ISSN 2723-6471

Predicting Gender from Online Dating Self-Introductions Using Machine Learning, Deep Learning, and DistilBERT

Lionel F. Gonzalez Casanova[1], Wen-Ju Chen[2], Hsi-Sheng Wei[3,*]

[1]*Bachelor's Program in Smart Sustainable Development and Management, National Taipei University, Sanxia, Taiwan, R.O.C.*

[2,3]*Department of Social Work, National Taipei University, Sanxia, Taiwan, R.O.C.*

(Received: June 20, 2025; Revised: August 10, 2025; Accepted: November 25, 2025; Available online: January 14, 2026)

Abstract

This study investigates a novel approach to automated gender classification in online dating profiles by comparing models that span traditional machine learning, deep learning, and transformer-based architectures. The dataset consists of self-introduction essays from publicly accessible repositories and enriched with psychological features (LIWC), lexical features (bag-of-words), and contextual representations (raw text). The primary objective is to evaluate predictive performance, robustness, and computational cost across these modeling strategies and to assess their trade-offs. A comprehensive preprocessing pipeline was implemented, including missing-value handling, text cleaning, LIWC feature extraction, Bag-of-Words vectorization, one-hot encoding of categorical variables, and class-imbalance mitigation through random oversampling. Text augmentation using synonym replacement was subsequently applied to increase data diversity while maintaining realistic linguistic patterns. Stratified five-fold cross-validation was used for traditional models and LIWC-only deep learning experiments, and StratifiedKFold (k = 5) was applied to LIWC + BoW configurations to ensure balanced splits. DistilBERT was fine-tuned on raw essay data using an 80/20 train–test split under GPU memory and batch-size constraints. Across three runs, DistilBERT achieved an average testing accuracy of 91% ± 1%, with precision, recall, F1-score, and ROC–AUC indicating balanced performance. A GRU trained on LIWC+BoW features reached 88.62% ± 0.53% accuracy, offering competitive results at substantially lower computational cost. An MLP trained solely on LIWC features provided a stable and interpretable baseline. Confusion matrices showed balanced predictions between male and female classes, highlighting the importance of feature representation and model selection. Overall, the findings demonstrate clear trade-offs between computational demand and semantic modeling capability. These results contribute to ongoing research on gender identification and guide future work on fairness, robustness, and explainability in AI-assisted user profiling. The study also underscores practical benefits for automated analysis of unstructured text in social and psychological applications, while recognizing ethical considerations related to non-binary and gender-fluid individuals.

Keywords: Gender Classification, Online Dating, LIWC, Bag-of-Words, XGBoost, GRU, MLP, DistilBERT, Deep Learning, Natural Language Processing

1. Introduction

In today's digital environment, an increasing number of people rely on online platforms to form social connections and seek potential romantic partners. A recent survey of 1,278 Italian respondents reported that 22.46% were current users of dating applications, while 30.75% identified as former users, indicating that more than half of the participants had engaged with such platforms at some point [1]. Although online dating offers a convenient medium for interpersonal interaction, it also presents notable risks. In particular, the fabrication of user identities and the misrepresentation of demographic information are common on social networking and dating platforms [2]. These deceptive behaviors have become widely recognized as a form of cybercrime [3] and can lead to substantial financial losses and psychological harm for affected individuals [4]. These concerns highlight the potential value of incorporating basic identity verification features, such as gender confirmation, as part of broader efforts to reduce misuse and improve user safety.

Gender identification algorithms that can accurately infer gender from limited user data also have practical utility in several domains, including targeted advertising, personalized content delivery, and tailored digital services. Investigating this challenge, therefore, provides insights into both user protection and technological innovation, and it forms the central focus of this study. While multiple-choice questions or checkboxes can be intentionally manipulated

***Corresponding author: Hsi-Sheng Wei (hswei@mail.ntpu.edu.tw)**
DOI: https://doi.org/10.47738/jads.v7i1.979

by respondents, open-ended writing tasks may inadvertently reveal subtle linguistic patterns associated with aspects of an individual's identity, including gender. Although such patterns are not definitive indicators, prior research has documented statistically significant trends in language use that correlate with gender differences [5].

For example, a large-scale analysis of titles and abstracts from more than six million PubMed articles published between 2002 and 2017 found that male researchers tend to portray their work more positively than female researchers [6]. Another study examining 342 personal statements from medical students applying to a urology residency program reported that female applicants used more social and emotionally focused language [7]. In the social sciences, these linguistic variations have traditionally been examined using tools such as the Linguistic Inquiry and Word Count (LIWC). In recent years, gender identification based on language use has gained increasing attention in the data science community, driven by advances in computational methods including Bag-Of-Words (BoW) models, transformer-based representations, and improvements through recurrent architectures and data augmentation strategies.

Although a variety of computational approaches have been proposed, prior research often relies on a single modeling technique and seldom investigates the benefits of integrating multiple representations or architectures. Moreover, many existing studies evaluate models using long-form textual data, which does not reflect the short and informal style typical of online dating platforms. To address these gaps, this study examines gender classification using self-introduction essays from the OkCupid platform, aiming to improve predictive accuracy and robustness through a multi-method approach.

Specifically, the study integrates psychological features extracted through LIWC with lexical features derived from BoW representations to capture both structured and unstructured dimensions of user language. These features are evaluated using traditional machine learning models such as XGBoost and LightGBM, as well as deep learning architectures including Fully Connected Neural Networks (FCNN), Gated Recurrent Units (GRU), and Multilayer Perceptrons (MLP), each suited to different aspects of feature representation. In parallel, a transformer-based model, DistilBERT, is fine-tuned directly on raw essay text to extract contextual semantic information. To further improve generalization, data augmentation techniques, such as synonym replacement, are applied during training.

By developing and evaluating multiple modeling pipelines across different input modalities and architectures, this study provides a more nuanced and practical framework for gender identification, particularly for short-form, user-generated content common on online dating platforms. The rest of this paper is organized as follows. Section 2 reviews the existing literature relevant to this work. Section 3 describes the datasets, methods, models, and experimental procedures. Section 4 presents the results and discussion. Section 5 concludes the paper and outlines directions for future research.

2. Related Work

2.1. Linguistic Inquiry and Word Count (LIWC)

LIWC is a widely used text analysis tool designed to quantify psychologically meaningful word categories [8]. It extracts features related to emotions, cognitive processes, and social relationships by classifying words into predefined linguistic categories. LIWC has been extensively applied in sociolinguistic and psychological research to investigate gendered language patterns. For example, analyses of large text corpora consistently show that women tend to use more language associated with psychological and social processes, whereas men more frequently reference impersonal and object-focused content [9]. In professional contexts such as letters of recommendation in emergency medicine, communal descriptors, such as those related to compassion and teamwork, appear more often for female applicants [10]. On social media platforms, self-identified women tend to emphasize personal relationships, while men's language contains higher levels of swearing and argumentative tone [11].

LIWC's ability to capture stylistic and psychological nuances makes it particularly valuable for analyzing online dating profiles, where users often express intentions, values, and emotional states in concise self-descriptions. Prior studies have also demonstrated that LIWC features can be effectively combined with machine learning algorithms for gender identification.

