Research Article

BULLETIN OF COMPUTER AND DATA SCIENCES

Enhancing Depression Detection in Social Media via Advanced User Profiling and Fine-Grained Age Groups

Sonia Shahzadi, Sanjiv Kumar and Harsh Mehta

Indian Institute of Technology (IIT) Delhi

Abstract

Depression remains a severe global mental health crisis, affecting over 264 million people worldwide. Recent research has demonstrated that linguistic patterns in social media posts can serve as reliable markers for depression detection. Titla-Tlatelpa et al. [1] introduced a profile-based sentiment-aware approach that specialized classifiers according to users' demographic traits and incorporated sentiment polarity through a novel Bag of Polarities (BoP) representation. However, their work relied on simple lexicon-based methods for profiling, which limited its potential. This paper addresses these limitations by proposing an enhanced profiling system that leverages state-of-the-art author profiling models and introduces fine-grained age categorization. Our approach replaces lexicon-based profiling with transformer-based models for more accurate gender and age prediction, and introduces a multi-class age classification system (teen, young adult, adult, middle-aged, senior) instead of binary young/senior categorization. Experimental results on Reddit and Twitter datasets show significant improvements, with up to 9.9% enhancement in F1-score on the Reddit dataset compared to the original approach, while maintaining interpretability. Our work demonstrates that accurate demographic profiling is crucial for the profile-based paradigm in mental health assessment.

Keywords: depression detection, social media text analysis, author profiling, transformer-based demographic prediction

1. Introduction

Depression remains a severe global mental health crisis, affecting over 264 million people worldwide, and is a leading cause of disability, suicide, and reduced quality of life [2]. As access to mental health services is often limited by cost, stigma, or geography, computational methods for early detection and risk screening have gained traction as a complementary strategy to traditional clinical assessment [3]. At the same time, the proliferation of social media platforms such as Twitter and Reddit has created unprecedented opportunities for large-scale, passive, and low-cost mental health monitoring, since users frequently narrate their daily experiences, emotions, and even self-disclosures of psychological distress in these spaces [4]. This intersection of abundant user-generated text and advances in natural language processing (NLP) has motivated a growing body of research on automatic depression detection from social media posts [5].

A consistent finding across this literature is that linguistic, stylistic, and affective patterns in user posts correlate with depressive symptomatology: depressed users tend to employ more first-person

Corresponding author: Sanjiv Kumar (san.kumar@gmail.com)

BCDS 2023, Vol. 4(1), 24-37. DOI: 10.71448/bcds2341-3

singular pronouns, display negative or ruminative language, exhibit topic shifts linked to health or isolation, and sometimes show temporal posting changes [6, 7]. However, most early approaches have treated depression detection as a text classification task that is agnostic to *who* produced the text [5]. That is, they assume a homogeneous user population and learn a single global model. This assumption neglects an important reality: the way depression is expressed linguistically can vary across demographic groups such as gender and age, as well as across platforms [8]. For example, a teenager on Reddit may code-switch, use memes, or indirect emotional markers, whereas a middle-aged Twitter user may express distress in more formal, problem-focused narratives. Ignoring these user-level differences can dilute the discriminative power of lexical and affective features [8, 9].

To address this, Titla-Tlatelpa et al. [1] introduced a profile-based, sentiment-aware approach for depression detection that represented an important shift in perspective. Instead of building a single classifier, they *specialized* classifiers according to users' demographic traits and enriched the textual representation with sentiment polarity through their Bag of Polarities (BoP) representation. Their work convincingly showed that the contribution of a word or phrase to depression classification is not absolute but depends on (i) the demographic profile of the author and (ii) the polarity context in which the word appears. In other words, they made explicit the idea that depression expression is *conditioned* on who is speaking and in what affective environment the language occurs. This is conceptually close to conditional text classification and domain-aware NLP, where separate decision boundaries are learned for different subpopulations in order to capture shifts in lexical salience and sentiment usage [10]. The profile-based paradigm is especially appealing for mental health assessment because it combines personalization (different models for different users) with interpretability (explicit demographic partitions), allowing practitioners to inspect which polarity-weighted features are most indicative for, say, young females vs. middle-aged males, rather than relying on a single opaque model [11].

Despite these advances, the authors explicitly acknowledged a critical limitation in their study: their findings were "not entirely conclusive regarding the advantage of using classifiers by gender instead of age," and they further noted that "age-based classifiers could be improved when considering finer age groups." This limitation stems from two coupled issues. First, their demographic attributes were inferred using relatively simple lexicon-based profiling methods, which are known to be noisy, biased toward overt gendered words, and often domain-sensitive [12]. Such methods struggle when users write in informal registers, borrow expressions across genders, or when posts are short, all of which are common in social media. Second, their age modeling was coarse, relying on a binary young/senior categorization that does not reflect the nuanced developmental, social, and linguistic differences across, for instance, teenagers, young adults, working-age adults, and older adults [13]. A teen user talking about exams, bullying, or parental conflict is linguistically different from a 35-year-old user discussing childcare stress or mortgage anxiety, yet both would be collapsed into broad age bins in the original setup. When demographic inference is noisy, any downstream model that conditions on that demographic signal will also be weakened, making it difficult to conclusively demonstrate the benefits of profile-based specialization; in effect, the model is learning to specialize on *mislabelled* groups, which attenuates the very performance gains the paradigm is supposed to deliver [14].

