Personality Traits and Demographics Analysis in Online Mental Health Discourse

Mario Ezra Aragón*, Marcos Fernandez-Pichel, and David E. Losada

***Abstract*—Social media provides valuable insights into users' thoughts, behaviors, and emotions, offering opportunities for mental health research. In this work, we explore how personality traits and demographic attributes manifest in the online behavior of individuals suffering from mental disorders. Focusing on the Big-5 personality dimensions, we analyze social media users associated with four mental health disorders (Anorexia, Depression, Gambling, and Self-harm), investigating how these traits differ across groups. Using the PANDORA dataset –which supplies annotations for personality traits, age, and gender–, we train models for personality prediction and author profiling. These models are subsequently transferred to various eRisk collections. Besides confirming known trends (e.g., high association between anorexia and certain young female groups, or between gambling and young males), our analysis reveals intriguing personality traits. For example, we found high neuroticism and agreeableness, and low extraversion and conscientiousness shared across most disorders. These trends underscore the relevance of these personality traits for these mental health problems. Finally, we conclude by analyzing demographic biases in risk detection systems, showing that the accuracy of alerts differs significantly across demographic groups.**

***Index Terms*—Personality Analysis, Author Profiling, Big-5, Mental Disorders, Social Media, Bias Analysis.**

I. Introduction

UNDERSTANDING human personality has long been a cornerstone of psychological research, offering valuable insights into individual differences in behavior, emotion, and cognition [1]–[4]. The Big-5 model [3], [5], [6], one of the most widely accepted frameworks for studying human personality, conceptualizes personality along five broad dimensions: Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (commonly abbreviated as OCEAN). Each dimension represents a continuum along which individuals may vary [7]. These five dimensions provide a structured and empirically validated means of capturing the variability in human personality [8]. Big-5 traits have been found to be predictive of various life outcomes, including academic performance, interpersonal relationships, and physical and mental health. However, personality traits are not isolated from other factors (e.g., demographics). Gender and age, for instance, have been shown to influence both the expression of personality traits and the prevalence of mental health disorders [9]–[11]. As such, any investigation into the psychological correlates of mental illness must take into account the relative prevalence of disorders across gender and developmental stages, as well as how these factors may interact with personality.

In recent years, there has been a growing interest in exploring the intersection between personality traits and mental health disorders [12]–[15]. Research suggests that specific personality profiles may predispose individuals to mental health conditions such as depression, anxiety, bipolar disorder, and schizophrenia [16]. For instance, high levels of Neuroticism have been consistently associated with mood and anxiety disorders, while low Conscientiousness and Extraversion have been linked to depressive symptoms and social withdrawal. Understanding these associations has practical implications for early detection, prevention, and personalized treatment.

The proliferation of social media platforms has opened new avenues for psychological research by providing access to vast quantities of real-time, user-generated data [17]–[21]. People often express their thoughts, emotions, and behaviors on platforms like Twitter, Facebook, and Reddit, inadvertently revealing underlying personality traits and potential signs of psychological distress. Researchers can analyze linguistic patterns and behavioral indicators by leveraging computational methods such as Natural Language Processing (NLP) and Machine Learning to infer personality dimensions and possible mental health issues from digital footprints.

Despite the growing adoption of computational approaches in mental health research, concerns remain about the potential biases embedded in algorithmic models. In particular, gender and age can significantly influence the accuracy of model predictions. For instance, models trained on imbalanced demographic data may exhibit systematic performance disparities across groups, potentially under-representing or misclassifying symptoms expressed by women, men, or by individuals at certain life stages. These biases affect the validity of predictive models and raise ethical concerns when such tools are proposed for real-world mental health screening or intervention settings. Consequently, it is essential to evaluate whether classification outcomes are equitably distributed across gender and age groups and to account for these factors when interpreting the associations between personality traits and mental disorders.

In this work, we aim to explore the personality traits of individuals suffering from mental disorders as manifested in their social media publications. By examining how individuals with different diagnoses express themselves online, we can

(*) Corresponding author
The authors are with Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela (USC), Santiago de Compostela, Spain (e-mail: ezra.aragon@usc.es, marcosfernandez.pichel@usc.es, david.losada@usc.es).

better understand the personality dimensions that may be correlated with specific disorders. Additionally, we consider how gender and age may relate, acknowledging their critical role in shaping personality and mental health. This research not only contributes to theoretical knowledge at the intersection of Personality Psychology and Psychopathology but also holds promise for developing automated tools for mental health monitoring and intervention. We can summarize our main contributions as follows:

- We analyze the Big-5 personality traits of social media users suffering from four mental health disorders and perform additional author profiling on these individuals. This provides valuable insights into the psychological and demographic characteristics of people undergoing mental health difficulties and reveals intriguing patterns about how their problems reflect on social media.
- We study the bias of numerous classification systems that were designed to trace signs of risk from user-generated publications. This sheds light on the relative performance of their detection mechanisms when presented with data from individuals of different genders, age groups, or personalities.
- Using the PANDORA dataset [22], we train predictive models targeting personality traits, gender, and age, and validate their generalizability on the eRisk datasets, thereby supporting the viability of cross-domain personality and demographic analysis.

II. Related Work

Researchers in Psychology have extensively studied human personality as a main factor influencing individuals [1], [2], [4]. One of the most accepted frameworks is the Big-5 model. In a seminal work, Tupes and Christal identified five personality dimensions in reports analyzing co-workers' personality descriptions [5]. Goldberg first coined the term "Big-5" and the acronym OCEAN [6]. However, the current version of the model, in conjunction with the NEO Personality Inventory-Revised (NEO-PI-R) questionnaire, was popularized later by Costa and McCrae [3]. Another recognized model is the Myers-Briggs Type Indicator (MBTI), based on Carl Jung's psychological theory and consisting of eight personality types [23]. Questionnaires play a crucial role in personality research, representing validated tools for diagnosis and research.

Our work is closely related to Text-Based Personality Computing (TBPC), which aims to infer personality traits from user-generated text (e.g., social media publications) [24]. This research area builds upon early studies of language and personality, particularly those by Pennebaker [25]. As noted in the recent review by Fang et al. [26], TBPC has evolved significantly over time. Initial approaches relied on bag-of-words (BoW) models and psycholinguistic tools such as LIWC [27] and the MRC Psycholinguistic Database [28]. More recent methods use contextual embeddings that better capture semantic nuances. One of the foundational datasets in this domain is the Essays dataset, consisting of 2,468 English essays [29]. This corpus is frequently cited in early TBPC research for its richness and coherence, which enable the extraction of stable linguistic patterns. However, its formal and controlled writing style limits its ability to reflect natural, spontaneous user interactions.

As TBPC research has progressed, there has been a growing trend toward using large-scale datasets collected from social media [30]. For example, the PANDORA dataset contains Reddit posts from users annotated with personality traits [22]. PANDORA includes user labels for multiple personality models, including the Big-5, MBTI, and HEXACO, making it particularly valuable for diverse research applications. The original PANDORA study benchmarked multiple machine learning models for predicting personality and demographic attributes from Reddit text, exploring representations based on n-grams, psycholinguistic dictionaries, named entities, and part-of-speech features. These benchmarks served as reference baselines for training personality inference models in downstream social media studies. Another valuable resource is the dataset compiled by Rangel et al. [31], which includes multilingual data (English, Spanish, Italian, and Dutch) from X (formerly Twitter) and is annotated with Big-5 personality scores. Besides personality classification, scientists have increasingly explored the links between personality traits and mental health conditions [12]–[15]. Prior findings suggest that specific personality profiles may predispose individuals to certain disorders such as depression, anxiety, bipolar disorder, and schizophrenia [16]. However, recent work raises concerns about gender bias in personality prediction systems, highlighting the need for fairness-aware approaches [32]. Our research is also related to the author profiling literature [31], [33]–[35], which analyzes texts to infer authors' demographic or psychological characteristics. Related to this, we exploit the PANDORA dataset not only for personality trait prediction, but also for training classifiers of gender and age. In parallel, a separate line of research has focused on mental health risk detection from Reddit data. The eRisk shared tasks released curated Reddit collections for early detection of conditions such as depression, anorexia, self-harm, and gambling disorder [36]–[38]. Work on eRisk has primarily investigated classification architectures for early risk detection and temporal modeling of user posts, but these datasets do not provide explicit personality annotations.