For example, the authors in [12] collected data from more than 9,000 bloggers across five languages and aggregated all posts written by each individual into a single document, typically between 1,000 and 2,000 words. Language-specific LIWC dictionaries were used to extract linguistic features, and LIWC-based models achieved gender classification accuracies ranging from 73.8% to 79.6%, depending on the language. In another study, [13] analyzed 18.5 million tweets from 11,155 users in Nigeria and computed LIWC outputs alongside unigram and hashtag features. The authors reported a 72.02% accuracy rate using LIWC alone, with additional gains observed when other textual features were incorporated.

These findings underscore the value of LIWC as a psychologically informed feature set, especially when integrated with complementary linguistic representations. In this study, LIWC is used to extract emotion-, cognition-, and social-related features from brief self-introduction essays, enabling a deeper exploration of linguistic markers relevant to gender identification in online dating environments.

2.2. Bag-of-Words

The bag-of-words (BoW) approach is a foundational method in natural language processing that represents text through word frequencies, disregarding word order but capturing lexical richness and distributional patterns. Despite its simplicity, BoW has been widely used in gender identification research due to its effectiveness, scalability, and interpretability. It allows models to detect statistically meaningful word-choice patterns across gender groups, including lexical cues that may be overlooked by psychologically oriented tools such as LIWC. Empirical studies have demonstrated the strength of BoW-based models. For example, the authors in [14] achieved 93.86% accuracy using BoW with logistic regression in a dataset containing ten authors and 500 articles per author. In [15], a multilingual SMS corpus combined BoW features with psycholinguistic measures, resulting in an accuracy of 80.29%. More recent research has integrated BoW with advanced algorithms such as LSTM networks and support vector machines, producing similarly competitive outcomes [16], [17].

Compared with LIWC, BoW captures a broader range of lexical and syntactic variation without relying on predefined psychological categories. This flexibility is particularly important for modeling natural language variation in informal, short-form texts such as dating profiles. For instance, [18] reported an 85.1% accuracy by combining BoW with stylometric and psycholinguistic cues in professional communication datasets. Likewise, [19] identified gender-associated word-use patterns, such as differences in function word or pronoun frequency, in WhatsApp messages. Together, these findings underscore the role of BoW as a valuable complement to LIWC in gender classification tasks. Whereas LIWC offers depth through psychologically interpretable features, BoW provides breadth through statistical sensitivity to lexical patterns. The present study combines both approaches, leveraging LIWC's psychologically grounded categories alongside BoW's expansive lexical coverage. This dual representation is expected to support more robust and generalizable models for gender identification, particularly when applied to brief, self-authored texts typical of online dating platforms.

2.3. Transformer Models

In recent years, transformer-based architectures have been increasingly applied to gender identification tasks because of their strong ability to model sequential data and capture subtle linguistic patterns. These models use attention mechanisms and contextual embeddings to interpret nuanced language cues that may signal gender. The authors in [20] demonstrated the effectiveness of such models in analyzing social media text, showing that transformer architectures such as DistilBERT outperform classical algorithms and earlier machine learning techniques, including Support Vector Machines (SVMs), in classification accuracy. Their results suggest that transformers are particularly well suited for processing informal, short, and context-dependent content, which characterizes much of contemporary online communication.

A growing body of empirical research further supports this perspective. For example, [21] applied transformer models to a community question-answering dataset from Yahoo! Answers that included more than 548,000 user profiles. The study found that full questions and answers produced the highest predictive accuracy, while brief self-descriptions offered limited utility because of their sparse availability. In another study, [22] proposed a multimodal transformer-based method that combined textual features from tweets with visual information extracted from profile and post images. Their model achieved an accuracy of 88.11% on a Kaggle dataset, outperforming single-modality models and

demonstrating the value of integrating textual and visual signals for gender prediction. Taken together, these studies illustrate the increasing sophistication and effectiveness of transformer-based methods in this field.

3. Methodology

This section describes the methods, datasets, models, and procedures used in this research. The discussion begins with an overview of the dataset and problem context, followed by detailed steps for data preprocessing and transformation. The section concludes with a description of the classification models employed.

3.1. Dataset and Problem Overview

Online dating platforms have become increasingly popular spaces for people seeking romantic connections, producing large volumes of self-authored text. These short and informal narratives offer unique insights into users' behaviors, preferences, and identities. Applying NLP to this content provides opportunities to better understand online communication and to support applications such as matchmaking, content moderation, and identity verification. This study focuses on gender identification in self-introduction essays from the OkCupid platform and investigates how linguistic features may reveal underlying identity cues in short-form, user-generated content.

To address this objective, the study adopts a multi-method approach that combines three complementary techniques: LIWC, BoW, and transformer-based language models. LIWC enables the extraction of psychologically meaningful features related to emotional tone, social orientation, and cognitive style. BoW offers a statistical view of lexical patterns by highlighting frequent word-use tendencies across gender groups. To complement these methods, contextual embeddings from transformer models such as DistilBERT are incorporated because they are well suited for capturing nuanced language use. These embeddings are further explored using sequential architectures such as LSTMs and neural baselines including MLP and FCNN, combined with data augmentation strategies to improve generalization. By integrating these approaches, the study aims to develop a more accurate and flexible gender classification framework that is tailored to the distinctive style and brevity of online dating text.

3.2. Data Preprocessing

OkCupid is a valuable source of data because of its high level of user engagement and the richness of the information collected. The platform employs a matching algorithm that relies on user responses to a wide range of personal and preference-based questions. These questions include demographic information such as age, gender, sexual orientation, and location, along with the importance users assign to their answers and the acceptability of potential matches' responses.

The original dataset used for this study contains information from 59,946 active users. Using the LIWC-22 software, we extracted 122 linguistic features from users' essay responses. To expand the sample size and strengthen model training, we created an augmented version of the dataset through synonym replacement in the user essays. This procedure generated semantically consistent but lexically varied texts, which preserved the natural writing style of dating profiles while increasing linguistic diversity and reducing the risk of overfitting. The augmented dataset contains 119,868 user instances, consisting of 71,646 male and 48,222 female entries, as shown in table 1. Importantly, the augmentation process preserved the class imbalance present in the original dataset (35,823 male and 24,111 female instances), ensuring that evaluation results remained representative of real-world distributions.

Throughout this study, both LIWC features and Bag-of-Words (BoW) representations were used to analyze textual data from user essays. A range of classification approaches was applied, including traditional machine learning models, neural network architectures, and a transformer-based model, in order to provide a comprehensive analysis of gender classification within the OkCupid dataset.

Table 1. Label Counts in the Original and Augmented Datasets

Label	Original Count	Augmented Count
Male	35,823	71,646
Female	24,111	48,222

3.2.1. Text Data Augmentation

To increase the size and diversity of the dataset, text data augmentation was performed using the nlpaug library. Synonym replacement was applied with the SynonymAug augmenter to generate additional variations of each essay. Although the augmented dataset initially retained the class imbalance of the original data, oversampling was later applied during training to reduce bias toward the minority class and improve model stability. This two-part strategy ensured that augmentation preserved the realism of user writing, while oversampling provided sufficient representation of the minority class during model fitting without disproportionately increasing the number of synthetic samples.

