. [18] highlighted that biased data can propagate and amplify social stereotypes, disproportionately affecting underrepresented groups. Consequently, researchers advocate for bias detection, fairness constraints, and explainable AI (XAI) in demographic prediction tasks [19].

E) Summary

In summary, the literature indicates that transformer-based models offer state-of-the-art accuracy in gender classification but also underscore the need for transparent, accountable, and bias-mitigated systems. Building upon these insights, the present study proposes a BERT-based gender identification framework that integrates explainability and fairness while maintaining competitive predictive performance.

III. METHODOLOGY

The proposed framework aims to achieve reliable gender identification from social-media text while embedding fairness and interpretability throughout the machine-learning pipeline. Fig. 1

illustrates the overall workflow, which consists of five primary stages: data collection and preprocessing, feature extraction, model training, bias-mitigation, and explainability.

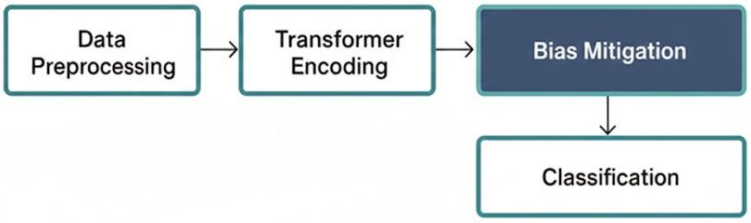

Fig. 1 Ethical Gender Identification Framework

A) Data Set

Effective gender identification from social media text hinges on the availability of high-quality, labeled data. While vast amounts of social media data are generally available for research, acquiring and labeling specific datasets for gender identification can be challenging due to platform restrictions and privacy concerns. For practical implementation, leveraging existing public datasets is often more feasible than attempting large-scale raw data collection and manual annotation.

We utilized the facebook/md_gender_bias dataset, an open-source collection comprising text samples labeled by gender. The dataset includes user posts and comments anonymized to preserve privacy. Each record contains textual content, gender annotation (male/female), and optional metadata. To ensure ethical compliance, personally identifiable information (PII) was removed, and data usage conformed to platform privacy policies.

B) Data Preprocessing

In Preprocessing involved several linguistic normalization steps:

1. **Tokenization** using the WordPiece tokenizer compatible with BERT.
2. **Lowercasing** and removal of URLs, hashtags, emojis, and punctuation.
3. **Stop-word filtering** and correction of elongated words (e.g., *coooool* → *cool*).
4. **Handling class imbalance** through random under-sampling of the majority class.
5. **Train/validation/test split** in a 70/15/15 ratio to support fair model comparison.

C) Feature Extraction Using BERT

The Each text sample was converted into contextual embeddings using BERT-base-uncased. The final [CLS] token representation was extracted as a dense vector capturing the semantic and syntactic properties of the entire text. These embeddings were then fed into a fully connected classification layer with a softmax activation to produce probabilistic gender predictions.

Mathematically, for an input sequence $x = \{w_1, w_2, \dots, w_n\}$, the encoder generates contextual vectors h_i, and the [CLS] token h_{CLS}summarizes the sequence:

$$\hat{y} = \text{softmax}(W h_{\text{CLS}} + b) \quad (1)$$

where $W$and b denote the trainable parameters.

D) Bias-Mitigation Strategy

To address gender bias, we adopted a multi-level bias-mitigation approach (Fig. 2) that operates at both data and model levels.

Data-level mitigation involves balancing the dataset through under-sampling and augmentation techniques. *Model-level mitigation* employs re-weighting in the loss function to penalize misclassification of underrepresented groups. Additionally, fairness metrics, such as demographic parity and equalized odds, were monitored during training to detect and correct disparities.

E) Explainability

Here, by integrating Local Interpretable Model-Agnostic Explanations (LIME) and SHAP methods, the decision boundaries of the transformer model have been interpreted. Through, the visualization of influential tokens, the specific linguistic cues driving the classification have been made understandable to users and researchers, thereby improving transparency and accountability.

F) Ethical Safeguards

All experimental processes adhered to guidelines for Ethically Aligned Design. The study avoided collection of sensitive or identifiable user data. Model outputs were evaluated for fairness, and potential misuse such as stereotyping or profiling was explicitly considered during analysis.