Prior research on text-based personality and mental health risk detection has largely evolved along separate lines. A few studies have explored associations between personality traits and mental health conditions. However, they typically focus on single datasets, specific disorders, or rely on self-reported personality assessments. In this work, we go beyond these settings by proposing a cross-domain computational framework that transfers personality and demographic prediction models trained on PANDORA to multiple independent eRisk collections. To the best of our knowledge, this is the first work to conduct this specific exploration. This enables a unified and large-scale comparative analysis of Big-5 personality traits across four mental health conditions. Additionally, we examine demographic factors and systematically analyze biases in state-of-the-art early risk detection systems, thereby linking person-

ality analysis with issues of demographic representativeness and fairness in social media–based mental health research.

III. Methodology

Our methodology consists of two main stages: i) training models with a large base of social media users, and ii) transferring the models to monitor personality dimensions and demographic factors in users affected by diverse mental health conditions. Specifically, we leverage the annotated PANDORA dataset [22] to build predictive models and apply them to the eRisk dataset for analyzing users with mental disorders. The overall process is illustrated in Figure 1.

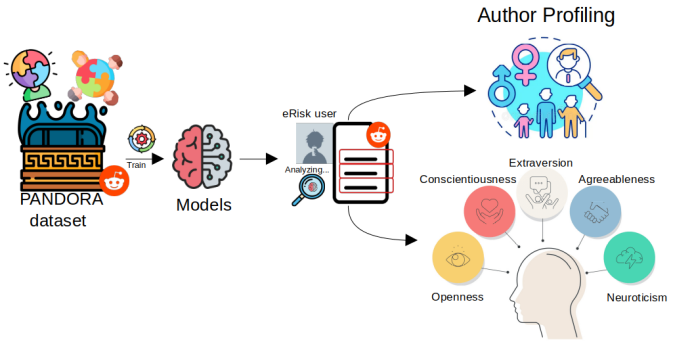

Fig. 1: The framework involves training models with the PANDORA dataset, followed by predicting demographic and personality traits among users from the eRisk data collections.

A. Training models using PANDORA

We begin by training machine learning models on the PANDORA dataset, a Reddit-based corpus with user-level annotations for personality traits, age, and gender. The PANDORA dataset [22] is constructed by linking Reddit users to externally completed personality questionnaires: dataset creators collected users who publicly shared links to their personality test results or self-reported personality scores, and associated their Reddit posting histories with these questionnaire-derived labels. For the Big-5 model, PANDORA provides normalized trait scores in the 0–100 range (in increments of 10), derived from standardized personality inventories. Demographic attributes such as age and gender are obtained from explicit self-disclosures in user posts or profiles. The final dataset comprises 1,600 users annotated with Big-5 traits, with subsets additionally labeled for age and gender, as reported in the original PANDORA study.

PANDORA is a well-known resource in computational personality analysis and also supports other personality models (e.g., MBTI and Enneagram). Its significance lies in enabling large-scale personality analysis across platforms. The creators of PANDORA demonstrated its utility through three experiments: i) using MBTI and Enneagram labels to infer Big-5 traits, ii) detecting bias in gender classification models based on associated personality traits, and iii) revealing how psycho-demographic variables can predict interests among Reddit users.

We extract multiple features from the users' posts, such as linguistic cues, lexical patterns, and embedding-based representations, and we exploit these features to train models that predict the Big-5 personality traits and demographic attributes such as age group and gender.

Our first analysis focuses on the task of gender prediction. Figure 2 shows the gender distribution in the original dataset.[1] We reproduced the same classical prediction methods implemented in [22]: a Support Vector Machine (SVM) classifier with TF-IDF representations of word n-grams. The original paper reported a test accuracy of 89.3% in gender classification. However, the dataset does not provide predefined training and testing splits. We performed 5-fold cross-validation using the entire dataset, obtaining an average accuracy of 87.5%, which demonstrates the model's reliability. Consequently, we adopted the resulting model to perform gender classification in our mental health dataset.

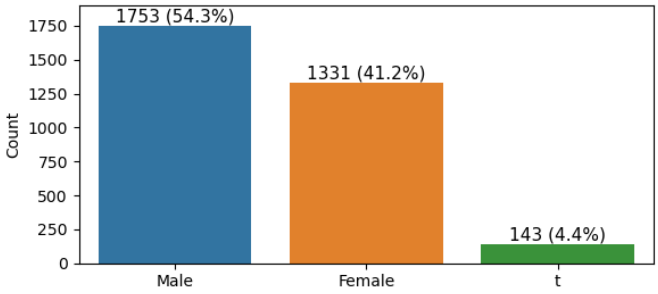

Fig. 2: PANDORA dataset gender distribution

The second analysis targets age prediction. In contrast to gender, the original PANDORA paper did not approach age prediction as a classification task. Instead, age labels were used as supplementary demographic information, primarily in bias-related experiments. For our purposes, we formulated age prediction as a multi-class classification problem and binned users into four age groups: 10–19, 20–29, 30–39, and 40+. We show the age distribution in Figure 3. We trained a multi-class SVM classifier using TF-IDF representations of word n-grams and evaluated the model using 5-fold cross-validation. The resulting weighted average F1-score was 0.618. While overall performance is limited, primarily due to class imbalance, the model nonetheless captures relevant age-related patterns, supporting its potential as a meaningful predictor of age.

In our final analysis, we examine the prediction of personality traits using the Big-5 model. For this task, we treated personality estimation as a regression problem and trained five models to predict the scores for each trait. We experimented with a wide range of model architectures and input representations, from traditional methods such as Bag-of-Words to advanced neural models, including Convolutional Neural Networks (CNNs), Long-Short-Term Memory networks (LSTMs), Bidirectional LSTMs (with and without attention

[1]The original dataset included a third gender label denoted as "t", which the original authors neither defined nor discussed. Due to the lack of information about this label, we discarded these cases and built a two-class classifier from the remaining examples.

TABLE I: Regression results for Big-5 traits over the original data (PANDORA). The table reports the Mean Squared Error (MSE), the Mean Absolute Error (MAE), and the Pearson correlation coefficient. "Paper best" shows the best Pearson scores reported in the original study.

	MSE			MAE			Pearson			
	BoW	**DisorBERT**	**mpnet**	**BoW**	**DisorBERT**	**mpnet**	**BoW**	**DisorBERT**	**mpnet**	**(Paper best)**
Openness	**0.074**	0.090	0.091	**0.220**	0.256	0.253	**0.310**	0.135	0.115	0.265
Conscientiousness	**0.091**	0.098	0.107	**0.249**	0.263	0.267	**0.268**	0.158	0.084	0.273
Extraversion	**0.088**	0.102	0.104	**0.244**	0.261	0.271	**0.326**	0.167	0.073	0.387
Agreeableness	**0.096**	0.106	0.108	**0.260**	0.276	0.281	**0.274**	0.129	0.109	0.270
Neuroticism	**0.108**	**0.108**	0.117	**0.279**	0.284	0.295	**0.232**	0.188	0.127	0.283

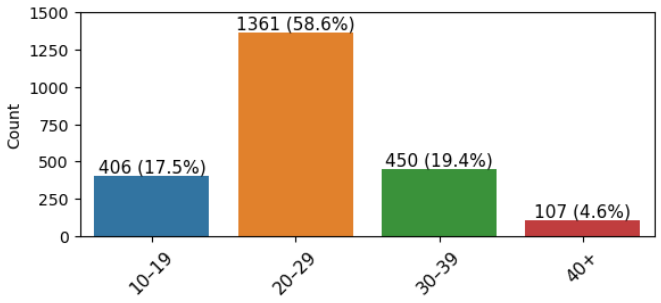

Fig. 3: PANDORA dataset age distribution.

mechanisms), and Transformer-based models. We selected the three top-performing models based on validation performance from these experiments. The experimental settings section details the specific configurations of these models, including hyperparameters and training procedures.