To avoid data leakage between training and testing sets, all augmented samples were generated only after the dataset had been partitioned. Augmentation was applied exclusively to the training folds during cross-validation or to the training portion in fixed-split experiments, such as those involving DistilBERT. This procedure ensured that no paraphrased or near-duplicate versions of an essay appeared in both the training and testing sets, thereby preserving the validity of the evaluation process. Each original essay was paraphrased approximately three times on average, resulting in a threefold increase in dataset size. However, the exact number of augmented instances produced per essay was not recorded in the current workflow, which limits reproducibility. Future revisions will include explicit logging of per-sample augmentation counts to improve methodological transparency.

3.3. Data Processing Pipeline

The data processing pipeline was designed to handle numerical, categorical, and textual inputs in preparation for classical machine learning, deep learning, and transformer-based models. In this study, the terms classical and traditional are used interchangeably. The objective was to ensure that inputs remained consistent and compatible across different model architectures rather than enforce a single fixed structure. For the numerical LIWC features, missing values were imputed using the mean of each column in order to preserve the original feature distribution. For the textual essays, a more practical strategy was used: missing entries were replaced with empty strings so that tokenizers and vectorizers could operate without error. This operation can be written as

$$X_{essays,i} \rightarrow \{X_{essays,i}, if\, X_{essays,i} ispresent\ ``\ ", if\, X_{essays,i} ismissing \quad (1)$$

Here, each essay entry $X_{essays,i}$ is checked individually. If content is present, the value is retained. If the entry is missing, it is replaced with an empty string (“ ”). This simple rule maintains compatibility with the tokenization and vectorization stages and avoids unintended transformations of the textual data.

After missing data were addressed, the essays were processed with the “CountVectorizer” to produce a sparse bag-of-words (BoW) matrix. The vocabulary was limited to the 1,000 most frequent terms in the corpus to balance information richness with computational efficiency. The resulting BoW vectors were concatenated with normalized LIWC features when training deep learning models. Although categorical fields such as orientation and relationship status were encoded using one-hot encoding to retain pipeline flexibility, these variables were ultimately not used in the neural or transformer-based experiments.

To address class imbalance, particularly the skew in gender distribution, the “RandomOverSampler” was applied to duplicate samples from the minority class. This oversampling step, combined with modest data augmentation, reduced bias toward the majority class and contributed to more stable decision boundaries. Numerical features were standardized with the “StandardScaler” (zero mean, unit variance), which improved convergence for architectures that are sensitive to input scale, including MLPs, GRUs, and LSTMs.

Evaluation strategies varied by model family. Classical machine learning approaches and LIWC-only neural models were assessed using stratified five-fold cross-validation, which provided reliable performance estimates with balanced label representation. The same stratified five-fold procedure was applied to deep learning models trained with LIWC and BoW features, primarily to reduce variance across folds. Transformer-based models required a different strategy because cross-validation was not feasible under the available GPU memory and runtime constraints. For DistilBERT, an 80/20 stratified train–test split was adopted, and several independent trials were conducted using different random seeds. Each run used a batch size of four and was trained for five epochs, which fit within the available GPU budget.

Although these mixed evaluation strategies were effective in practice, they introduce a level of inconsistency that makes direct comparisons across all model families less straightforward. Future work conducted in a distributed or cloud-based environment would make it possible to apply consistent stratified cross-validation to transformer models as well, thereby improving comparability and methodological coherence across the full experimental pipeline.

3.4. Classification Models

This study designs and evaluates two traditional machine learning models, five deep learning models, and one transformer-based model to predict user gender from dating profiles. Each model offers distinct advantages for processing textual and linguistic features, allowing for a comprehensive assessment of different computational approaches.

3.4.1. Baseline Traditional ML Models

This study evaluated a diverse set of models to predict user gender from dating profile text, with the goal of comparing how different model families, including traditional machine learning, deep learning, and transformer-based architectures, process linguistic and textual features. For the traditional machine learning baseline, we selected extreme gradient boosting (XGBoost) and light gradient boosting machine (LightGBM). XGBoost served as a strong reference point because of its robust performance on tabular data and its ability to model nonlinear feature interactions, even in the presence of label imbalance. LightGBM, which uses a histogram-based training strategy, provided an efficient alternative that scales effectively to high-dimensional sparse inputs such as bag-of-words features. Together, these models offered a useful benchmark for assessing how well simpler, feature-engineered approaches perform before introducing more complex neural and transformer-based architectures.

3.4.2. Deep Learning-Based Models

The deep learning group consisted of five architectures: a FCNN, a Long Short-Term Memory (LSTM) network, a bidirectional LSTM (biLSTM), a GRU, and a MLP. The FCNN served as a baseline dense model that treats input features independently. It was applied to both the LIWC-only dataset and the combined LIWC and BoW feature set to assess the extent to which sequential information contributed to performance. The LSTM, which is designed to retain long-range dependencies in text, helped address limitations of standard recurrent networks such as vanishing gradients. The biLSTM extended this capability by processing each essay in both forward and backward directions, allowing the model to capture contextual cues that may be missed by a single-direction network. The GRU, a streamlined variant of the LSTM, provided faster training while maintaining strong performance on text classification tasks. The MLP was included to evaluate how a traditional feedforward architecture behaved when trained on structured LIWC features as well as on the combined LIWC and BoW representation.

For the transformer-based category, the study used DistilBERT, a compact version of BERT that retains much of the original model's language understanding ability while offering faster and more resource-efficient operation. Because DistilBERT is pretrained on large general-purpose corpora, it can extract semantic and contextual information directly from raw essays without relying on handcrafted linguistic features. This capability made DistilBERT particularly effective for identifying subtle patterns in writing style and content that may not be explicitly captured by LIWC or BoW representations.

To ensure transparency and reproducibility, the main training parameters for DistilBERT are summarized in table 2. The selected configuration balances computational feasibility with model robustness and reflects the practical constraints involved in fine-tuning transformer-based architectures on limited GPU resources. The detailed setup supports reproducibility and comparability across model families.

Table 2. Key DistilBERT Fine-Tuning Parameters

Parameter	Description
Pretrained mode	DistilBERT-base-uncased
Optimizer	Adam
Learning rate	2 × 10−5

 ISSN 2723-6471

Loss function	Binary Cross-Entropy
Batch size	4
Epochs	5
Pooling layer	Global Max Pooling
Hidden layer	Dense (64 units, ReLU activation)
Output layer	Dense (1 unit, Sigmoid activation)
Class-imbalance handling	Balanced class weights
Early stopping	Not applied (for comparability)
Validation strategy	Fixed validation split (no stratified CV due to computational cost)

All models were implemented using TensorFlow v2.10.0, pandas v3.9.19, and scikit-learn v1.6.1. Experiments were run on a local machine equipped with an NVIDIA GPU with 8 GB of VRAM, which supported efficient training for both deep learning and transformer-based models. Preprocessing steps, including tokenization, vectorization, and scaling, were applied consistently across all architectures.

Figure 1 presents the overall research workflow, which includes data preprocessing, feature extraction, model training, and evaluation. This visual summary clarifies how different feature representations, such as LIWC, BoW, and transformer embeddings, are incorporated into the experimental pipeline.