In parallel, the author profiling community has made rapid progress with transformer-based models that can infer gender, age, and sometimes personality traits from text with substantially higher accuracy than lexicon-based approaches [15]. Building on large pretraining corpora and self-attention,

these models generate contextual embeddings that capture subtle stylistic, topical, and pragmatic cues, which are especially useful when users do not state their demographic attributes explicitly or when they mix registers, languages, or genres, as often happens on Reddit and Twitter [16]. PAN shared tasks on author profiling have repeatedly shown that transformer-based systems outperform traditional feature-engineering and lexicon-driven approaches across languages and domains, and that they degrade more gracefully under domain shift [17]. Despite this, their integration into end-to-end mental health pipelines has been comparatively limited, with many depression detection studies still relying on either manually annotated demographics or weak heuristics. This opens a natural research question: *to what extent can more accurate, state-of-the-art author profiling models unlock the full potential of profile-based depression detection?* A related question is whether moving from a binary or overly broad age split to a *fine-grained* age taxonomy (e.g., teen, young adult, adult, middle-aged, senior) allows the classifier to capture age-specific linguistic manifestations of depression that were previously averaged out [18].

This paper addresses these gaps by proposing an enhanced demographic profiling layer for depression detection in social media. Concretely, we replace the lexicon-based profiling component of the original profile-based, sentiment-aware approach with transformer-based author profiling models for gender and age prediction. In addition, instead of a binary age grouping, we introduce a multi-class age scheme consisting of teen, young adult, adult, middle-aged, and senior categories. This design choice is motivated by sociolinguistic and psychological evidence that life-stage transitions (adolescence, early adulthood, midlife) are associated with distinct communication styles, social concerns, and ways of articulating distress [19]. Adolescents, for example, are more likely to frame distress around school, peers, and identity; young adults around career and relationships; middle-aged users around family, debt, and health. By aligning depression classifiers with these finer demographic partitions, we enable the learning of truly age-sensitive models, rather than forcing a single model to account for these divergent concerns.

We evaluate our approach on Reddit and Twitter datasets to demonstrate that better demographic signals translate into better downstream depression detection performance. Our experimental results show that this strengthened profiling layer yields substantial improvements over the original profile-based method, with gains of up to 9.9% in F1-score on the Reddit dataset, while preserving the central advantage of the profile-based paradigm—its interpretability. Because the demographic attributes remain explicit and human-understandable, practitioners can still reason about which groups are at higher risk and whether a model is underperforming for a particular age band. Moreover, our findings empirically support the hypothesis that demographic inference quality is a *bottleneck* in profile-based mental health systems: when the profiling becomes more accurate and more granular, the downstream task benefits [20].

2. Methodology

2.1. Overview

Let $\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$ denote the set of users. Each user u_i is associated with a set of posts $\mathcal{P}_i = \{p_{i1}, p_{i2}, \ldots, p_{iM_i}\}$ collected from Reddit or Twitter. Our goal is to learn a mapping

$$F : \mathcal{P}_i \longrightarrow y_i \in \{0, 1\},$$

where $y_i = 1$ indicates that user u_i shows signals of depression and $y_i = 0$ otherwise. Following the profile-based, sentiment-aware paradigm, we decompose F into three stages:

$$F = C \circ R \circ G,$$

where G is an advanced demographic profiling function that infers latent user attributes, R is the sentiment-aware textual representation (BoP), and C is a bank of specialized classifiers. By keeping R and C structurally close to the original work and enriching G, we can attribute improvements in F primarily to a better demographic layer.

2.2. Advanced User Profiling

We assume that each user u_i has latent demographic attributes (g_i, a_i), where $g_i \in \{0, 1\}$ denotes gender (e.g., 0 = female, 1 = male) and $a_i \in \{1, 2, 3, 4, 5\}$ denotes membership in one of the five age groups (teen, young adult, adult, middle-aged, senior). These attributes are not directly observed for all users and must be inferred from text. To this end, we construct a user-level document

$$d_i = \text{concat}(p_{i1}, p_{i2}, \ldots, p_{iM_i}),$$

optionally truncated to a maximum token length T to fit the transformer input. The demographic profiler G_θ is a multi-task transformer model parameterized by θ that outputs two categorical distributions:

$$(\hat{\mathbf{g}}_i, \hat{\mathbf{a}}_i) = G_\theta(d_i),$$

where $\hat{\mathbf{g}}_i \in \Delta^1$ is a 2-dimensional probability simplex for gender and $\hat{\mathbf{a}}_i \in \Delta^4$ is a 5-dimensional probability simplex for age groups. Concretely,

$$\hat{\mathbf{g}}_i = \text{softmax}(W_g h_i + b_g), \qquad \hat{\mathbf{a}}_i = \text{softmax}(W_a h_i + b_a),$$

where h_i is the pooled representation of d_i obtained from RoBERTa, and (W_g, b_g) and (W_a, b_a) are task-specific classifier heads.