We begin our analysis by examining the distribution of personality traits in the PANDORA dataset. Figure 4, displays the distribution for each of the Big-5 personality traits. Each subplot represents the number of users (y-axis) assigned a given trait probability (x-axis).

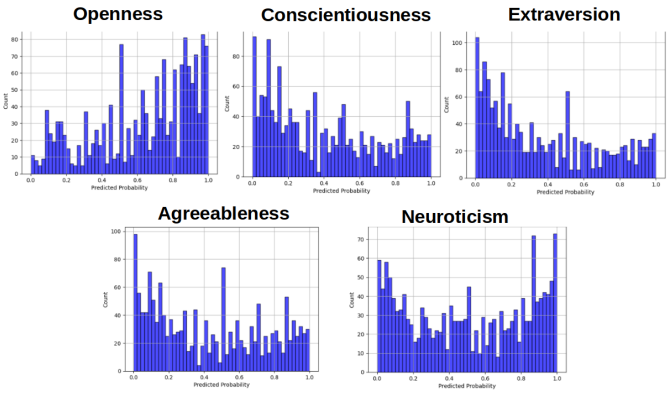

Fig. 4: PANDORA dataset Big-5 distribution.

The distribution for Openness is left-skewed, indicating many users were predicted to have high levels of openness. In contrast, Conscientiousness, Extraversion, and Agreeableness all exhibit right-skewed distributions, suggesting that lower levels of these traits are more prevalent among users in the dataset. Notably, Neuroticism exhibits a bimodal distribution, with peaks at high and low probability levels, which may reflect divergent emotional stability levels within the Reddit user base. These distributions align with established patterns of online behavior. Platforms such as Reddit commonly attract user populations that display greater openness to experience but reduced levels of agreeableness and conscientiousness. This initial trait profiling serves as a foundation for further downstream analyses and facilitates transfer to additional datasets, including eRisk.

Table I presents the regression results on the original PANDORA dataset for the three best-performing models –BoW (Bag-of-Words), DisorBERT, and mpnet– evaluated using Mean Squared Error (MSE), Mean Absolute Error (MAE), and Pearson correlation coefficient. Across all traits and metrics, the BoW model consistently outperforms both DisorBERT and mpnet. Specifically, BoW achieves the lowest MSE and MAE values across all five traits, indicating better capacity to estimate trait scores. Regarding Pearson correlation, which captures the linear association between predicted and ground-truth values, BoW obtains the highest scores across all dimensions. For example, BoW achieves a Pearson score of 0.326 for Extraversion, compared to 0.167 (DisorBERT) and 0.073 (mpnet). This pattern is consistent across the remaining traits, suggesting that BoW predicts values closer to the ground truth and preserves the correct relative ordering of samples more effectively than the other models. The table includes a "Paper best" column reflecting the highest reported Pearson correlation for each trait in the original PANDORA study [22]. These values are taken directly from the original publication and are included for reference, while all other results in the table correspond to our reproduced 5-fold cross-validation experiments on the full PANDORA dataset following the same evaluation protocol. The authors used multiple textual representations in the original study, including n-grams, psycholinguistic dictionaries, named entities, and part-of-speech patterns. The goal of this experiment is not merely to replicate prior results, but to identify a reliable personality-prediction model that can be transferred to the eRisk datasets. Notably, our BoW variant outperforms this model on two traits (underlined scores) and, overall, is competitive with the best Big-5 predictor proposed by the authors of PANDORA. For instance, the best previously reported Pearson score for Extraversion is 0.387, while BoW achieves a competitive 0.326 without task-specific fine-tuning or additional supervision. Despite its simplicity, BoW exhibits solid robustness, as evidenced by this comparison.

In contrast, DisorBERT and mpnet, which are embedding-based sentence representation models, underperform significantly. Although these models have shown promise in a range of NLP tasks, their relatively poor performance in this setting suggests a mismatch between their sentence embeddings models and the task of predicting personality traits. This underperformance may be attributed to the lack of fine-tuning on domain-specific data and the potential inadequacy of pre-trained semantic representations to capture subtle personality cues in text.

Among the five traits, Extraversion consistently yields the highest Pearson correlations across all models, implying that linguistic indicators of Extraversion are more salient and readily learnable. In contrast, traits like Neuroticism and Agreeableness exhibit lower correlations, indicating that these personality dimensions are more challenging.

These findings lead to some key insights. First, the widely cited quote "**Simplicity is the ultimate sophistication**" proves to be applicable here. The performance of the BoW model reinforces the value of employing simpler, more interpretable methods in personality prediction tasks. Lexical features appear well-suited to capturing personality-related signals in text. Second, **neural embeddings are suboptimal:** the relatively poor results of the embedding representations suggest that off-the-shelf embeddings lack the granularity or domain alignment required for accurate personality assessment. Third, **model generalization matters:** the Pearson metric reveals that while errors (MSE, MAE) might be comparable, BoW better captures the ranking and relative order of cases in personality trait analysis. This indicates a potential generalization gap in embedding models concerning ordinal consistency.

B. Cross-domain prediction

Once the models were trained using the PANDORA dataset, we transferred them to monitor users from a collection in a different domain, focused on early risk detection (eRisk). The eRisk datasets contain publications from Reddit users, who are grouped into a positive group (users suffering from a certain mental health condition) and a control group (general users from the platform). Unlike PANDORA, the eRisk datasets do not include explicit annotations for personality traits or demographic information. Therefore, we leverage the models trained with PANDORA to infer these attributes from the posting histories of the eRisk users. This allows us to extract estimates of personality dimensions (Big-5 traits), age groups, and gender.

This stage aligns with the broader task of author profiling, which involves characterizing individuals based on both psychological and demographic dimensions. The inferred personality traits offer insights into the users' cognitive and emotional patterns, shedding light on how certain characteristics correlate with specific mental health conditions. Similarly, demographic predictions –particularly those related to age and gender– enable subgroup analyses, revealing population-specific trends or disparities.

By combining psychological and demographic profiling, we aim to build a comprehensive representation of users within the eRisk collection. This holistic perspective allows for a more nuanced interpretation of how mental health challenges manifest in online behavior and supports more targeted and ethical computational mental health research on social media platforms.

IV. Experimental Settings

A. Test Datasets (eRisk)

We employed datasets from the eRisk evaluation tasks conducted between 2017 and 2023 [36]–[42]. These initiatives were designed to promote the early detection of signs of mental health disorders such as anorexia, depression, gambling addiction, and self-harm. In Table II, we present an overview of the datasets, including the distribution of classes.

TABLE II: eRisk datasets users distribution.

	Positive	Control
Anorexia	134	1153
Depression	312	2795
Gambling	348	6253
Self-harm	297	1914

The eRisk datasets contain numerous publications from Reddit users (including comments and posts). Each dataset is divided into two categories: i) users identified as experiencing one of the target conditions (anorexia, depression, gambling, or self-harm), and ii) a control group made up of individuals not affected by these disorders. Positive cases were determined based on explicit self-reports of a clinical diagnosis (textual instances in which users stated that a medical professional had diagnosed them).[2] The control group includes random users from various subreddits and other individuals who discuss the target conditions, including clinicians who offer support and guidance in mental health communities. This means simple keyword-based detection is insufficient; effective systems must interpret users' behavior and context more deeply. Such data collection methodology thus results in test collections that better capture real-world complexity and variability.

