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

[1] Dr. Mohammed Ali Eltaher [2] Mr. Ibtusam Abdlsalam Belghassem
[3] Mr. Fatimah Basheer Alqadhi
https://orcid.org/ 0009-0006-4841-2751 1 Author
https://orcid.org/ 0009-0005-4638-0971 2 Author
https://orcid.org/ 0009-0002-9694-1747 3 Author
[1 2 3] Department of Data Science & Artificial Intelligent, Faculty of Information Technology, Tripoli University, Tripoli, Libya
[1] m.eltaher@uot.edu.ly [2] i.alashoury@out.edu.ly [3] f.alqadhi@out.edu.ly

تنفيذ نموذج لتحديد الجنس من خلال نصوص وسائل التواصل الاجتماعي باستخدام نماذج المحولات: منظور عملي وأخلاقي

د. محمد علي الطاهر [1] أ. أبتسام عبد السلام ابوالقاسم [2] أ. فاطمة بشير القاضي [3]
قسم علم البيانات والذكاء الاصطناعي، كلية تقنية المعلومات، جامعة طرابلس، طرابلس، ليبيا.[1 2 3]
تاريخ الاستلام: 2026-06-01، تاريخ القبول: 2026-06-10، تاريخ النشر: 2026-06-25

Abstract:

Gender identification for social user network from user-generated text has significant applications in social computing, advertising, and online safety. Though, traditional classifiers often fail to capture linguistic subtleties and raise ethical concerns about bias and privacy. On this study, we propose a transformer-based framework for gender identification on social network text which highlighting transparency and fairness. We assess multiple transformer architectures, including BERT and RoBERTa, comparing their performance against traditional machine-learning baselines using the facebook/md-gender-bias dataset. The proposed method achieves improved for 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 classification, transformer models, BERT, social media network 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 spread of user-generated content across platforms such as Facebook ,Twitter and Instagram offers an extraordinary opportunity to analyze demographic attributes through natural language processing (NLP). Among these, gender identification has established extensive attention because linguistic tendencies often associate with gender linked communication styles [1]. Reliable gender inference can support personalized recommendation systems, sociolinguistic studies, and online security applications.

Despite progress, current models suffer from two 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 more and more critical [2]. Inaccurate and biased gender predictions risk reinforcing stereotypes or infringing on personal autonomy.

Current 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].

Nevertheless, the deployment of transformer models in sensitive tasks like gender identification requires a lot of attention to ethical and social implications. Algorithmic fairness, dataset transparency and explainability necessity be embedded over the pipeline [5]. Accordingly, this study presents a practical and ethically aware transformer-based gender identification framework by addressing both performance and responsible-AI considerations.

The main contributions of this work are as follows. First, a social media text dataset derived from the facebook /md_gender_bias dataset has been constructed and preprocessed for

transformer-based gender identification. Second, an end-to-end pipeline has been engineered. This integrates BERT-based feature extraction with specialized fairness-oriented, bias-mitigation and interpretability modules. Third, multiple transformer variants have been rigorously, assessed and benchmarked against traditional machine learning baselines. Fourth, a comprehensive ethical investigation 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 reviews related work in transformer-based gender identification 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 proposes future directions.

II. RELATED WORK

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

A) Traditional Approaches

Content-based models mainly utilized bag-of-words or n-gram features extracted from tweets, blog posts, or comments. For example, Kavuri and Kavitha [8] examined gender prediction on online forums using stylistic features. On the other hand, Lim et al. [9] applied lexical models to Twitter data. Despite achieving accuracies beyond 70%, these methods were limited by sparse feature representations and their incapability to capture contextual sense.

B) Neural and Embedding-Based Models

With the advent of deep learning, researchers started 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 enhanced performance through sequence modeling [11]. Yet, 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, mainly 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 and enabling robust representation learning across various domains [13]. Studies such as [14], [15] and [16] demonstrated the power of transformers for author profiling, sentiment analysis, and gender identification tasks. For illustration, Alzahrani and Jololian [14] accomplished over 86% accuracy in gender classification using BERT models on multilingual datasets.