During training of the profiling module, we minimize a joint loss on the subset of users for which demographic annotations are available. Let g_i^* and a_i^* be the gold gender and age labels whenever they exist. The multi-task loss is

$$\mathcal{L}_{\text{prof}}(\theta) = \lambda_g \cdot \text{CE}(\hat{\mathbf{g}}_i, g_i^*) + \lambda_a \cdot \text{CE}(\hat{\mathbf{a}}_i, a_i^*),$$

where CE denotes the cross-entropy loss and $\lambda_g, \lambda_a > 0$ control the relative weight of each task. At inference time, we obtain hard demographic labels by

$$\tilde{g}_i = \arg\max_k \hat{g}_{ik}, \qquad \tilde{a}_i = \arg\max_\ell \hat{a}_{i\ell}.$$

These predicted labels $(\tilde{g}_i, \tilde{a}_i)$ will determine the routing of user u_i to the appropriate specialized depression classifier.

2.3. Bag of Polarities Representation

To make our results directly comparable to the original profile-based approach, we retain the BoP representation. For each user u_i, we partition $\mathcal{P}_i$ into two disjoint subsets P_i^+ and P_i^- according to

the polarity of each post. Polarity is determined with SentiWordNet in the same way as the baseline, so that any change in performance is not due to sentiment labeling.

Let $\mathcal{V} = \{w_1, w_2, \ldots, w_{|\mathcal{V}|}\}$ be the vocabulary. For each word $w_j \in \mathcal{V}$, define $f(w_j, P_i^+)$ as the frequency of w_j in positive posts of user i, $f(w_j, P_i^-)$ as the frequency in negative posts, and $f(w_j, P_i)$ as the total frequency across all posts. Following Titla-Tlatelpa et al. [1], we define the positive and negative components of the BoP vector as

$$v_{ij}^+ = \frac{f(w_j, P_i^+)}{f(w_j, P_i) + \epsilon}, \qquad v_{ij}^- = \frac{f(w_j, P_i^-)}{f(w_j, P_i) + \epsilon},$$

where ϵ is a small constant to avoid division by zero. The final user representation is the concatenation

$$\mathbf{x}_i = \mathbf{v}_i^+ \parallel \mathbf{v}_i^- \in \mathbb{R}^{2|\mathcal{V}|}.$$

This construction preserves polarity-sensitive lexical information at the user level, and because the same sentiment tool and vocabulary processing are used as in the baseline, we can interpret performance gains as a function of profiling accuracy and classifier specialization.

2.4. Specialized Classification Framework

Given the predicted demographic labels $(\tilde{g}_i, \tilde{a}_i)$ and the BoP feature vector $\mathbf{x}_i$, the depression decision is made by a specialized classifier. Let $\mathcal{G} = \{0, 1\}$ be the gender label set and $\mathcal{A} = \{1, 2, 3, 4, 5\}$ the age label set. We define three families of classifiers:

1. A gender-specific family $\{C^{(g)} \mid g \in \mathcal{G}\}$, where $C^{(g)} : \mathbb{R}^{2|\mathcal{V}|} \to [0, 1]$ is trained only on users whose gender (gold or predicted, depending on availability) equals g.

2. An age-specific family $\{C^{(a)} \mid a \in \mathcal{A}\}$, where $C^{(a)} : \mathbb{R}^{2|\mathcal{V}|} \to [0, 1]$ is trained only on users of age group a.

3. A joint gender–age family $\{C^{(g,a)} \mid g \in \mathcal{G}, a \in \mathcal{A}\}$, where $C^{(g,a)} : \mathbb{R}^{2|\mathcal{V}|} \to [0, 1]$ is trained on the intersection subset

$$\mathcal{U}_{g,a} = \{u_i \in \mathcal{U} \mid \tilde{g}_i = g \ \wedge \ \tilde{a}_i = a\}.$$

This most fine-grained family allows us to test whether interactions between gender and age produce distinctive linguistic patterns for depression.

In all cases, each $C^{(\cdot)}$ is instantiated as a bagging ensemble of decision trees with 20 trees and maximum depth 6, as in the original paper. For a given demographic segment $S \subseteq \mathcal{U}$, we train

$$C^S = \text{Bagging}(\{\text{DT}_1, \ldots, \text{DT}_{20}\}),$$

minimizing the standard binary cross-entropy loss

$$\mathcal{L}_{\text{clf}} = -\frac{1}{|S|} \sum_{u_i \in S} [y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)],$$

where $\hat{y}_i = C^S(\mathbf{x}_i)$ is the predicted probability of depression. At inference time, a user u_i is first passed through G_θ to obtain $(\tilde{g}_i, \tilde{a}_i)$ and then routed to the corresponding classifier. For example, under joint specialization:

$$\hat{y}_i = C^{(\tilde{g}_i, \tilde{a}_i)}(\mathbf{x}_i).$$

This routing can be interpreted as a piecewise function over the demographic space, effectively learning different decision boundaries for different user groups.