B. Training Configuration (PANDORA)

Preprocessing: We began with basic text preprocessing, converting all words to lowercase and removing elements such as URLs, emoticons, and hashtags. This step was designed to eliminate non-essential components that do not contribute meaningfully to our analysis.

Training and prediction: For the age and gender prediction tasks, we obtained the complete representation of the post history of the users using TF-IDF of word unigrams and trained an SVM classifier[3].

For the regression task (Big-5), we used unigrams and bigrams and selected the most relevant features (200,000) employing

[2]Vague statements such as "I think I have depression" or "I am depressed" were not treated as sufficient evidence of diagnosis.

[3]The GitHub link will be shared upon acceptance.

the SelectKBest method (from the sklearn library [43]). The model is a feedforward neural network with three fully connected layers. It maps the input features to 128 units, then to 64, and finally to a single output. Each hidden layer employs ReLU activation and a 0.3 dropout rate for regularization. A sigmoid activation in the output layer enables binary classification. We used the Adam optimizer with a learning rate of $1e^{-3}$ and binary cross-entropy as the loss function. For the embedding models, the best approach is to encode the entire sentences for each user and then compute the mean, max, and min values across the entire post history.

Parameters: We employed the frameworks offered by scikit-learn v1.7 [43], HuggingFace v4.24.0 [44], and PyTorch v1.13.0 [45] for the downstream task's training and conducted the training over 100 epochs on a GeForce RTX 3070 with 8GB memory.

C. Classification Models

This subsection describes the alternative representations used for the different tasks.

Bag-of-Words: (BoW) We implemented a traditional Bag-of-Words model using unigram (and bigrams for regression) features weighted by TF-IDF. These features were classified using an SVM with a linear kernel. While we initially tested other kernels and conventional classifiers, the linear SVM consistently outperformed them. As a result, we selected it as our reference baseline for traditional methods.

DisorBERT: DisorBERT [46] is a customized variant of BERT that underwent a dual-domain adaptation process. First, the model was adapted to social media language to capture informal and conversational textual patterns. Next, it specializes in the mental health domain, enhancing its ability to recognize domain-specific terminology and linguistic patterns related to psychological disorders. Both adaptation steps used a domain-specific lexicon to guide the masking process, ensuring the model focuses on relevant language features during training.

Mpnet: It is a pre-training method [47] that exploits the advantages of masked language modeling and permuted language modeling for Natural Language Understanding.

Given its simplicity, we experimented with BoW for the three tasks, while DisorBERT and Mpnet were used exclusively for the personality analysis task.

V. Evaluation Results and Discussion

Figure 5 presents the gender predictions across several mental health categories of the eRisk dataset using the best performing models (discussed in Section III). We can observe a consistent trend: the model predicts the female category as the most prevalent in most of the positive classes (i.e., users suffering from a mental disorder). In contrast, the male category tends to dominate within the control classes. The gambling dataset represents the only exception to this trend. Specifically:

- In the Anorexia dataset, **76%** of positive cases are predicted as female and **24%**, compared to only **9%** female and **91%** male in the control group. This imbalance naturally derives from the characteristics of individuals suffering from this eating disorder.
- In Depression, **58%** of positive users are classified as female and **42%** as male, whereas the control group is **87%** male and **13%** female.
- For Gambling, positive cases are overwhelmingly male (**91%**), and the control group also skews male (**82%**).
- In the Self-harm datasets, **68%** of positive users are predicted as female and **32%** as male, while the control group consists of **86%** male predictions.

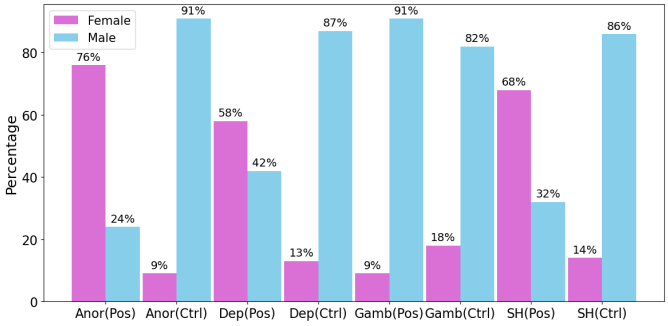

Fig. 5: Gender predictions over the eRisk collections

Figure 6 displays the age distribution across the data collections. The age distribution across positive cases indicates that the majority of users are concentrated in the 20–29 age range, particularly within Gambling (80%), Anorexia (76%), and Depression (69%). Self-harm presents a notable deviation from this pattern, with a comparatively larger proportion of younger users (41% aged 10–19) alongside 54% in the 20–29 group.

The control groups, in contrast, tend to show a less skewed age distribution, with more presence of people from the two oldest groups (30-39 and 40+). In any case, all age distributions show a strong presence of young individuals, which is a natural feature of social media activity.

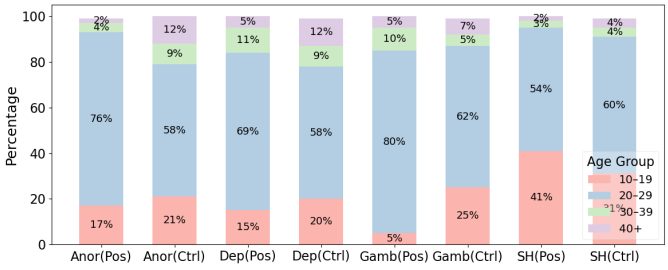

Fig. 6: Age predictions over the eRisk collections

A. Comparison with Real-World Trends

The gender and age distributions inferred by our models closely mirror well-established mental health patterns, supporting the validity of the datasets and the models used.

Gender Differences: The gender predictions are consistent with trends published in the mental health literature:

- Anorexia is more prevalent among females and it particularly affects adolescents and young adults (epidemiological data suggest that most anorexia cases occur in females[4]). Our model reflected this, with 76% of positive cases classified as female.
- Women experience higher rates of depression diagnoses, shaped in part by underlying psychological and societal dynamics. The 58% female representation in depression-positive cases aligns well with global prevalence rates.[5]
- Gambling disorders typically have a higher prevalence in men and are often associated with impulsivity and risk-taking factors [48]. Again, our models were consistent with this trend, predicting 91% of positive gambling users as male.
- Adolescent females have been shown to exhibit higher rates of non-suicidal self-injury compared to their male peers [49]. The 68% female classification in self-harm positive cases supports this pattern.

The respective control groups are estimated to be overwhelmingly populated by males. This partially reflects the demographic imbalance in Reddit usage (gender-based data from 2025 indicate that males make up a substantially greater portion of users compared to females).[6]

Age Trends: Our age predictions also align with known age-related vulnerabilities [50]:

- Many mental health disorders begin before the age of 30, particularly during adolescence and early adulthood.[7] The concentration of anorexia, depression, and self-harm in the 10–29 year-old range is consistent with findings from global mental health surveys and clinical observations.
- The gambling data show a broader age distribution, including 18% of the cases in the 40+ range, which may reflect the chronic nature of gambling behaviors or delayed disorder manifestation [51].
- Self-harm peaks in teenage years, explaining why 41% of the positive users fall into the 10–19 age group.

These findings confirm the predictive capability of our models, with gender and age predictions from the PANDORA-trained models on the eRisk collection aligning with known real-world distributions of mental health disorders. This underscores the potential of demographic inference in understanding the context of mental health within online communities.