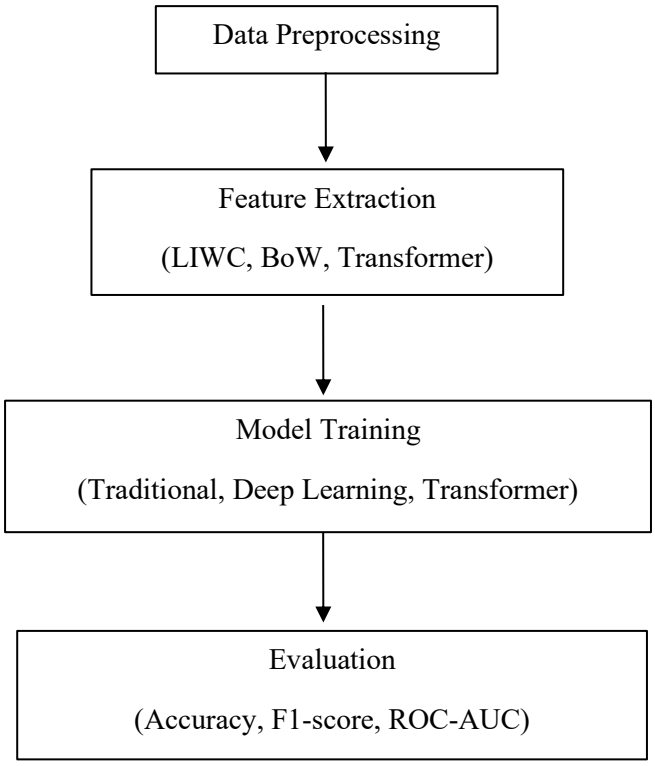

Figure 1. Research workflow from preprocessing to evaluation, integrating all feature representations and model types.

4. Results and Discussions

To comprehensively assess model performance on the augmented OkCupid dataset, we evaluated a range of traditional machine learning models, deep learning architectures, and a transformer-based model (DistilBERT). The evaluation considered precision, recall, F1-score, accuracy, and ROC-AUC across different configurations and feature combinations. To ensure robustness, results are reported with mean and standard deviation values over multiple runs

where applicable. Computational costs, including training time, inference latency, and GPU memory usage, were also measured to highlight trade-offs between performance, efficiency, and scalability.

Compared to prior studies, the hybrid approach presented here extends LIWC-based research, which has traditionally relied on linear or shallow models, by integrating psychologically informed features with bag-of-words representations and deep learning architectures such as MLP. While transformer-based models captured semantic nuances effectively, they required substantial computational resources. The integrated methodology used in this study provides strong performance at lower computational cost while maintaining interpretability, helping to bridge the gap between human-understandable linguistic cues and predictive accuracy.

4.1. Evaluation of LIWC-CV

The first set of experiments used LIWC features exclusively and employed five-fold cross-validation. As shown in table 3, traditional models such as XGBoost achieved substantially better performance than FCNN and LGBM. Among all models, the MLP achieved the highest predictive performance, with precision, recall, and F1-score values of 0.899 ± 0.004 and an accuracy of 89.88.

Table 3. LIWC-CV Testing Model Results (Average ± Std Over Five Folds)

Model	Precision	Recall	F1-score	Accuracy
FCNN	0.703 ± 0.006	0.703 ± 0.006	0.702 ± 0.006	69.83% ± 0.61
XGBoost	0.797 ± 0.005	0.797 ± 0.005	0.797 ± 0.005	79.69% ± 0.46
LGBM	0.719 ± 0.004	0.719 ± 0.004	0.719 ± 0.004	71.83% ± 0.42
MLP	0.899 ± 0.004	0.899 ± 0.004	0.899 ± 0.004	89.88% ± 0.36

Table 4 summarizes the corresponding computational costs. Tree-based models such as XGBoost and LGBM showed very short training times per fold (1.51 and 1.31 seconds, respectively), reflecting their efficiency when applied to low-dimensional tabular inputs. In contrast, deep learning models required substantially more time. The MLP averaged 0.82 seconds per epoch and 146.36 seconds per fold because of the iterative weight updates performed across multiple layers. The FCNN was faster, taking 33 seconds per fold, which is consistent with its shallower architecture. These findings illustrate a clear trade-off between predictive performance and computational cost. Although the MLP achieved the highest accuracy and F1-score, XGBoost and LGBM delivered competitive performance with significantly greater computational efficiency.

Table 4. LIWC-CV Model Training Times (Epoch and Total Clock Comparisons)

Model	Avg Epoch Time (s)	Avg Time Per Fold (s)	Approx. Total (Five Folds)
FCNN	33 ± 1	165 ± 5	825s
XGBoost	–	1.51 ± 0.02	7.55s
LGBM	–	1.31 ± 0.02	6.55s
MLP	0.82 ± 0.03	146.36 ± 3	731.8s

4.2. Evaluation on LIWC + BoW with Cross-Validation

In the second stage, the LIWC features were augmented with a Bag-of-Words (BoW) representation derived from user essays. This combined feature set was evaluated across five deep learning models using stratified cross-validation, with ten training epochs and a reduced batch size of 16 (reduced from 32) to accommodate GPU memory limitations.

As shown in table 5, all five models outperformed their counterparts trained on LIWC-only features. The gated recurrent unit (GRU) achieved the highest accuracy (88.62% ± 0.53) and the highest F1-score (0.886 ± 0.004), followed closely by the bidirectional LSTM (biLSTM) and the LSTM. The multilayer perceptron (MLP) also performed competitively, with an average accuracy of 85.30% ± 0.55.

Table 6 summarizes the computational costs. The recurrent models (LSTM, biLSTM, and GRU) required substantially longer training times per epoch, approximately 82 ± 2 seconds, while feedforward architectures such as the FCNN and MLP required about 33 to 34 seconds per epoch. This comparison highlights a clear trade-off between predictive performance and computational efficiency. Overall, the results demonstrate the effectiveness of combining psychological features (LIWC) with text-based representations (BoW), while also emphasizing the need to consider model complexity and training time.

Table 5. LIWC-BoW-CV Testing Model Results (Mean ± Std Over Five Folds)

Model	Precision	Recall	F1-score	Accuracy
FCNN	0.859 ± 0.005	0.861 ± 0.004	0.860 ± 0.004	86.15% ± 0.42%
LSTM	0.884 ± 0.004	0.883 ± 0.004	0.884 ± 0.004	88.42% ± 0.55%
biLSTM	0.886 ± 0.004	0.885 ± 0.004	0.885 ± 0.004	88.50% ± 0.54%
GRU	0.886 ± 0.004	0.886 ± 0.004	0.886 ± 0.004	88.62% ± 0.53%
MLP	0.854 ± 0.005	0.853 ± 0.005	0.853 ± 0.006	85.30% ± 0.55%

Table 6. LIWC-BoW-CV Model Training Time Per Epoch

Model	Avg Epoch Time (s)	Approx. Total Training Time (Ten Epochs, Five Folds)
FCNN	33 ± 1	1,650s
LSTM	82 ± 2	4,100s
biLSTM	82 ± 2	4,100s
GRU	82 ± 2	4,100s
MLP	34 ± 1	1,700s

4.3. Evaluation of DistilBERT

For the final stage, DistilBERT, a transformer-based model, was applied using only the raw essay texts and no handcrafted features. Because transformer models require substantial memory, the training configuration used a batch size of four and five training epochs. To ensure stability and reproducibility, three independent runs were conducted with different random seeds, and the mean and standard deviation of the results are reported.