D) Ethical Considerations in Gender Prediction

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

E) Summary

In summary, the literature shows that transformer-based models offer state-of-the-art accuracy in gender classification but also underline the need for transparent, accountable, and bias-mitigated systems. Building upon these insights, the current 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 that consists of five primary stages: data collection and preprocessing, feature extraction, model training, bias-mitigation, and explainability.

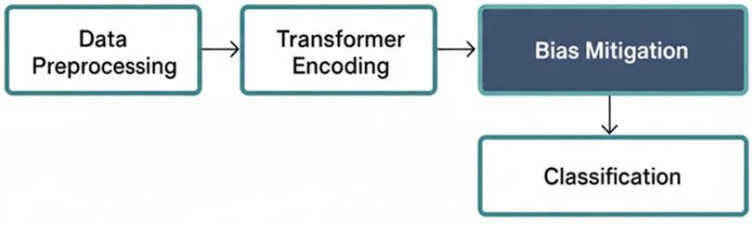

Fig. 1 Ethical Gender Identification Framework

A) Data Set

Effective gender identification from social media text hinges on the availability of high-quality labeled data. Although huge 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 applied 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

The preprocessing for data involved some linguistic normalization steps as following:

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

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:

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

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, thus improving transparency and accountability.

F) Ethical Safeguards

The entire experimental processes observed to guidelines for ethically aligned design. The study avoided collection of sensitive or identifiable user data. Model outputs were assessed for fairness, and potential misuse such as stereotyping or profiling was explicitly considered throughout analysis.

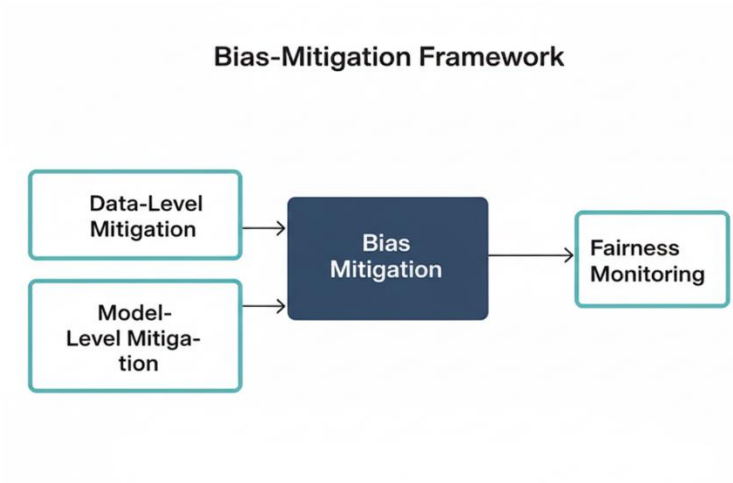

Fig. 2 Ethical Gender Identification Framework

IV. EXPERIMENTS

At this section, we define the experimental setup, model configurations, and evaluation metrics employed to evaluate the proposed transformer-based gender classification framework.

A) Experimental Setup

All experiments were conducted using the facebook/md gender bias dataset described in Section III part A. Implementation was execute out in Python 3.10 using the PyTorch deep-learning framework. In addition, the Hugging Face Transformers library was used for model

training and inference.

B) Baseline Models

In comparison, three traditional machine-learning classifiers were implemented using TF-IDF feature representations:

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

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

C) Evaluation Metrics

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

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

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

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

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

where TP is true positives, FP is false positives and FN represent false negatives.

The emphasis on F1-score and the difference between precision and recall over simple accuracy for classification tasks, particularly with potentially imbalanced datasets, reflects a sophisticated understanding of model assessment out there superficial performance. In gender classification, datasets may naturally exhibit imbalances in the representation of different genders. In such situations, 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 inclusive evaluation framework is critical for a reliable and fair assessment of the model's capabilities.