B. Big-5 Analysis in Mental Disorders

For our following analysis, we leverage the best-performing model and present an analysis of the distribution of the Big-5 personality traits among individuals with mental health disorders, including anorexia, gambling, depression, and self-harm, compared to control groups. The box plots (Figures 7, 8, 9, 10, and 11) illustrate the predicted probabilities for each trait across the disorders, revealing distinct patterns and trends. To provide statistical evidence on the difference between the positive and control group, we apply the Mann–Whitney U test and report in the figures the resulting p-values. The Mann–Whitney U test is a non-parametric statistical test used to determine if there is a significant difference between the distributions of two independent groups. A small p-value (usually less than 0.05) indicates sufficient evidence to claim that the positive group differs systematically from the control group.

Key Trends and Observations:

Openness:

1) There is no consistent trend across disorders. Some groups (e.g., anorexia) show slight elevations, while others (e.g., depression) exhibit minimal differences from controls.

2) This suggests openness may not be a primary trait linked to these specific disorders or may interact with other factors.

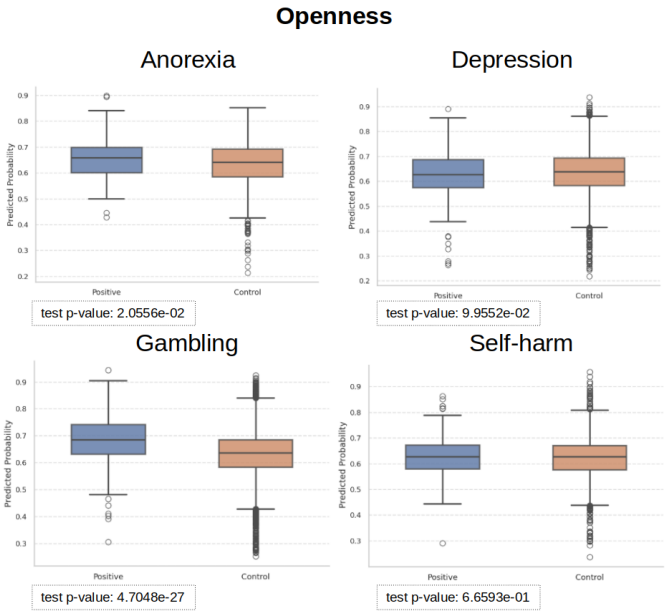

Fig. 7: Openness distribution over the eRisk collections.

Conscientiousness:

1) Lower levels are observed in individuals from the positive groups of the four disorders. This indicates impulsivity and reduced self-regulation associated with these groups of individuals.

2) Control groups not only show higher medians but also more variability and extreme cases.

Extraversion:

1) The positive groups (with the exception of Gambling) show lower levels of extraversion compared to the respective control groups. This supports the idea that social withdrawal and low positive affect are features associated with anorexia, depression, and self-harm.

[4]https://withinhealth.com/learn/articles/anorexia-nervosa-an-statistics-gender-race-and-socioeconomics

[5]https://www.who.int/news-room/fact-sheets/detail/depression

[6]https://explodingtopics.com/blog/reddit-users

[7]https://www.brainsway.com/knowledge-center/depression-across-age-groups/

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2026.3674676

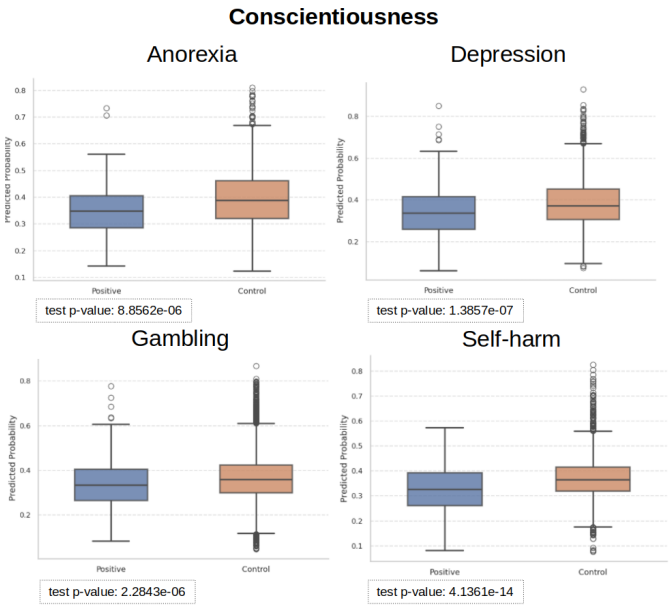

Fig. 8: Conscient. distribution over the eRisk collections.

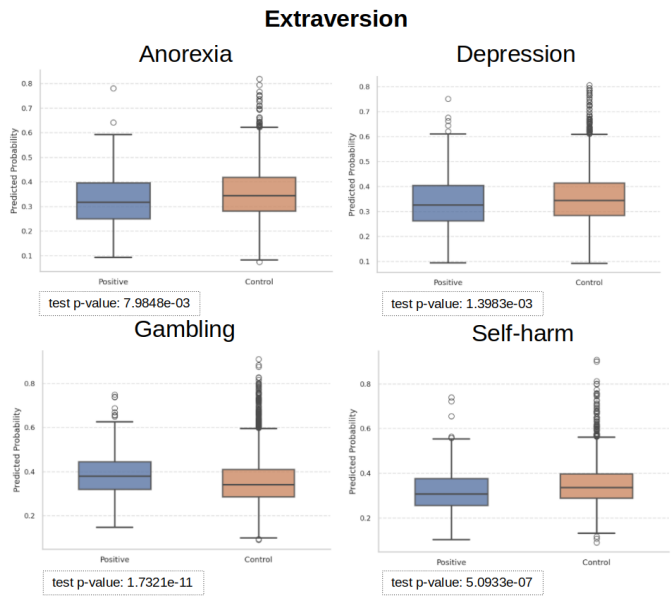

Fig. 9: Extraversion distribution over the eRisk collections.

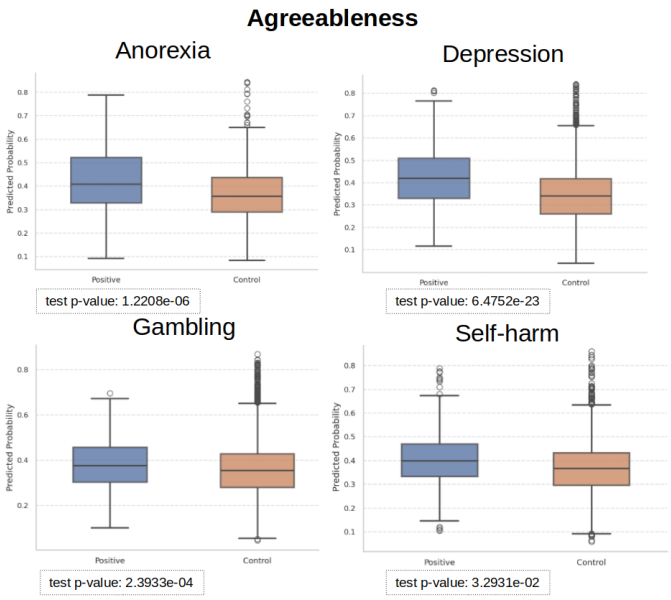

Fig. 10: Agreeableness distribution over the eRisk collections.