The performance results in table 7 show that DistilBERT achieved an average accuracy of 98.7% ± 0.5 on the training set and 91.0% ± 1.0 on the testing set across the three runs. The model also produced highly consistent ROC-AUC scores, with mean values of 0.9994 ± 0.0003 for training and 0.966 ± 0.003 for testing. These results confirm both the robustness of the model and its strong discrimination capability. The relatively small variances further indicate that the performance is not attributable to random initialization.

Overall, these findings demonstrate the strength of transformer-based models in capturing contextual information and semantic relationships in free-form text. Even under constrained conditions such as a small batch size, DistilBERT remains highly effective at detecting subtle semantic cues within user-generated essays.

Table 7. DistilBERT Performance Over Three Runs (Mean ± STD)

Phase	Precision	Recall	F1-score	Accuracy	ROC-AUC
Training	0.987 ± 0.005	0.987 ± 0.005	0.987 ± 0.005	98.7% ± 0.5	0.9994 ± 0.0003
Testing	0.907 ± 0.015	0.910 ± 0.010	0.907 ± 0.015	91.0% ± 1.0	0.966 ± 0.003

In addition to predictive performance, we evaluated the computational cost of fine-tuning and inference on the testing sets. As shown in table 8, fine-tuning DistilBERT for five epochs required approximately 111.5 minutes in total (about 1,340 seconds per epoch). Converting outputs to binary predictions required 132 seconds, and the full testing procedure

per run took 35 seconds. All timings were measured on an NVIDIA GeForce RTX 3060 Ti GPU. Reporting these computational costs helps illustrate the trade-offs between model performance and resource requirements.

Table 8. Computational Cost of DistilBERT Over Three Runs

Phase	Epochs/Batches	Time (s)	Hardware
Training	Five epochs	1340 ± 2 per epoch	NVIDIA GeForce RTX 3060 Ti
Binary Predictions	One full run (2,944 samples)	132 ± 1	NVIDIA GeForce RTX 3060 Ti
Testing	One full run (736 batches)	35 ± 1	NVIDIA GeForce RTX 3060 Ti

4.4. Summary of Observations

Across the different model settings, the transformer-based approach consistently produced the strongest results when applied directly to the raw text. In contrast, both traditional models and the deep learning architectures showed clear improvements when bag-of-words representations were combined with LIWC features. For the MLP in particular, an examination of the learned weights revealed the ten most influential BoW features for each gender class, which helped identify the vocabulary items that contributed most to the final predictions. Limiting the vocabulary to the 1,000 most frequent terms helped reduce noise from rare words, although this cutoff likely excluded some subtle gender-related cues, as summarized in table 9.

Table 9. Top Ten Discriminative BoW Words Per Class (MLP + LIWC + BoW).

Word/Token ID	Class 1 (Male) Weight	Class 0 (Female) Weight
036063068785690475	0.1139	-
03997887859792479	0.1124	-
038442017187226486	0.1101	-
027077104539492282	0.1076	-
041088903324556725	0.1047	-
03769429003479613	0.1039	-
024801899909367677	0.1013	-
04108636898104777	0.0995	-
04038287991908424	0.0985	-
024324807640241097	0.0973	-
0309018640623947	-	–0.1097
03887718572475974	-	–0.1053
02477061731067412	-	–0.1045
03610205139637079	-	–0.1041
04050354369590088	-	–0.1028
0415520162281991	-	–0.1018
024264528851192017	-	–0.1014
03770246454091266	-	–0.1009
040353969019539304	-	–0.0976
04028838211324002	-	–0.0970

Using stratified cross-validation helped maintain balanced evaluation splits and reduced potential bias in performance estimates. Hyperparameters such as batch size and number of epochs were adjusted to reflect the computational

constraints of each model family, with transformer fine-tuning requiring the most conservative settings. DistilBERT, although a lighter version of BERT, remained substantially more computationally demanding than the tree-based baselines. Training the model for five epochs required approximately 111.5 minutes in total (about 1,340 seconds per epoch), while inference on the test sets required about 132 seconds to generate predictions and an additional 35 seconds for full evaluation. All computational times were recorded on an NVIDIA GeForce RTX 3060 Ti GPU with 8 GB of VRAM. These results highlight a familiar trade-off in natural language processing: transformer models provide richer semantic understanding but require considerably more time and resources.

Patterns in model errors indicated that approximately 14% of the validation essays were misclassified. Most of these errors originated from extremely short entries or neutral, low-information essays in which the models had very little linguistic content to analyze. As shown in table 10, placeholder or single-word submissions, many of which appeared as 0.0 after preprocessing, accounted for a substantial portion of these misclassifications. The statistical reliability of the MLP under five-fold stratified cross-validation further supported its stability. The model achieved an average accuracy of 0.858, with fold-level F1-scores ranging from 0.852 to 0.863. This narrow performance range suggests consistent behavior across folds and reinforces the robustness of combining LIWC features with bag-of-words representations.

Table 10. Examples of Longest Misclassified Essays. Only a Subset Is Shown.

Essay	True Label	Predicted Label	Length (words)
0.0	1	0	1
0.0	1	0	1
0.0	0	1	1
0.0	1	0	1
0.0	0	1	1

4.5. Performance Metrics Considered

To evaluate model performance in a consistent and informative manner, several classification metrics were used, each highlighting a different aspect of model behavior. The confusion matrix provided a direct view of model errors by listing the counts of true positives, true negatives, false positives, and false negatives. This representation made it easier to see which types of mistakes occurred most frequently for each class. Recall (also called sensitivity) was used to measure how well a model identified all relevant instances. It is defined as the number of true positives divided by the sum of true positives and false negatives. A high recall value indicates that the model rarely misses relevant cases, which is important in settings where false negatives carry significant cost. Precision complemented this metric by indicating how reliable the positive predictions were. It is calculated as true positives divided by the sum of true positives and false positives, so higher values reflect fewer false alarms. To capture both perspectives at the same time, the F1-score was included. This metric, the harmonic mean of precision and recall, is especially useful when class distributions are not perfectly balanced. Although accuracy summarizes the proportion of correctly classified samples overall, it can be misleading when one class is dominant. To account for a model's performance across all possible thresholds, the ROC-AUC score was also reported as a measure of discrimination between classes independent of any single decision cutoff. In addition to predictive performance, computational requirements were documented, including both training and inference times measured on an NVIDIA GeForce RTX 3060 Ti GPU. Reporting these values highlights the trade-offs between accuracy and efficiency, which is especially important for resource-intensive models such as DistilBERT.

From the full set of evaluated models, three representative examples were selected to illustrate performance across confusion matrices, ROC-AUC scores (when applicable), and computational costs. The first example was a Multilayer Perceptron (MLP) trained exclusively on LIWC features and evaluated using stratified five-fold cross-validation. Because it relies entirely on structured linguistic indicators, this model served as a useful baseline. Performance was

reported as fold-level averages with standard deviations, reflecting the model's consistently strong precision and recall and demonstrating that psychologically motivated features alone can yield reliable predictions.