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

D) Training and Validation

Throughout training, initial stopping was employed based on validation loss to avoid 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. The entire transformer models substantially outperformed traditional 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 therefore establishes an approximate 7 % improvement in F1-score compared with the strongest classical baseline (SVM).

B) Bias-Mitigation Performance

Fairness metrics were assessed to measure the effectiveness of bias-mitigation strategies

introduced in Section III part 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 approve that the mitigation procedures substantially reduced gender specific prediction bias without degrading accuracy.

C) Explainability Analysis

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

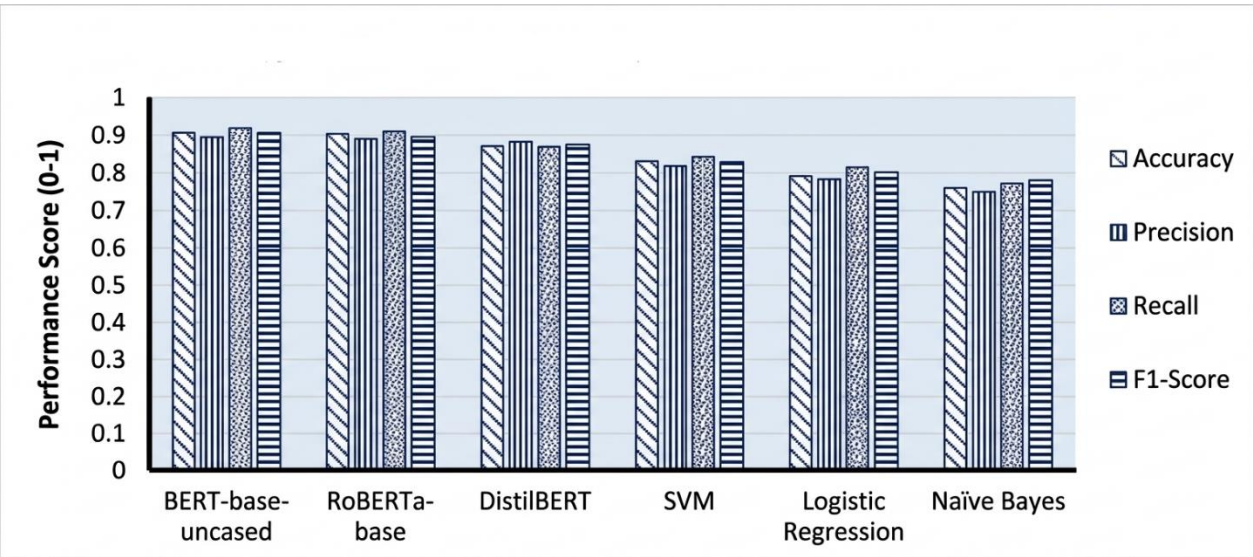

Fig. 3 Bar Chart of Model Metrics

VI. DISCUSSION

A) Interpretation of Findings

The experimental results presented in Section V approve that transformer-based models, particularly BERT-base-uncased, accomplish the highest accuracy and fairness balance for gender classification on social media text. This enhancement arises from the model's capacity to capture contextual and semantic dependencies through multi head self-attention, and 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, showing that ethical AI design principles can coexist with predictive

efficiency. Yet, slight variations in fairness metrics show that data imbalance and language stereotypes remain partially embedded in training corpora.

B) Comparison With Prior Work

Comparing 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 engaging contextual encoders [14]–[16] stated similar accuracy levels but often mitted systematic ethical assessment.

By integrating explainability and fairness monitoring, the current 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. Whereas automated inference of demographic attributes can improve personalization or sociolinguistic research, it also risks reinforcing gender stereotypes and exposing sensitive user information.

Hence, deployment should observe to privacy-preserving principles such as data anonymization, minimal attribute retention, and restricted downstream use. Furthermore, 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

Upcoming work should follow intersectional fairness evaluating gender identification in conjunction with race, age, or cultural background to guarantee inclusivity across demographic subgroups.