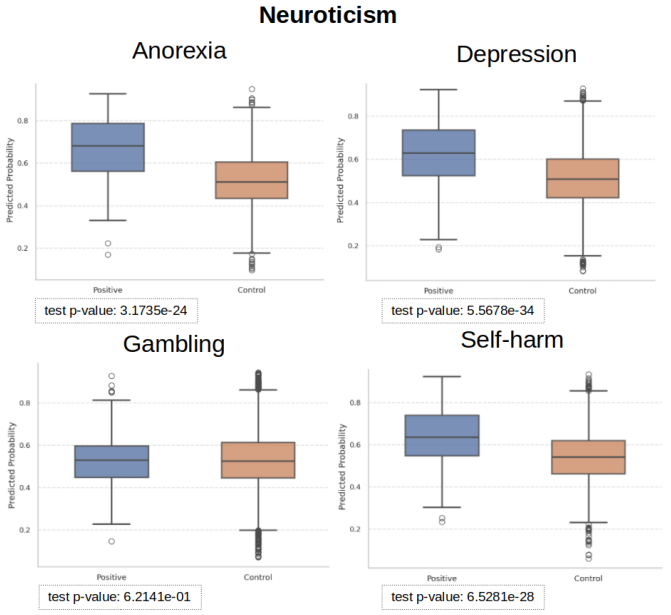

Fig. 11: Neuroticism distribution over the eRisk collections.

Agreeableness:

1) For the four disorders, the levels of agreeableness are higher for the positive groups, compared to their control group counterparts.

Neuroticism:

1) This trait is consistently elevated across most disorders (anorexia, depression, self-harm) compared to controls, suggesting a strong association between high neuroticism and these mental health conditions.

2) The difference observed aligns with existing literature linking neuroticism to emotional instability, vulnerability, and mood disorders.

Summing up, we found high neuroticism & agreeableness but low extraversion & conscientiousness shared across most disorders. These trends underscore the relevance of these personality traits for these mental health problems.

Let us compare these findings with studies about the Big-5 personality traits in the literature. Some common patterns exist between our analysis and the trends discussed in previous studies. We found elevated **Neuroticism** across all disorders, aligning with the well-established role of this trait as a predictor of poor mental health outcomes. In fact, individuals with high levels of Neuroticism are more likely to experience negative emotions, have difficulty coping with stress, and may struggle with impulsivity and psychological distress [15]. Similarly, our

findings on low **Conscientiousness** among individuals with psychological disorders align with the literature. In clinical studies, people who score low in Conscientiousness have been found to have worse psychological outcomes [15], [52]–[54]. In the literature, a negative association between **Extraversion** and psychological disorders was found [15] (extroverts tend to enjoy social interactions, experience positive emotions more easily, and have better mental health outcomes). In our study, the positive groups (except Gambling) show lower extraversion levels than the respective control groups. Regarding **Openness**, we find inconsistent results in the literature. Kang et al. [15] argued that low Openness levels are associated with depression and anxiety, but [55] suggested a positive association between this trait and depression. Our results for this trait are also inconclusive, as we did not observe clear differences between the positive and control groups. Finally, our results on **Agreeableness** differ from the findings from prior research, which typically report low or unaffected levels in people suffering from psychological problems. Our data reveal otherwise: individuals in the positive groups tend to show higher levels of Agreeableness. We hypothesize that this might be due to the peculiarities of social media users participating in mental health support communities. Users on social networks (especially Reddit communities related to mental health) tend to exhibit a heightened sense of community and social cohesion, thus reflecting higher levels of agreeableness in their interactions with other social media users.

VI. Bias Analysis

Understanding how machine learning models behave across different demographic groups is essential to ensure fair and reliable outcomes. In particular, bias analysis helps uncover whether certain user groups benefit disproportionately from better performance, while others may be systematically disadvantaged. With this motivation in mind, we carried out a bias analysis on the eRisk data collections used in our study.

To conduct this analysis, we analyzed the predictions of the different systems submitted by the participants of eRisk.[8] Our goal was to investigate whether these predictive systems exhibited biases in favor of specific demographic groups (e.g., users of a particular gender or age). To this end, we first labeled the positive group of each collection using the gender and age classifiers described above (see Section III). Next, we split the positive group into demographic categories (e.g., female positive users and male positive users) and evaluated the performance of each system within each group using the F1 metric. This allowed us to identify which user groups were more easily and accurately classified, thereby shedding light on potential biases in model behavior.

Figure 12 shows the F1 scores achieved by participant models across three eRisk tasks broken down by gender. Each system is represented as a point (x, y) whose coordinates are the F1 values for males and females, respectively. The diagonal line represents equal performance for male and female users. Points above the diagonal indicate higher performance for positive female users, while points below indicate better detection of positive male users.

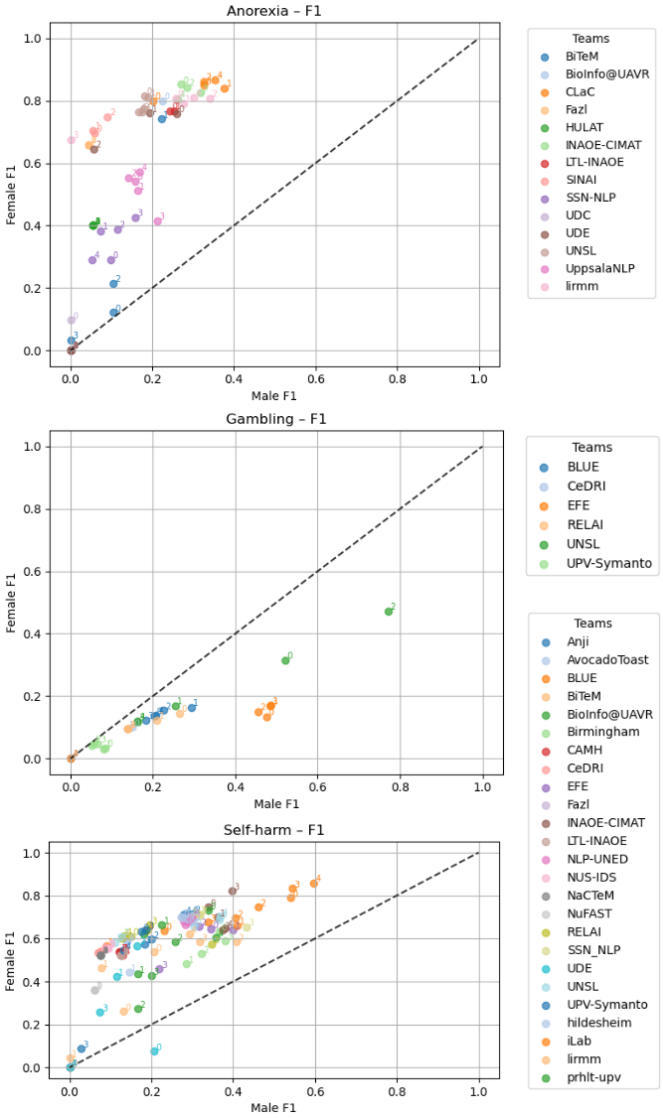

Fig. 12: Gender bias of the participants over different eRisk collections.

In anorexia and self-harm, all systems show a clear tendency to be more accurate on the female cases (most points cluster above the diagonal, often with significant performance gaps compared to those achieved for the male cases). This means that, when a user exhibits signs of anorexia or self-harm, the risk is more likely to be detected if the user is a woman. For gambling, instead, the systems show higher accuracy for the positive male cases.

These results suggest that models trained on eRisk collections are biased toward specific gender groups. One possible explanation relates to the underlying data distribution: disorders such as anorexia and self-harm have a higher prevalence

[8] We would like to thank the organizers of eRisk for providing us with access to these predictions. Specifically, we had access to the system's predictions for three of the eRisk tasks (self-harm, anorexia, and gambling). In the case of self-harm, we had access to systems' predictions for several editions of the task. In the case of anorexia and gambling, we had access to systems' predictions for one edition.

among women, both in epidemiological studies and in online communities. Consequently, models could implicitly learn patterns more strongly associated with female users, reducing sensitivity for male cases. We can observe this trend in Figure 5: for instance, 76% of anorexia-positive users and 68% of self-harm–positive users are female, while control groups are predominantly male (e.g., 91% of anorexia controls and 86% of self-harm controls). Such imbalances likely reinforce gender-specific signal learning, amplifying disparities in model performance.