The second representative model was a GRU trained on the combined LIWC and Bag-of-Words inputs, which incorporated both structured and lexical–sequential information from user essays. Using stratified five-fold cross-validation, the GRU achieved competitive accuracy while maintaining reasonable computational efficiency. Results were summarized using average and standard deviation statistics, along with recorded training and inference times.

The third model was DistilBERT, which was fine-tuned on the raw essay text using an 80/20 stratified train–test split. Starting from the distilbert-base-uncased pretrained checkpoint, the model was optimized using Adam with a learning rate of $2 * 10^{-5}$ and binary cross-entropy loss. A batch size of four was used for five epochs, followed by a global max-pooling layer, a hidden dense layer with 64 ReLU-activated units, and a sigmoid output layer for binary classification. Class weights were applied to address label imbalance, and early stopping was intentionally not used to maintain consistency with the other architectures. Despite these constraints, DistilBERT delivered the strongest overall performance, demonstrating the advantage of transformer architectures in extracting nuanced semantic cues from unstructured dating-profile text.

4.5.1. Confusion Matrix Analysis

Figure 2 presents the confusion matrix for the MLP model trained on LIWC features using stratified five-fold cross-validation. The model correctly identified 63,985 female users (true negatives) and 64,808 male users (true positives). It misclassified 7,661 female users as male (false negatives) and 6,838 male users as female (false positives). These results indicate a strong balance in classification performance across both classes.

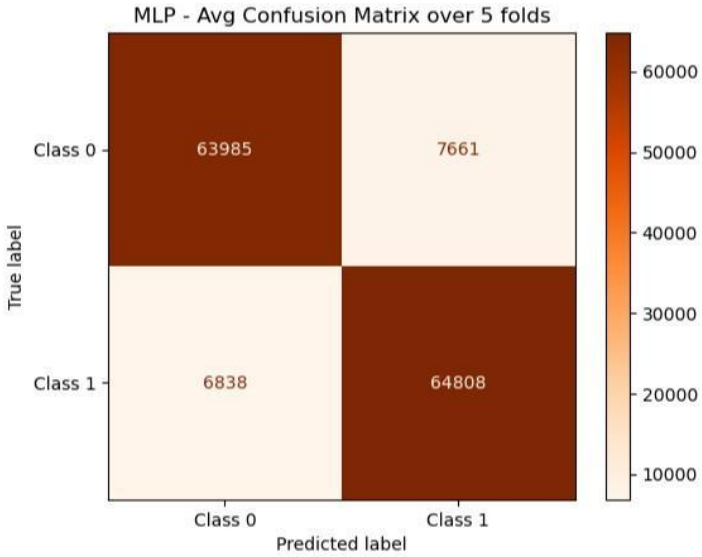

Figure 2. Confusion matrix: MLP model using LIWC and stratified five-fold CV. Class 0 = female, class 1 = male.

Figure 3 presents the confusion matrix for the GRU model trained on the combined LIWC and BoW feature set, also evaluated using stratified five-fold cross-validation. The model correctly predicted 62,601 female instances (true negatives) and 63,913 male instances (true positives). It misclassified 9,047 female users as male (false negatives) and 7,731 male users as female (false positives), demonstrating effective integration of linguistic and textual features in its predictions.

For DistilBERT, which was trained using only the essay texts and did not include LIWC or BoW features, evaluation was conducted using an 80/20 train–test split. Despite the small batch size of four imposed by GPU memory constraints, the model demonstrated strong performance on the testing set, achieving an average accuracy of 91.0% ± 1.0 across three independent runs. The ROC-AUC score was also consistent, with a mean of 0.966 ± 0.003, indicating stable discrimination capability.

 ISSN 2723-6471

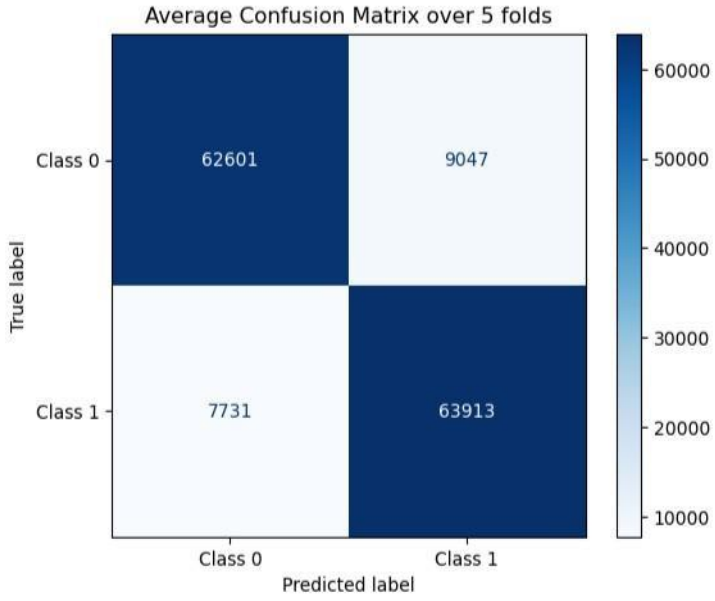

Figure 3. Confusion matrix: GRU model using LIWC+BoW and stratified five-fold CV. Class 0 = female, class 1 = male.

Figure 4 presents the confusion matrix for the testing set from Run 3 as a representative example. In this run, the model correctly classified 8,582 female users (true negatives) and 13,059 male users (true positives). It misclassified 974 female users as male (false positives) and 934 male users as female (false negatives). This figure illustrates the model's ability to capture semantic nuances from essay texts while maintaining high classification accuracy.

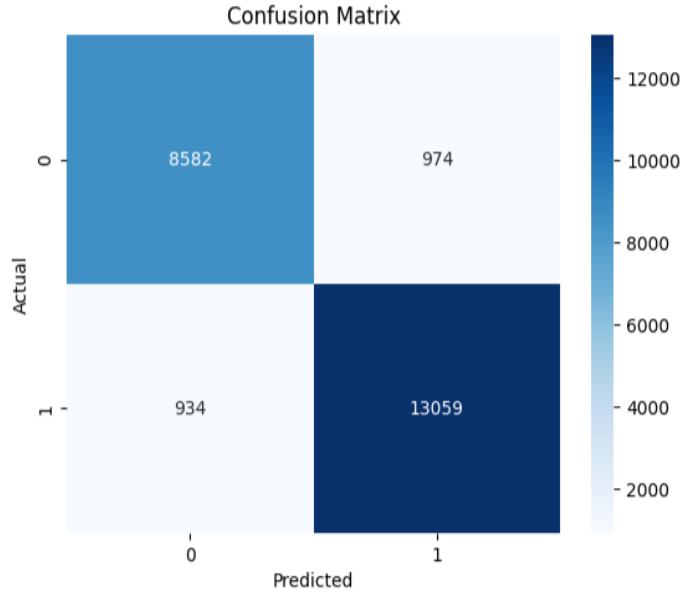

Figure 4. Confusion matrix: DistilBERT model using only essays (train–test split). Class 0 = female, class 1 = male.