Methods 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 assurance that gender prediction tools remain equitable, transparent, and beneficial.

VII. CONCLUSION AND FUTURE WORK

This study offered a transformer-based framework for gender identification from social media network text that integrates ethical principles into model design and evaluation. By engaging BERT-base-uncased as the main 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 delivers transparency and accountability, key requirements for responsible deployment of demographic prediction models.

The results reveal 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. Nevertheless, residual bias and sensitivity to linguistic stereotypes specify that ethical oversight and fairness-aware optimization must remain integral parts of model development.

The Future work will focus on the following. Fist, extending the approach to multilingual and cross-platform data for broader generalization. Second, inspecting adversarial debiasing and federated learning strategies to enhance fairness and privacy preservation. Third, incorporating real-time explainability dashboards for model auditing and human-centered interpretability.

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

ACKNOWLEDGMENTS:

This research arises 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 thankfully acknowledge the group's collective expertise and the rigorous academic environment that enriched the development of this work.

REFERENCES:

[1] Ferguson, S. A., & Olechowski, A. (2023). Are We Equal Online?: An Investigation of Gendered Language Patterns and Message Engagement on Enterprise Communication Platforms. Proceedings of the ACM on Human-Computer Interaction, 7(CSCW2), 1-29.

[2] Blodgett, S. L., Barocas, S., Daumé Iii, H., & Wallach, H. (2020, July). Language (technology) is power: A critical survey of “bias” in NLP. In Proceedings of the 58th annual meeting of the association for computational linguistics (pp. 5454-5476).

[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) (pp. 4171-4186).

[4] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

[5] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. In Proceedings of the conference on fairness, accountability, and transparency (pp. 220-229).

[6] Burger, J. D., Henderson, J., Kim, G., & Zarrella, G. (2011, July). Discriminating gender on Twitter. In Proceedings of the 2011 conference on empirical methods in natural language processing (pp. 1301-1309).

[7] Argamon, S., Koppel, M., Fine, J., & Shimoni, A. R. (2003). Gender, genre, and writing style in formal written texts. To appear in Text, 23, 3.

[8] S. Kavuri, K., & Kavitha, M. (2020). A stylistic features based approach for author profiling. In Recent Trends in Communication and Intelligent Systems: Proceedings of ICRTCIS 2019 (pp. 185-193). Singapore: Springer Singapore. [9] Lim, K. W., & Buntine, W. (2014, November). Twitter opinion topic model: Extracting product opinions from tweets by leveraging hashtags and sentiment lexicon. In Proceedings of the 23rd ACM international conference on conference on information and knowledge management (pp. 1319-1328).

[10] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[11] Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

[12] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942.

[13] Nemani, P., Joel, Y. D., Vijay, P., & Liza, F. F. (2024). Gender bias in transformers: A comprehensive review of detection and mitigation strategies. Natural Language Processing Journal, 6, 100047.

[14] Alzahrani, E., & Jololian, L. (2021). How different text-preprocessing techniques using the BERT model affect the gender profiling of authors. arXiv preprint arXiv:2109.13890.

[15] Minot, J. R., Cheney, N., Maier, M., Elbers, D. C., Danforth, C. M., & Dodds, P. S. (2021). Interpretable bias mitigation for textual data: Reducing gender bias in patient notes while maintaining classification performance. arXiv preprint arXiv:2103.05841.

[16] Schwarzenberg, P., & Figueroa, A. (2023). Textual pre-trained models for gender identification across community question-answering members. IEEE Access, 11, 3983-3995.

[17] Shahbazi, N., Lin, Y., Asudeh, A., & Jagadish, H. V. (2023). Representation bias in data: A survey on identification and resolution techniques. ACM Computing Surveys, 55(13s), 1-39.

[18] IEEE, Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, 2nd ed., IEEE Standards Assoc., 2021.

[19] Crawford, K., & Paglen, T. (2021). Excavating AI: The politics of images in machine learning training sets. Ai & Society, 36(4), 1105-1116.