Figure 13 presents the distribution of F1 scores across different age groups for the anorexia, gambling, and self-harm tasks. Each boxplot summarizes the performance of the participating systems for a given age group, allowing us to compare how consistently systems detect positive cases across the lifespan. For anorexia and self-harm, the systems perform best for users in the first two age groups. The most accurate predictions for gambling are in the ranges 20-29 and 30-39. As with gender, this might be related to the distribution of positive and negative cases across age groups.

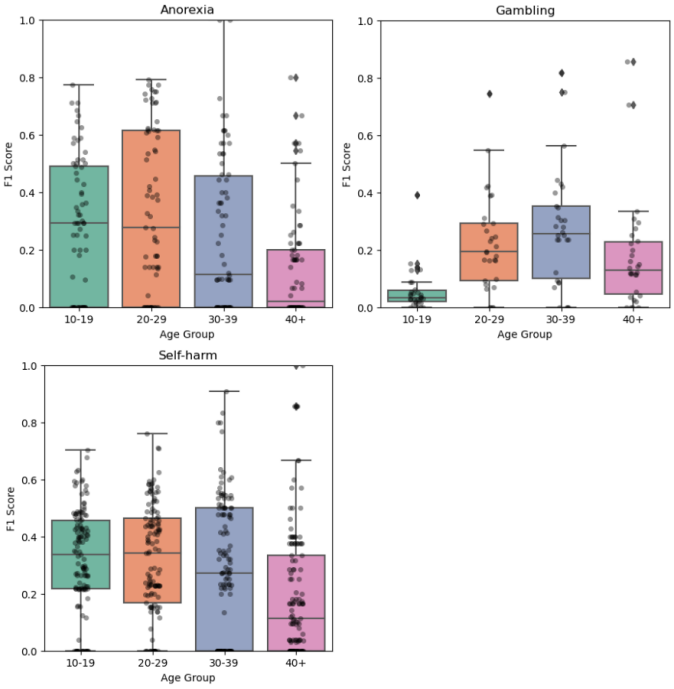

Fig. 13: Age bias of the participants over different eRisk collections.

Overall, this analysis indicates a potential age-related bias in detection effectiveness. Younger and middle-aged users suffering from these disorders are more accurately identified, whereas older users are consistently harder to detect across tasks. This may stem from differences in language use: younger individuals might express symptoms or risky behaviors more openly, making them easier for models to capture. At the same time, older users may communicate in subtler or less typical patterns, leading to under-detection. This interpretation aligns with the dataset distribution (see Figure 6): eRisk collections are dominated by younger participants, while older users are comparatively underrepresented. As a result, models are more exposed to linguistic patterns typical of younger groups during training, thereby reinforcing their sensitivity to these age groups and reducing their effectiveness in later adulthood cases.

Our findings show that eRisk systems tend to perform better for female and younger users, while male and older users are comparatively harder to detect. From a medical perspective, this is concerning, as it mirrors well-known clinical gaps [56]: men are often underdiagnosed, and mental health issues in later adulthood frequently go unnoticed or untreated. For gambling, the consistently low performance across age groups highlights both the difficulty of the task and the diverse profiles of those affected. These disparities emphasize the need for balanced datasets and bias-aware methodologies, ensuring that early detection systems can capture the linguistic and behavioral markers of all demographic groups and provide equitable support.

VII. Limitations and Ethical Considerations

Here, we highlight some of the most critical limitations of our research:

Dataset Biases and Generalizability: The PANDORA and eRisk datasets are derived from Reddit, a platform with well-known demographic and participation skews. This may limit the generalizability of our findings to broader populations, including older adults, non-English-speaking users, and individuals from marginalized, culturally diverse, or lower-connectivity communities who are less likely to participate in Reddit discussions or to express mental health concerns in comparable ways. Moreover, linguistic markers of personality and psychological distress can vary across cultures and social contexts. Annotations rely on self-reported or inferred labels, which may introduce noise or inaccuracies in personality and mental health classifications. Therefore, future work should examine these trends using datasets from other sources and, ideally, written in multiple languages. In addition to demographic skews, Reddit exhibits platform-specific dynamics that may further affect data representativeness. Certain subreddits are shaped by political or ideological orientations, content visibility is influenced by reward mechanisms, and a non-negligible fraction of accounts may be automated or semi-automated. These factors can alter posting behavior, topic prevalence, and linguistic style, thereby introducing additional noise or biases in the inferred personality and mental health signals.

Model Performance: While the Bag-of-Words (BoW) model outperformed advanced neural architectures in personality prediction, its reliance on lexical features may miss subtle linguistic cues (e.g., sarcasm, context) relevant to mental health. Age- and gender-prediction models demonstrated modest accuracy (e.g., F1-Scores of 0.618 for age groups), suggesting potential for improvement in demographic inference. Since the eRisk datasets do not provide demographic ground truth labels, we conducted a small-scale manual check to assess the plausibility

of cross-domain gender predictions transferred from PANDORA. For each mental health condition, we selected the five users with the highest predicted probability of being male and the five with the highest predicted probability of being female, as determined by the PANDORA-trained gender classifier. Two people independently annotated the gender of these users based on their Reddit posts. The inter-rater agreement (Cohen's κ) between the two annotators is reported in Table III. Furthermore, we observed high agreement between human annotations and model predictions (ranging from 80% to 100% correct predictions per dataset). These results support the use of transferred demographic predictions for exploratory analysis, while underscoring the need for annotated target-domain data to rigorously evaluate cross-domain generalization.

TABLE III: Cohen's inter-rater agreement.

Task	**Anorexia**	**Depression**	**Gambling**	**SH**
Cohen's κ	0.80	1.00	0.80	0.78

Cross-Dataset Variations: Applying models trained on PANDORA to eRisk assumes that language patterns are comparable across datasets. Divergences in user behavior or platform-specific norms (e.g., subreddit communities) could affect this transferability. Although both datasets originate from the same source, we will continue validating these models to ensure their transferability across different communities, forums, and user segments.

Trait Interpretation: The study has focused on broad Big-5 dimensions and general demographic characteristics, potentially overlooking disorder-specific substrates or interactions between traits. Therefore, our current work will be extended by developing more fine-grained methods that can, for instance, track more specific psychological and personality traits.

The nature of our work presents the following *Ethical Considerations*:

Privacy and Consent: We acknowledge the ethical challenges of analyzing social media content, particularly regarding user privacy. To address these concerns, our study exclusively utilized publicly available, pre-existing datasets. At no point did we engage directly with or contact social media users. The dataset included only public user interactions, which were used in full compliance with the original platforms' terms of service and user agreements.

Stigmatization and Harm: Associating personality traits with mental disorders risks reinforcing stereotypes (e.g., labeling individuals with high neuroticism as "unstable"). Our findings should be framed cautiously to avoid deterministic interpretations. Furthermore, predictive models could be misused for surveillance or profiling vulnerable groups. Therefore, the actual implementation of this type of analysis should fully consider all these ethical aspects and incorporate the corresponding safeguards.

Bias and Fairness: Gender and age prediction imbalances are a natural consequence of the biases in training data or platform demographics (e.g., male-dominated control groups). Future work should continue to audit models for fairness across subgroups. The exclusion of the ambiguous gender label ("t") may inadvertently marginalize non-binary users, and, therefore, our work needs to be complemented by new studies that take a broader account of the gender identities present in our society. Finally, the exclusion of underrepresented groups from both data collection and model evaluation may inadvertently reinforce existing disparities in mental health research and technology-based interventions. Ensuring that computational mental health tools are developed and validated across diverse cultural and social contexts remains an essential challenge for future research.

Clinical Utility: While this study has identified some correlations, it cannot establish causality between traits and disorders. A formal clinical screening would require further validation against diagnostic criteria and direct involvement of mental health professionals.