4.5.2. Summary of Findings

LIWC features provided a psychologically grounded starting point and produced interpretable baselines, but their dictionary-driven design creates inherent limitations. Because the vocabulary is fixed, LIWC cannot easily capture evolving slang, informal phrasing, or platform-specific language patterns that appear frequently in online dating environments, particularly among younger users. This constraint reduces linguistic coverage compared to models that learn directly from raw text.

Both the MLP and the GRU benefited from stratified cross-validation, which strengthened the stability of their evaluation scores and reduced the influence of sampling variability. The GRU model, trained on a combination of LIWC and bag-of-words features, achieved an F1-score of approximately 88.6%. This performance surpassed that of the LIWC-only MLP and demonstrated that combining structured psychological features with data-driven lexical signals improves robustness. DistilBERT delivered the strongest generalization performance across three independent runs, achieving approximately 91% accuracy. These results highlight the value of transformer-based models in capturing fine-grained semantics and contextual cues, even when trained under constraints such as small batch sizes and limited GPU memory.

The bag-of-words representation was intentionally restricted to the top 1,000 most frequent tokens to limit dimensionality and sparsity. When combined with LIWC and regularized deep learning architectures, this approach helped control overfitting while preserving a degree of interpretability. Among the neural models, the GRU, biLSTM, and LSTM consistently outperformed feedforward networks such as the FCNN and MLP. This outcome aligns with expectations because recurrent architectures are better suited for modeling dependencies within user-generated text.

Approximately 14% of validation essays were misclassified. Most errors occurred in extremely short essays or placeholder text that offered minimal linguistic content, limiting the models' ability to infer gender-associated patterns. This suggests that essay length, richness of expression, and narrative tone are influential factors and should be examined more closely in future work.

The confusion matrices showed relatively balanced precision and recall for both male and female classes, indicating no substantial prediction bias. Hybrid combinations involving transformer models, such as LIWC with DistilBERT or BoW with DistilBERT, were not explored in this study because of computational constraints and the additional engineering required to merge numerical features with contextual embeddings. Future research could investigate these hybrid designs to determine whether structured linguistic features provide complementary value to the semantic representations learned by transformers.

Ethical considerations are also important. LIWC's fixed dictionary and binary framing may not fully capture the linguistic expressions of non-binary or gender-fluid users. This limitation underscores the broader need for more inclusive lexicons and modeling strategies when examining gender inference in digital contexts.

4.6. Ethical Considerations

The model demonstrates strong predictive accuracy, but several ethical concerns require careful attention, particularly with respect to potential misuse and broader societal impacts. The dataset used in this study, collected more than a decade ago, adheres to a binary gender framework that recognizes only male and female categories. This limited conceptualization excludes non-binary, transgender, and gender-fluid individuals and increases the risk of misclassification. It also reinforces normative assumptions about gender identity [23], which may contribute to the marginalization of underrepresented groups [24]. Gender classification technologies can also be used in non-consensual or harmful ways [25]. Possible misuses include surveillance, targeted marketing, and automated decision-making systems that affect individuals without their awareness or agreement. When applied in sensitive contexts such as online dating, attempts to infer gender from text further raise concerns about user privacy and personal autonomy.

This study relied solely on publicly available and anonymized data for academic purposes, yet it is important to acknowledge that technical success does not guarantee ethical soundness. Responsible AI research must consider the downstream risks associated with deploying such models outside controlled environments. Future work should prioritize fairness and inclusivity by incorporating gender-diverse datasets, avoiding assumptions of binary classification, and developing transparent models that can be audited for bias. The integration of explainable AI methods may also help ensure that predictions remain interpretable and accountable.

5. Conclusion

This study investigated automated gender classification in online dating profiles using a diverse set of models ranging from traditional machine learning techniques to advanced deep learning and transformer-based architectures. The

dataset consisted of self-introduction essays from OkCupid and was enriched with psychological features from LIWC, lexical features from bag-of-words (BoW), and contextual representations derived directly from raw text.

A complete preprocessing pipeline was developed, including missing value handling, text cleaning, LIWC feature extraction, bag-of-words vectorization, one-hot encoding of categorical variables, and feature standardization. Stratified five-fold cross-validation was applied to traditional and deep learning models to ensure robust evaluation, while DistilBERT relied on an 80/20 train–test split because of GPU memory constraints. Class imbalance was addressed through random oversampling, complemented by synonym-based text augmentation to increase dataset diversity.

Deep learning models trained on LIWC and BoW features showed clear advantages. The MLP trained on LIWC provided a stable and interpretable baseline, whereas the GRU trained on the combined LIWC and BoW feature set achieved superior F1-scores and robustness by leveraging sequential information in user-generated text. Challenges associated with BoW sparsity and high dimensionality were mitigated by limiting the vocabulary size and combining BoW with LIWC features, which helped control overfitting while retaining interpretability.

Fine-tuning DistilBERT on raw essays required approximately 111.5 minutes over five epochs for a single fold, with 132 seconds needed to generate binary predictions and 35 seconds for full evaluation. These timings were measured on an NVIDIA GeForce RTX 3060 Ti with 8 GB of VRAM. Across three independent runs, DistilBERT achieved a mean testing accuracy of 91.0% ± 1.0, outperforming both classical and deep learning models and capturing semantic and contextual nuances that feature-based methods cannot.

Confusion matrix analysis confirmed balanced performance across male and female classes. Misclassifications were concentrated in short or placeholder essays, which constituted approximately 14% of the validation set and provided very limited linguistic content. Additional evaluation metrics such as ROC curves and AUC can further support the analysis of model behavior under class imbalance.

Ethical considerations remain a critical aspect of this work. The binary structure of the dataset excludes non-binary and gender-fluid individuals, and LIWC's fixed dictionary limits its adaptability to evolving and informal language. These constraints highlight the need for more inclusive datasets and complementary data-driven methods such as transformers to better reflect contemporary communication styles.

Future research may explore hyperparameter optimization, additional ensemble strategies, and broader demographic attributes such as age or personality traits. The integration of Explainable AI techniques, including SHAP and LIME, could improve interpretability, although applying these methods to transformer models requires specialized optimization because of their token-level embedding structure. Hybrid integrations involving both feature-based inputs and transformer embeddings, such as LIWC with DistilBERT or BoW with DistilBERT, were not systematically examined in this study and represent an important direction for future work.

Finally, the study is limited to the OkCupid dataset, which may introduce platform-specific biases. Cross-domain validation on additional social platforms is necessary to test generalizability and reduce the risk of overfitting to platform-specific writing conventions. Future work will also examine demographic fairness by evaluating performance across demographic subgroups to ensure equitable predictions and identify potential sources of bias.