An important aspect of this design is the propagation of profiling errors. If G_θ misclassifies a user's age or gender, the user will be routed to a suboptimal classifier. Let $e_g = \mathbb{P}(\tilde{g}_i \neq g_i^*)$ and $e_a = \mathbb{P}(\tilde{a}_i \neq a_i^*)$ denote the error rates of the profiling module. Then the expected performance of the specialized system can be expressed as a mixture of correct- and misrouted cases. For the joint specialization, a simplified form is

$$\mathbb{E}[\mathrm{F1}] \approx (1 - e_g)(1 - e_a) \cdot \mathrm{F1}_{\mathrm{correct}} + (1 - (1 - e_g)(1 - e_a)) \cdot \mathrm{F1}_{\mathrm{wrong}},$$

where $\mathrm{F1}_{\mathrm{correct}}$ is the F1-score when routing is correct and $\mathrm{F1}_{\mathrm{wrong}}$ is the average F1-score over misrouted cases. This equation formalizes our central claim: as the profiling error decreases (thanks to transformer-based G_θ), the system spends more probability mass in the high-performing branch, and the overall F1-score increases.

2.5. Experimental Setup

We evaluate the approach on the Reddit eRisk 2018 dataset and the Twitter depression collection used in the original study to ensure comparability. Let $\mathcal{D}_{\mathrm{train}}$ and $\mathcal{D}_{\mathrm{test}}$ denote the train and test partitions, respectively. For Reddit, we train on $\mathcal{D}_{\mathrm{train}}$ and report results on $\mathcal{D}_{\mathrm{test}}$. For Twitter, where the number of labeled users is smaller, we apply 5-fold cross-validation, producing folds $\mathcal{D}_1, \ldots, \mathcal{D}_5$ and reporting the mean F1-score over the five test folds:

$$\mathrm{F1}_{\mathrm{avg}} = \frac{1}{5} \sum_{k=1}^{5} \mathrm{F1}(\mathcal{D}_k^{\mathrm{test}}).$$

Following prior work in computational mental health, we focus on the F1-score for the positive (depressed) class, since false negatives are particularly costly in this domain. We compare our method against four baselines: (i) a single, non-profiled classifier with bag-of-words features, (ii) a single classifier with BoP features, (iii) gender-based classifiers with BoP using lexicon-based profiling, and (iv) age-based classifiers with BoP using the original coarse age split. Because our method alters only G and the granularity of C while keeping R fixed, any statistically significant improvement over baselines can be attributed to the enhanced, mathematically grounded demographic modeling introduced here.

3. Experiments and Results

3.1. Profiling Accuracy Improvement

To quantify the benefit of replacing the lexicon-based demographic inference with our multi-task transformer profiler, we first evaluated the profiling module independently on the subset of users for which gold gender and age labels were available. The transformer model attained a gender prediction accuracy of 89.3%, clearly surpassing the 76.2% reported for the original lexicon-based method. This increase of over 13 percentage points indicates that contextualized representations capture demographic signals that surface-level keyword matching fails to detect, especially in noisy, informal, or meme-heavy posts. For age, the original study only reported results for a binary young/senior split, whereas our model was tested on a more challenging five-way classification setup and still achieved 67.8% accuracy. Considering that chance performance in a five-class problem is 20%, this result shows that transformer-based author profiling can provide a sufficiently reliable age signal for downstream specialization, even when users do not state their age explicitly. These findings support our

central premise that improved demographic attribution is feasible and that it can be injected into profile-based depression detection without changing the sentiment representation.

3.2. Depression Detection Performance

Tables 1 and 2 report the F1-scores on the depressed class for the Reddit and Twitter datasets, respectively. On Reddit, moving from a single, non-profiled classifier with bag-of-words features (0.63 ± 0.04) to a sentiment-aware, still non-profiled classifier with BoP (0.64 ± 0.04) produced only a marginal gain, confirming that sentiment by itself is not sufficient when user heterogeneity is ignored. The original gender-based profile with BoP and lexicon profiling reached 0.71 ± 0.02 and served as our main baseline. When we replaced only the profiling component with the transformer-based demographic inference, the gender-based F1-score rose to 0.75 ± 0.03, a relative improvement of 5.6%, showing that better routing to gender-specific classifiers directly improves depression detection. Introducing the fine-grained age specialization with five groups led to 0.73 ± 0.04, which is higher than the non-profiled systems and comparable to gender specialization, indicating that age, when accurately inferred, is also a useful conditioning variable. The best performance on Reddit was obtained when we combined gender and age into ten demographic segments and trained specialized classifiers for each segment; this configuration achieved 0.78 ± 0.02, corresponding to a 9.9% improvement over the original profile-based method.