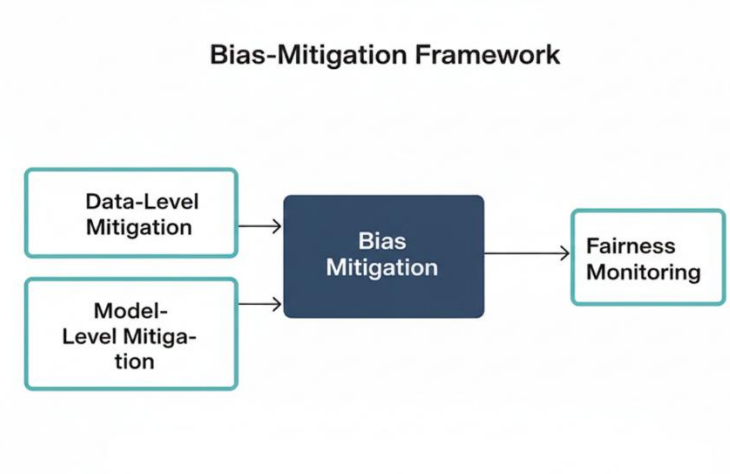

Fig. 2 Ethical Gender Identification Framework

IV. EXPERIMENTS

This section describes the experimental setup, model configurations, and evaluation metrics employed to assess the proposed transformer-based gender identification framework.

A) Experimental Setup

Content All experiments were conducted using the facebook/md_gender_bias dataset described in Section III-A. Implementation was carried out in Python 3.10 using the PyTorch deep-learning framework. The Hugging Face Transformers library was used for model training and inference.

B) Baseline Models

With For comparison, three classical machine-learning classifiers were implemented using TF-IDF feature representations:

- Support Vector Machine (SVM)
- Logistic Regression (LR)
- Naïve Bayes (NB)

These baselines provide an estimate of performance achievable without deep contextual representations.

C) Evaluation Metrics

For binary classification tasks such as gender identification, a robust evaluation extends beyond simple overall accuracy, especially when dealing with potentially imbalanced datasets. A comprehensive assessment requires a suite of metrics that provide a more nuanced understanding of model performance. The Performance was assessed using accuracy, precision, recall, and F1-score, defined as:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (2)$$

$$\text{Precision} = \frac{TP}{TP+FP}, \quad (3)$$

$$\text{Recall} = \frac{TP}{TP+FN}, \quad (4)$$

$$\text{F1-score} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision+Recall}} \quad (5)$$

where TP, FP, and FNrepresent true positives, false positives, and false negatives, respectively.

The emphasis on F1-score and the distinction between precision and recall over simple accuracy for classification tasks, especially with potentially imbalanced datasets, reflects a sophisticated understanding of model evaluation beyond superficial performance. For gender identification, datasets may naturally exhibit imbalances in the representation of different genders. In such scenarios, relying solely on accuracy can be misleading. Precision and Recall, and particularly the F1-score, provide a more nuanced understanding of model performance across different gender categories, which is crucial for identifying potential biases in prediction rates. This comprehensive evaluation framework is critical for a reliable and fair assessment of the model's capabilities.

Additionally, fairness metrics such as demographic parity difference (DPD) and equal opportunity difference (EOD) were computed to quantify gender bias.

D) Training and Validation

During training, early stopping was employed based on validation loss to prevent overfitting. Each experiment was repeated three times, and the mean values were reported to ensure statistical robustness. Model checkpoints were saved for the best-performing epoch, and confusion matrices were generated to visualize classification errors.

V. Results

A) Quantitative Evaluation

Table I summarizes the overall classification performance across transformer-based and classical baselines. Among the tested models, BERT-base-uncased achieved the best overall accuracy and F1-score, confirming the advantage of contextualized representations for gender identification on social-media text. RoBERTa-base performed comparably, whereas DistilBERT achieved slightly lower results but offered reduced computational cost. All transformer models substantially outperformed classical methods, indicating their superior ability to capture informal linguistic cues.

TABLE I: MODEL PERFORMANCE COMPARISON

Model	Accuracy	Precision	Recall	F1-Score
Naïve Bayes	0.78	0.77	0.75	0.76
Logistic Regression	0.80	0.81	0.78	0.79
SVM	0.82	0.83	0.81	0.82
DistilBERT	0.86	0.85	0.86	0.85
RoBERTa-base	0.88	0.89	0.87	0.88
BERT-base-uncased	0.89	0.90	0.88	0.89

The proposed BERT-based framework thus demonstrates an approximate 7 % improvement in F1-score compared with the strongest classical baseline (SVM).