6. Declarations

6.1. Author Contributions

Conceptualization: L.F.G.C., W.J.C., and H.S.W.; Methodology: H.S.W.; Software: L.F.G.C.; Validation: L.F.G.C. and H.S.W.; Formal Analysis: L.F.G.C. and H.S.W.; Investigation: L.F.G.C.; Resources: H.S.W.; Data Curation: H.S.W.; Writing Original Draft Preparation: L.F.G.C. and H.S.W.; Writing Review and Editing: H.S.W. and L.F.G.C.; Visualization: L.F.G.C.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

This study utilized publicly accessible datasets, which can be obtained from: https://sites.google.com/site/assistmentsdata/home/2009-2010-assistment-data. Furthermore, the prediction outputs

produced in this research are hosted in an open-access repository at: https://github.com/fandysetyoutomo-dev/predict_skill_mastery.git

6.3. Funding

We sincerely acknowledge the Ministry of Higher Education, Science, and Technology of Indonesia for its financial support of this research. Such assistance has been invaluable in facilitating the research process and manuscript preparation, enabling us to achieve the best possible outcomes.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] L. Flesia, V. Fietta, C. Foresta, and M. Monaro, "What are you looking for? Investigating the association between dating app use and sexual risk behaviors," *Sexual Medicine*, vol. 9, no. 4, pp. 1-12, 2021, doi: 10.1016/j.esxm.2021.100405.

[2] C. Lauckner, "'Catfishing,' cyberbullying, and coercion: An exploration of the risks associated with dating app use among rural sexual minority males," *Journal of Gay & Lesbian Mental Health,* vol. 23, no. 3, pp. 289–306, 2019, doi: 10.1080/19359705.2019.1587729.

[3] J. A. Snyder and K. A. Golladay, "Risk factors and characteristics of catfishing fraud victimization," *Deviant Behavior*, early access, vol. 47, no. 1, pp. 64-842024, doi: 10.1080/01639625.2024.2416071.

[4] J. M. Drew and J. Webster, "The victimology of online fraud: A focus on romance fraud victimisation," *Journal of Economic Criminology*, vol. 3, no.1, pp.1-12, 2024, doi: 10.1016/j.jeconc.2024.100053.

[5] J. Coates, W*omen, Men and Language: A Sociolinguistic Account of Gender Differences in Language*, 3rd ed. New York, NY, USA: Routledge, 2015, doi: 10.4324/9781315645612.

[6] M. J. Lerchenmueller, O. Sorenson, and A. B. Jena, "Gender differences in how scientists present the importance of their research: Observational study," *BMJ (Clinical Research Ed.),* vol. 367, no. 1, pp1-12, 2019, doi: 10.1136/bmj.l6573.

[7] A. Demzik et al., "Gender-based differences in urology residency applicant personal statements," *Urology*, vol. 150, no.1, pp. 2–8, 2021, doi: 10.1016/j.urology.2020.08.066.

[8] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods," *Journal of Language and Social Psychology*, vol. 29, no. 1, pp. 24–54, 2010, doi: 10.1177/0261927X09351676.

[9] M. L. Newman, C. J. Groom, L. D. Handelman, and J. W. Pennebaker, "Gender differences in language use: An analysis of 14,000 text samples," *Discourse Processes*, vol. 45, no. 3, pp. 211–236, 2008, doi: 10.1080/01638530802073712.

[10] S. Li, A. L. Fant, D. M. McCarthy, D. Miller, J. Craig, and A. Kontrick, "Gender differences in language of standardized letter of evaluation narratives for emergency medicine residency applicants," *AEM Education and Training*, vol. 1, no. 4, pp. 334–339, 2017, doi: 10.1002/aet2.10057.

[11] G. Park et al., “Women are warmer but no less assertive than men: Gender and language on Facebook,” *PLOS ONE*, vol. 11, no. 5, pp.1-12, 2016, doi: 10.1371/journal.pone.0155885.

[12] T. Isbister, L. Kaati, and K. Cohen, “Gender classification with data independent features in multiple languages,” in *Proc. Eur. Intell. Security Informatics Conf. (EISIC), Athens, Greece*, vol. 2017, no. 1, pp. 54–60, 2017, doi: 10.1109/EISIC.2017.16.

[13] C. Fink, J. Kopecky, and M. Morawski, “Inferring gender from the content of tweets: A region specific example,” *Proc. Int. AAAI Conf. Web Social Media*, vol. 6, no. 1, pp. 459–462, 2021, doi: 10.1609/icwsm.v6i1.14320.

[14] N. K. Alhuqail, “Author identification based on NLP,” European Journal of Computer Science and Information Technology, vol. 9, no. 1, pp. 1–26, 2021.

[15] A. Safdar, O. Akhter, O. Inayat, and A. Khalid, “Using bag-of-words and psycho-linguistic features for MAPonSMS,” in *Working Notes of FIRE 2018 – Forum for Information Retrieval Evaluation, Gandhinagar*, India, vol. 2018, no. Dec., pp. 247–256, 2018.

[16] P. Tüfekci and M. B. Kösesoy, “Biological gender identification in Turkish news text using deep learning models,” *Multimedia Tools and Applications*, vol. 83, no. 17, pp. 50669–50689, 2024, doi: 10.1007/s11042-023-17622-w.

[17] T. Dalyan, H. Ayral, and Ö. Özdemir, “A comprehensive study of learning approaches for author gender identification,” *Information Technology and Control*, vol. 51, no. 3, pp. 429–445, 2022, doi: 10.5755/j01.itc.51.3.29907.

[18] N. Cheng, R. Chandramouli, and K. P. Subbalakshmi, “Author gender identification from text,” *Digital Investigation,* vol. 8, no. 1, pp. 78–88, 2011, doi: 10.1016/j.diin.2011.04.002.

[19] T. K. Koch, P. Romero, and C. Stachl, “Age and gender in language, emoji, and emoticon usage in instant messages,” *Computers in Human Behavior*, vol. 126, Art. no. 106990, no. 1, pp. 1-12, 2022, doi: 10.1016/j.chb.2021.106990.

[20] G. Vonitsanos, A. Kanavos, and P. Mylonas, "Decoding gender on social networks: An in-depth analysis of language in online discussions using natural language processing and machine learning," in *Proc. IEEE Int. Conf. Big Data (BigData),* Sorrento, Italy, vol. 2023, no. 1, pp. 4618–4625, doi: 10.1109/BigData59044.2023.10386655.

[21] P. Schwarzenberg and A. Figueroa, “Textual pre-trained models for gender identification across community question-answering members,” *IEEE Access*, vol. 11, no. 1, pp. 3983–3995, 2023, doi: 10.1109/ACCESS.2023.3235735.

[22] Z. M. Nia et al., “Twitter-based gender recognition using transformers,” *Mathematical Biosciences and Engineering*, vol. 20, no. 9, pp. 15962–15981, 2023, doi: 10.3934/mbe.2023711.

[23] O. Keyes, “The misgendering machines: Trans/HCI implications of automatic gender recognition,” *Proc. ACM* Hum.-Comput. Interact., vol. 2, no. CSCW, pp. 1–22, 2018, doi: 10.1145/3274357.

[24] K. Thieme, M. A. S. Saunders, and L. Ferreira, "From language to algorithm: Trans and non-binary identities in research on facial and gender recognition," *AI Ethics*, vol. 5, no.1, pp. 991–1008, 2025, doi: 10.1007/s43681-023-00375-5.

[25] S. Shrestha and S. Das, “Exploring gender biases in ML and AI academic research through systematic literature review,” *Frontiers in Artificial Intelligence*, vol. 5, no.1, pp. 1-18, 2022, doi: 10.3389/frai.2022.976838.