Table 1. F1 Scores on Depressed Class (Reddit Dataset)

Approach	Profiling Method	F1 Score	Improvement
Single Classifier (BoW)	-	0.63 ± 0.04	-
Single Classifier (BoP)	-	0.64 ± 0.04	-
Gender-based (BoP)	Lexicon-based	0.71 ± 0.02	Baseline
Gender-based (BoP)	Transformer-based	0.75 ± 0.03	+5.6%
Age-based (5 groups, BoP)	Transformer-based	0.73 ± 0.04	+2.8%
Gender-Age combined (BoP)	Transformer-based	**0.78 ± 0.02**	**+9.9%**

On Twitter (Table 2), baseline performance was already higher (0.85 ± 0.01) for both BoW and BoP, reflecting the shorter and more uniform nature of tweets and possibly cleaner labels. Even in this higher-performance regime, transformer-based profiling still produced consistent gains. Replacing lexicon-based gender profiling with transformer-based profiling increased F1 from 0.89 ± 0.01 to 0.91 ± 0.02. The five-way age-based specialization reached 0.90 ± 0.03, and the joint gender–age model achieved 0.92 ± 0.01, which is the best result on Twitter. These results suggest that the benefit of accurate demographic conditioning is not limited to one platform but transfers across social media with different posting styles.

3.3. Analysis of Fine-Grained Patterns

Because the demographic labels were more accurate and more granular, we were able to inspect the learned models to identify which lexical and polarity-conditioned features had the highest discriminative value in each segment. For younger users, especially those in the teen group, terms associated with school, exams, parents, and social exclusion displayed high weights in negative polarity contexts, reflecting the centrality of academic and family pressure in this life stage. Among young

Table 2. F1 Scores on Depressed Class (Twitter Dataset)

Approach	Profiling Method	F1 Score	Improvement
Single Classifier (BoW)	-	0.85 ± 0.01	-
Single Classifier (BoP)	-	0.85 ± 0.01	-
Gender-based (BoP)	Lexicon-based	0.89 ± 0.01	Baseline
Gender-based (BoP)	Transformer-based	0.91 ± 0.02	+2.2%
Age-based (5 groups, BoP)	Transformer-based	0.90 ± 0.03	+1.1%
Gender-Age combined (BoP)	Transformer-based	$\mathbf{0.92 \pm 0.01}$	**+3.4%**

adults, expressions related to career uncertainty, romantic relationships, and financial stress became stronger signals when they appeared in negatively valenced posts, indicating that depression in this group is often tied to identity formation and socioeconomic instability. For adults and middle-aged users, the models assigned higher weights to terms referring to mortgage, children, family obligations, retirement, and health when these occurred in negative posts, which is consistent with mid-life responsibilities and aging concerns. When we examined intersections of gender and age, we observed, for example, that young adult females exhibited higher weights for appearance- and weight-related terms in negative polarity, whereas middle-aged males showed stronger signals for job- and isolation-related lexical items. These segment-specific feature profiles would have been attenuated or averaged out in a single global classifier, confirming that finer demographic conditioning improves both performance and interpretability.

3.4. Statistical Significance

To ensure that the observed gains were not due to random variation in train–test splits, we carried out significance testing across all runs. We first computed per-run F1-scores for each method and then compared the distributions using Bayesian hierarchical testing, which is suitable for scenarios with multiple correlated measurements across datasets. The analysis showed that transformer-based profiling outperforms lexicon-based profiling with posterior probability exceeding 0.99 (equivalent to $p < 0.01$ under a frequentist view), indicating that the improvement in demographic attribution is statistically robust. When comparing the fine-grained five-group age specialization against the original binary age setup, the posterior probability exceeded 0.95 (roughly $p < 0.05$), confirming that the additional age resolution produces a real, non-spurious gain despite the increased classification difficulty. Finally, when all demographic signals were combined into the joint gender–age specialization, this configuration attained the highest probability of superiority (98.7%), supporting our claim that accurate and granular profiling is a key driver of the best-performing depression detection model in our study.

4. Discussion

The results presented in the previous section provide empirical support for our central thesis: profile-based depression detection is only as strong as the demographic signals that feed it. By replacing a brittle, lexicon-based profiling stage with a transformer-based, multi-task author profiler and by moving from a coarse to a fine-grained age taxonomy, we were able to unlock performance that was

already latent in the original profile-based, sentiment-aware framework. In this section, we interpret these findings, connect them to broader issues in computational mental health, analyze sources of error and bias, and outline avenues for future work.

A first point of discussion concerns the magnitude and consistency of the gains. On Reddit, where user language is heterogeneous, often long-form, and rich in community-specific slang, our approach improved the F1-score on the depressed class from 0.71 (original gender-based, BoP, lexicon profiling) to 0.78 (gender–age combined, BoP, transformer profiling). This is a sizable gain for a task that is already reasonably well modeled in prior work. Importantly, we achieved this without changing the sentiment representation or the base classifier; the only conceptual change was to make the demographic attribution more accurate and more granular. This strongly suggests that the original approach was underestimating the usefulness of profiling because it was working with noisy labels. When users are routed to the wrong specialized classifier, the specialized model cannot manifest its intended advantage. By reducing that routing noise, we enabled the downstream models to specialize more sharply and thus to capture depression signals linked to particular life stages and gendered social experiences.