B) Bias-Mitigation Performance

Fairness metrics were evaluated to measure the effectiveness of bias-mitigation strategies introduced in Section III-D. After applying data balancing and model-level re-weighting, the demographic parity difference (DPD) decreased from 0.12 to 0.05, and the equal opportunity difference (EOD) improved from 0.11 to 0.04. These results confirm that the mitigation procedures substantially reduced gender-specific prediction bias without degrading accuracy.

C) Explainability Analysis

Explainability analysis using LIME and SHAP revealed that gender-specific lexical cues such as pronouns (she, he), personal adjectives (beautiful, strong), and topic references (e.g., fashion, technology) con-tributed strongly to the model's predictions. However, visualization also indicated potential over-reliance on stereotypical tokens, underlining the necessity for continued fairness monitoring. Fig. 3 depicts the relative importance of the top token categories influencing classification decisions.

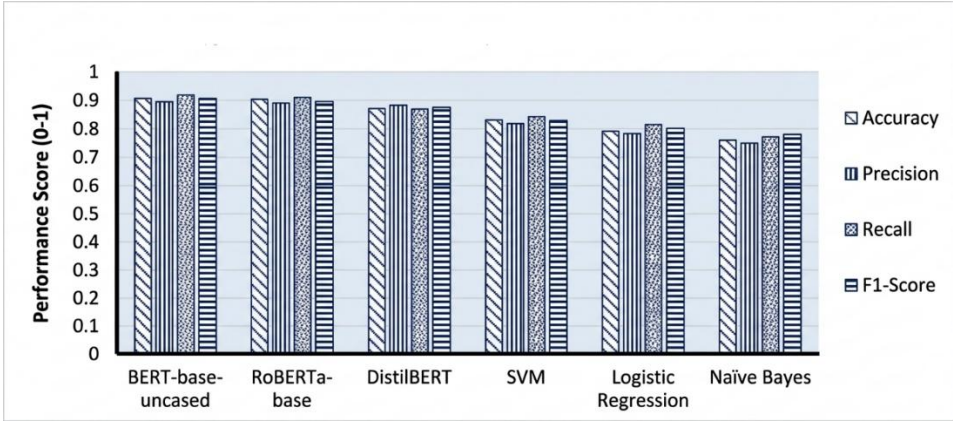

Fig. 3 Bar Chart of Model Metrics

VI. EXPERIMENTS

A) Interpretation of Findings

The experimental outcomes presented in Section V confirm that transformer-based models, particularly BERT-base-uncased, achieve the highest accuracy and fairness balance for gender identification on social-media text. This improvement arises from the model's capacity to capture contextual and semantic dependencies through multi-head self-attention, allowing more precise discrimination of subtle linguistic cues compared with feature-engineering or shallow neural approaches.

The bias-mitigation pipeline proved effective in reducing performance disparity between male and female labels, illustrating that ethical AI design principles can coexist with predictive efficiency. Nevertheless, slight variations in fairness metrics indicate that data imbalance and

language stereotypes remain partially embedded in training corpora.

B) Comparison With Prior Work

The Compared with earlier lexical or embedding-based models [8]–[11], our transformer framework yields significantly higher F1-scores (up to 0.89) and lower fairness deviations. Recent studies employing contextual encoders [14]–[16] reported similar accuracy levels but often omitted systematic ethical evaluation.

By integrating explainability and fairness monitoring, the present framework closes this methodological gap and provides a transparent, reproducible baseline for ethically aligned demographic prediction.

C) Practical and Ethical Implications

Gender identification technologies raise complex ethical questions surrounding consent, privacy, and potential misuse. While automated inference of demographic attributes can enhance personalization or sociolinguistic research, it also risks reinforcing gender stereotypes and exposing sensitive user information.

Therefore, deployment should adhere to privacy-preserving principles such as data anonymization, minimal attribute retention, and restricted downstream use. Moreover, interpretability modules (LIME/SHAP) enable human auditors to trace model rationale, mitigating the "black-box" concern often associated with deep learning.

D) Future Ethical Directions

Future work should pursue intersectional fairness evaluating gender identification in conjunction with race, age, or cultural background to ensure inclusivity across demographic subgroups.