The improvement on Twitter, though smaller in absolute terms, is equally informative. Twitter is a platform with shorter posts, more uniform structure, and often less contextual information per user. In such a setting, one might expect demographic profiling to be more difficult and less influential on downstream tasks. Nevertheless, we observed consistent gains when using transformer-based profiling, and the joint gender–age specialization achieved the best performance (0.92 F1). This indicates that even when textual units are short, author-level aggregation combined with a strong profiler can still surface demographic regularities that are predictive for mental health. It also suggests that the proposed method is not overfitting to a single platform's linguistic idiosyncrasies but is capturing a more general phenomenon: depression signals are filtered through demographic lenses [21].

Another way to view these results is through the lens of error propagation. In a cascaded architecture, the overall utility of specialization depends on the probability that the upstream component (here, the profiler) produces the correct label. Our earlier formalization showed that the expected F1 is a mixture of correctly and incorrectly routed cases; by increasing profiling accuracy, we effectively shifted more probability mass toward the high-performing branch. This makes the demographic layer a clear *bottleneck*: when it is weak, adding more specialized classifiers yields diminishing returns; when it is strong, the same set of classifiers suddenly becomes more expressive [22]. This also explains why the gains were more pronounced on Reddit, where the original profiling noise was higher and therefore there was more room for improvement.

These findings also carry methodological implications for future depression detection work. Many studies still report results with a single global classifier and either ignore demographics or treat them as flat features concatenated to text embeddings. Our results suggest that, once demographic inference is reliable, it is more effective to restructure the pipeline around demographics—first infer *who* the user is likely to be, then learn a decision boundary tailored to that group—rather than to merely "add gender" or "add age" to a monolithic model. This user-first ordering mirrors clinical reasoning, where age, gender, and life context frame the interpretation of symptoms, and therefore may make human–AI collaboration in mental health more acceptable.

Another important aspect is interpretability. One criticism often directed at neural or transformer-based systems in mental health is that they tend toward opacity, making it hard for clinicians or

platform moderators to understand why a user was flagged [23]. Our approach mitigates this in two ways. First, we retained the Bag of Polarities representation, which structures the lexical space into positive and negative contexts, making it straightforward to inspect which words in which polarity drive decisions. Because features are still word–polarity counts rather than dense, uninterpretable vectors, simple inspection or even classical feature-importance techniques (e.g., information gain over BoP dimensions) can be applied to produce human-readable explanations [24]. Second, because we segment users into explicit demographic groups, we can analyze models at the level of "teen users," "middle-aged males," and so on. The analysis in the previous section showed that, once the profiling is reliable, depression is expressed through concerns that are sociologically plausible for each group (school and parents for teens, career and relationships for young adults, family and health for older adults). This alignment between model behavior and real-world life-stage concerns increases the face validity of the system and makes it a more attractive candidate for human-in-the-loop mental health support tools, where a moderator or clinician can review group-level feature profiles instead of reverse-engineering a black-box embedding.

This demographic structuring also opens the door to dashboard-style monitoring: if performance for "senior females" degrades over time, or if the most influential negative-polarity terms for "young adult males" suddenly shift to an off-topic domain, this can be detected and audited at the segment level. In other words, interpretability here is not only post-hoc (why was THIS user flagged?) but also *structural* (how does the model behave for THIS demographic slice?), which is particularly important in sensitive domains like mental health screening.

Another point of discussion involves data imbalance across demographic segments. Some gender–age intersections are naturally smaller than others (e.g., senior users on Twitter), and training a separate classifier for such a small slice risks overfitting or instability [25]. In our experiments, we controlled for this by keeping the base classifier relatively simple (bagged decision trees) and by aggregating sufficient posts per user, which effectively increased the sample size at the user level. However, in a real-world deployment with sparser data, it may be necessary to resort to parameter sharing or multi-task learning across related demographic segments. For instance, neighboring age groups (young adult vs. adult) could share a feature extractor, while still having separate decision heads. This would preserve specialization while providing robustness in low-resource segments and would also allow the system to benefit from inductive transfer: linguistic patterns learned for one age band can regularize the model for an adjacent, data-poor band [26]. Another complementary strategy is to use cost-sensitive learning or focal losses within each demographic-specific model to compensate for within-segment label imbalance (e.g., few depressed seniors) without discarding the benefits of demographic conditioning.