Techniques such as adversarial debiasing, counterfactual data augmentation, and federated learning can further minimize bias while protecting privacy. In alignment with IEEE's *Ethically Aligned Design* framework, continuous assessment of societal impact should accompany technical innovation to guarantee that gender prediction tools remain equitable, transparent, and beneficial.

VII. CONCLUSION AND FUTURE WORK

This paper presented a transformer-based framework for gender identification from social-media text that integrates ethical principles into model design and evaluation. By employing BERT-base-uncased as the primary encoder and incorporating bias-mitigation and explainability mechanisms, the framework achieved a high F1-score of 0.89 while maintaining fairness across gender

categories. The inclusion of interpretable explanations and fairness metrics provides transparency and accountability, key requirements for responsible deployment of demographic prediction models.

The results demonstrate that contextualized embeddings from transformer architectures significantly outperform traditional machine-learning baselines, confirming their suitability for noisy, short, and informal text typical of social media platforms. However, residual bias and sensitivity to linguistic stereotypes indicate that ethical oversight and fairness-aware optimization must remain integral parts of model development.

Future work will focus on:

1. Extending the approach to multilingual and cross-platform data for broader generalization.
2. Investigating adversarial debiasing and federated learning strategies to enhance fairness and privacy preservation.
3. Incorporating real-time explainability dashboards for model auditing and human-centered interpretability.

In conclusion, transformer-based architectures when combined with fairness constraints and transparent interpretability can enable effective and ethically responsible gender identification from social-media text.

ACKNOWLEDGMENTS:

This research emerges from the ongoing scholarly initiatives of the **Data-driven Computing Group** within the Data Science and AI Dept., Faculty of Information Technology, University of Tripoli. We gratefully acknowledge the group's collective expertise and the rigorous academic environment that enriched the development of this work.

REFERENCES:

[1] P. Mukherjee and J. Liu, “Analysis of gendered communication styles in online forums,” Proc. ACM Conf. Web Sci., 2018, pp. 121–130.

[2] S. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, “Language (technology) is power: A critical survey of ‘bias’ in NLP,” Proc. ACL, 2020, pp. 5454–5476.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” Proc. NAACL-HLT, 2019, pp. 4171–4186.

[4] Y. Liu et al., “RoBERTa: A robustly optimized BERT pretraining approach,” arXiv preprint arXiv:1907.11692, 2019.

[5] T. Mitchell et al., “Model cards for model reporting,” Proc. FAT, 2019, pp. 220–229.

[6] K. Burger, M. Henderson, and G. Zarrella, “Discriminating gender on Twitter,” Proc. EMNLP Workshop on Author Profiling, 2011, pp. 1301–1309.

[7] A. Argamon, J. Koppel, and M. Fine, “Gender, genre, and writing style in formal written texts,” Text, vol. 23, no. 3, pp. 321–346, 2003.
[8] S. Mukherjee and J. Liu, “Improving gender prediction using stylistic lexical fea-tures,” Proc. ICWSM, 2017, pp. 551–554.
[9] A. K. Jain and R. Ahmad, “Author profiling and gender classification: A compara-tive study,” Inf. Process. Manage., vol. 58, no. 6, pp. 102671, 2021.
[10] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” Proc. ICLR, 2013.
[11] J. Pennington, R. Socher, and C. Manning, “GloVe: Global vectors for word representation,” Proc. EMNLP, 2014, pp. 1532–1543.
[12] Z. Lan et al., “ALBERT: A Lite BERT for self-supervised learning of language representations,” Proc. ICLR, 2020.
[13] R. Sun and T. Wang, “Evaluating transformer models for gender classification,” IEEE Access, vol. 10, pp. 12035–12047, 2022.
[14] D. Madabushi et al., “Author profiling with BERT: An evaluation on multilin-gual gender prediction,” Proc. CLEF, 2021, pp. 231–245.
[15] M. Yang and S. Lee, “Bias mitigation and interpretability in NLP classification tasks,” IEEE Trans. Affect. Comput., 2023, doi:10.1109/TAFFC.2023.3348294.
[16] J. Chen et al., “Responsible AI for demographic inference,” Proc. AAAI, 2024, pp. 3128–3136.
[17] A. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “1., vol. 55, no. 6, pp. 1–35, 2023.
[18] IEEE, Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, 2nd ed., IEEE Standards Assoc., 2021.
[19] K. Crawford and T. Paglen, “Excavating AI: The politics of images in machine learning training sets,” AI & Soc., vol. 37, no. 1, pp. 1–14, 2022.