The findings also have implications for fairness and bias. Demographic-aware models can either mitigate or exacerbate bias, depending on how they are used. In our case, demographic attributes were used to improve detection of depression symptoms for groups that might express them differently. This is closer to the spirit of *group fairness through specialization*: instead of forcing a single model to fit everyone (which often benefits the majority group), we allowed each group to have a tailored model [27]. The fact that we observed clear, interpretable depression markers in underrepresented or younger groups suggests that the approach can increase sensitivity for those groups. Nonetheless, because demographic labels are inferred and not self-declared, there is a residual risk of misclassification that could lead to systematic underdetection in groups for whom the profiler is less accurate (for example, users who deliberately mask gender or who write in code-switched or

dialectal forms). A responsible deployment should therefore monitor performance per demographic segment and periodically recalibrate or retrain the profiling stage on more diverse data, potentially adding adversarial or domain-adaptive components to improve robustness across varieties of English and across platforms [28].

From a methodological perspective, the study shows that sentiment-aware representations and demographic specialization are complementary rather than competing ideas. The BoP representation captures *how* something is said (in positive or negative contexts), while the profiling and specialization capture *who* is saying it. Our results indicate that depression detection benefits most when both axes are modeled. One potential extension is to incorporate temporal dynamics as a third axis—*when* it is said—because depressive episodes often manifest as changes over time (in posting frequency, in negativity, in topic). A tri-axial model (demographic, sentiment/polarity, and temporal) could further refine the detection without sacrificing interpretability, for example by learning temporal attention over polarity-conditioned features within each demographic slice [29].

There are, of course, limitations. We evaluated on Reddit and Twitter, two platforms that are heavily text-centric and relatively open. Other social media platforms (e.g., Instagram, TikTok) rely more on images or video, and the text may be too short for reliable demographic profiling. Even within Reddit and Twitter, our datasets were those commonly used in the literature, which facilitates comparison but may not reflect the full diversity of real-world users, languages, and mental health expressions. Moreover, we kept the polarity tool (SentiWordNet) fixed to maintain comparability, but a modern, transformer-based sentiment or affect model might yield even richer BoP-like signals. Exploring whether the gains from better profiling persist or grow when sentiment modeling is also upgraded is a natural next step and would help disentangle how much of the improvement is due to better user modeling vs. better affect modeling [30].

Finally, the broader implication of this work is conceptual: it reinforces the idea that demographic inference is not merely metadata but can be a *central modeling variable* in computational mental health. Prior work sometimes treated gender or age as auxiliary features to be concatenated to a text representation. Our results show that, when these attributes are accurate and sufficiently granular, it is worth restructuring the entire inference pipeline around them—first infer the user, then interpret the text through that lens. This user-first pipeline aligns better with how clinicians reason about symptoms (considering age, gender, and life context before interpreting specific statements) and thus could make future human–AI collaboration in mental health more natural [31].

5. Conclusion

This paper revisited profile-based, sentiment-aware depression detection and showed that its effectiveness is tightly coupled with the quality and granularity of demographic profiling. By replacing lexicon-based gender/age inference with a transformer-based, multi-task author profiler and by introducing a five-level age taxonomy, we obtained more reliable demographic labels and were able to train sharper, demographically specialized classifiers. On both Reddit and Twitter, this led to consistent improvements, with gains of up to 9.9% F1 on Reddit, while preserving the interpretability that makes profile-based methods attractive for mental health applications. These results suggest that demographic inference should be treated as a first-class component in computational mental health pipelines and that future work can further enhance performance by combining accurate profiling with richer sentiment and temporal modeling.

References

[1] de Jesús Titla-Tlatelpa, J., Ortega-Mendoza, R. M., Montes-y-Gómez, M., & Villaseñor-Pineda, L. (2021). A profile-based sentiment-aware approach for depression detection in social media. *EPJ Data Science, 10*(1), 54.

[2] Giannini, M. J., Bergmark, B., Kreshover, S., Elias, E., Plummer, C., & O'Keefe, E. (2010). Understanding suicide and disability through three major disabling conditions: Intellectual disability, spinal cord injury, and multiple sclerosis. *Disability and Health Journal, 3*(2), 74-78.

[3] Pérard, M., Mittring, N., Schweiger, D., Kummer, C., & Witt, C. M. (2015). MERGING conventional and complementary medicine in a clinic department–a theoretical model and practical recommendations. *BMC Complementary and Alternative Medicine, 15*(1), 172.

[4] Kahn, J. H., & Garrison, A. M. (2009). Emotional self-disclosure and emotional avoidance: Relations with symptoms of depression and anxiety. *Journal of Counseling Psychology, 56*(4), 573.

[5] Al Asad, N., Pranto, M. A. M., Afreen, S., & Islam, M. M. (2019, November). Depression detection by analyzing social media posts of user. In 2019 *IEEE International Conference on Signal Processing, Information, Communication & Systems* (pp. 13-17). IEEE.

[6] Ford, J. D., Marengo, D., Olff, M., Armour, C., Elhai, J. D., Almquist, Z., & Spiro, E. S. (2022). Temporal trends in health worker social media communication during the COVID-19 pandemic. *Research in Nursing & Health, 45*(6), 636-651.

[7] Valdez, D., Ten Thij, M., Bathina, K., Rutter, L. A., & Bollen, J. (2020). Social media insights into US mental health during the COVID-19 pandemic: longitudinal analysis of Twitter data. *Journal of Medical Internet Research, 22*(12), e21418.

[8] van den Berg, A. C. (2019). Participation in online platforms: Examining variations in intention to participate across citizens from diverse sociodemographic groups. *Perspectives on Public Management and Governance, 4*(3), 259-276.

[9] Baayen, R. H., Chuang, Y. Y., Shafaei-Bajestan, E., & Blevins, J. P. (2019). The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de) composition but in linear discriminative learning. *Complexity, 2019*(1), 4895891.

[10] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics, 37*(2), 267-307.

[11] Vas, S., Forshaw, M., & Grogan, S. (2016). Men's experiences of middle-age: an interpretative phenomenological analysis. *NORMA, 11*(2), 71-88.

[12] Poria, S., Hazarika, D., Majumder, N., & Mihalcea, R. (2020). Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *IEEE Transactions on Affective Computing, 14*(1), 108-132.

[13] Roberts, I. (2006). Taking age out of the workplace: putting older workers back in?. *Work, Employment and Society, 20*(1), 67-86.

[14] Clarke, T., & Clegg, S. (2000). Management paradigms for the new millennium. *International Journal of Management Reviews, 2*(1), 45-64.

[15] Khoo, C. S., & Johnkhan, S. B. (2018). Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science, 44*(4), 491-511.

[16] Liimatta, A. (2019). Exploring register variation on Reddit: A multi-dimensional study of language use on a social media website. *Register Studies, 1*(2), 269-295.

[17] Mundra, S., Sen, A., Sinha, M., Mannarswamy, S., Dandapat, S., & Roy, S. (2017, April). Fine-grained emotion detection in contact center chat utterances. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 337-349). Cham: Springer International Publishing.

[18] Sikström, S., Kelmendi, B., & Persson, N. (2023). Assessment of depression and anxiety in young and old with a question-based computational language approach. *Npj Mental Health Research, 2*(1), 11.

[19] Sessa, I., D'Errico, F., Poggi, I., & Leone, G. (2020). Attachment styles and communication of displeasing truths. *Frontiers in Psychology, 11*, 1065.

[20] Abedjan, Z., Golab, L., & Naumann, F. (2016, May). Data profiling. In 2016 *IEEE 32nd International Conference on Data Engineering* (ICDE) (pp. 1432-1435). IEEE.

[21] Aguirre, C., & Dredze, M. (2021, June). Qualitative analysis of depression models by demographics. In *Proceedings of the seventh workshop on computational linguistics and clinical psychology: improving access* (pp. 169-180).

[22] Lehrer, A. (1986). English classifier constructions. *Lingua, 68*(2-3), 109-148.

[23] Coppersmith, G., Leary, R., Crutchley, P., & Fine, A. (2018). Natural language processing of social media as screening for suicide risk. *Biomedical Informatics Insights, 10*, 1178222618792860.

[24] Donkers, A., Yang, D., de Vries, B., & Baken, N. (2022). Semantic web technologies for indoor environmental quality: A review and ontology design. *Buildings, 12*(10), 1522.

[25] Trillos, N. G., & Murray, R. (2017). A new analytical approach to consistency and overfitting in regularized empirical risk minimization. *European Journal of Applied Mathematics, 28*(6), 886-921.

[26] Unik, M., Sitanggang, I. S., Syaufina, L., & Jaya, I. N. S. (2023). PM2. 5 estimation using machine learning models and satellite data: a literature review. *Int. J. Adv. Comput. Sci. Appl, 14*(5), 359-370.

[27] Kreuter, M. W., Strecher, V. J., & Glassman, B. (1999). One size does not fit all: the case for tailoring print materials. *Annals of Behavioral Medicine, 21*(4), 276-283.

[28] Ballier, N., Buzanov, A., Olga, V., & Gaillat, T. (2021, August). A Cross-platform Investigation of Complexity for Russian Learners of English. In EUROCALL 2021.

[29] Carta, S. (2012). *Documentary Film, Observational Style and Postmodern Anthropology in Sardinia: A Visual Anthropology* (Doctoral dissertation, University of Birmingham).

[30] Eke, C. I., Norman, A. A., Shuib, L., & Nweke, H. F. (2019). A survey of user profiling: State-of-the-art, challenges, and solutions. *IEEE Access, 7*, 144907-144924.

[31] Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C., & Althoff, T. (2023). Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence, 5*(1), 46-57.

How to cite this article: Sonia Shahzadi, Sanjiv Kumar and Harsh Mehta (2023). Enhancing Depression Detection in Social Media via Advanced User Profiling and Fine-Grained Age Groups. *Bulletin of Computer and Data Sciences*, 4(1), 24-37. DOI: 10.71448/bcds2341-3

Received: 05/01/2023 **Revised:** 10/02/2023 **Accepted:** 29/03/2023 **Publish:** 30/04/2023

AND DATA SCIENCES BULLETIN OF COMPUTER *Bulletin of Computer and Data Sciences* is a peer-reviewed open access journal.