VIII. Conclusion

This study has explored the intersection of Big-5 personality traits and mental health disorders through the lens of social media language, leveraging the PANDORA and eRisk datasets to analyze users with anorexia, depression, gambling disorder, and self-harm tendencies. Our findings revealed distinct personality profiles associated with these conditions. For example, we found notably elevated neuroticism and reduced extraversion across disorders, as well as lower conscientiousness and agreeableness in specific groups (e.g., gambling and self-harm). Demographic trends derived from our data, such as the predominance of younger females in anorexia and self-harm communities, are well aligned with the clinical literature. This alignment underscores the value of these computational methods for mental health studies. However, dataset biases (e.g., Reddit's demographic skew) and the limited performance of certain profiling models point to the need for more diverse data and improved modeling approaches. Future work could also include conducting longitudinal studies to assess the stability of personality traits and explore causal relationships, as well as expanding this research to underrepresented disorders and populations.

Acknowledgement

We thank the support obtained from MICIU/AEI/10.13039/501100011033 (PID2022-137061OB-C22, supported by ERDF) and Xunta de Galicia-Consellería de Cultura, Educación, Formación Profesional e Universidades (ED431G 2023/04, ED431C 2022/19, supported by ERDF). The authors also acknowledge the project "Cátedra de IA aplicada a la Medicina Personalizada de Precisión" (Cátedras ENIA, TSI-100932-2023-3); Cátedras ENIA is funded by the Ministerio de Transformación Digital y Función Pública (Secretaría de Estado de Digitalización e Inteligencia Artificial); and by the NextGeneration EU-fund. The first author also thanks the support obtained from the Juan de la Cierva Grant (JDC2023-052296-I), funded by MCIN/AEI/10.13039/501100011033 and by the FSE+.

REFERENCES

[1] W. Mischel and Y. Shoda, "A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure." *Psychological review*, vol. 102, no. 2, p. 246, 1995.

[2] P. T. Costa and R. R. McCrae, "Trait theories of personality," in *Advanced personality*. Springer, 1998, pp. 103–121.

[3] P. T. Costa Jr and R. R. McCrae, *Neo Personality Inventory.* American Psychological Association, 2000.

[4] J. M. Digman, "Higher-order factors of the big five." *Journal of personality and social psychology*, vol. 73, no. 6, p. 1246, 1997.

[5] E. C. Tupes and R. E. Christal, "Recurrent personality factors based on trait ratings," *Journal of personality*, vol. 60, no. 2, pp. 225–251, 1992.

[6] L. R. Goldberg, "An alternative "description of personality": The big-five factor structure," in *Personality and personality disorders*. Routledge, 2013, pp. 34–47.

[7] O. P. John, S. Srivastava *et al.*, "The big-five trait taxonomy: History, measurement, and theoretical perspectives," 1999.

[8] N. Li, H. Zhang, and L. Feng, "Incorporating forthcoming events and personality traits in social media based stress prediction," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 603–621, 2023.

[9] R. R. McCrae, P. T. Costa Jr., F. Ostendorf, A. Angleitner, M. Hřebíčková, M. D. Avia, J. Sanz, M. L. Sánchez-Bernardos, M. E. Kusdil, R. Woodfield, P. R. Saunders, and P. B. Smith, "Nature over nurture: Temperament, personality, and life span development." *Journal of Personality and Social Psychology*, vol. 78, no. 1, pp. 173–186, 2000. [Online]. Available: https://doi.org/10.1037/0022-3514.78.1.173

[10] R. D. Goodwin and I. H. Gotlib, "Gender differences in depression: the role of personality factors," *Psychiatry Research*, vol. 126, no. 2, pp. 135–142, 2004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165178104000393

[11] K. Gönç and H. Dibeklioğlu, "Affect and personality aided modeling of transcribed speech for depression severity estimation," *IEEE Transactions on Affective Computing*, vol. 16, no. 3, pp. 2334–2351, 2025.

[12] M. A. Bucher, T. Suzuki, and D. B. Samuel, "A meta-analytic review of personality traits and their associations with mental health treatment outcomes," *Clinical psychology review*, vol. 70, pp. 51–63, 2019.

[13] R. F. Krueger and N. R. Eaton, "Personality traits and the classification of mental disorders: Toward a more complete integration in dsm–5 and an empirical model of psychopathology." *Personality Disorders: Theory, Research, and Treatment*, vol. 1, no. 2, p. 97, 2010.

[14] S. M. Lamers, G. J. Westerhof, V. Kovács, and E. T. Bohlmeijer, "Differential relationships in the association of the big five personality traits with positive mental health and psychopathology," *Journal of Research in Personality*, vol. 46, no. 5, pp. 517–524, 2012.

[15] W. Kang, F. Steffens, S. Pineda, K. Widuch, and A. Malvaso, "Personality traits and dimensions of mental health," *Scientific Reports*, vol. 13, no. 1, p. 7091, 2023.

[16] F. D. Mann, O. E. Atherton, C. G. DeYoung, R. F. Krueger, and R. W. Robins, "Big five personality traits and common mental disorders within a hierarchical taxonomy of psychopathology: A longitudinal study of mexican-origin youth," *J Abnorm Psychol*, vol. 129, no. 8, pp. 769–787, Sep. 2020.

[17] G. Park, H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, M. Kosinski, D. J. Stillwell, L. H. Ungar, and M. E. Seligman, "Automatic personality assessment through social media language." *Journal of personality and social psychology*, vol. 108, no. 6, p. 934, 2015.

[18] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman *et al.*, "Personality, gender, and age in the language of social media: The open-vocabulary approach," *PloS one*, vol. 8, no. 9, p. e73791, 2013.

[19] V. Kulkarni, M. L. Kern, D. Stillwell, M. Kosinski, S. Matz, L. Ungar, S. Skiena, and H. A. Schwartz, "Latent human traits in the language of social media: An open-vocabulary approach," *PloS one*, vol. 13, no. 11, p. e0201703, 2018.

[20] E. A. Ríssola, D. E. Losada, and F. Crestani, "A survey of computational methods for online mental state assessment on social media," *ACM Trans. Comput. Healthcare*, vol. 2, no. 2, Mar. 2021. [Online]. Available: https://doi.org/10.1145/3437259

[21] E. A. Ríssola, J. Parapar, D. E. Losada, and F. Crestani, *A Survey of the First Five Years of eRisk: Findings and Conclusions*. Cham: Springer International Publishing, 2022, pp. 31–57. [Online]. Available: https://doi.org/10.1007/978-3-031-04431-1_3

[22] M. Gjurković, V. M. Karan, I. Vukojević, M. Bošnjak, and J. Snajder, "PANDORA talks: Personality and demographics on Reddit," in *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, L.-W. Ku and C.-T. Li, Eds. Online: Association for Computational Linguistics, Jun. 2021, pp. 138–152. [Online]. Available: https://aclanthology.org/2021.socialnlp-1.12/

[23] I. B. Myers, *A guide to the development and use of the Myers-Briggs type indicator: Manual*. Consulting Psychologists Press, 1985.

[24] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 273–291, 2014.

[25] J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer, "Psychological aspects of natural language use: Our words, our selves," *Annual review of psychology*, vol. 54, no. 1, pp. 547–577, 2003.

[26] Q. Fang, A. Giachanou, A. Bagheri, L. Boeschoten, E. J. van Kesteren, M. S. Kamalabad, D. L. Oberski *et al.*, "On text-based personality computing: Challenges and future directions," *Findings of the Association for Computational Linguistics, ACL 2023*, pp. 10 861–10 879, 2023.

[27] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: Liwc 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001, 2001.

[28] M. Wilson, "Mrc psycholinguistic database: Machine-usable dictionary, version 2.00," *Behavior research methods, instruments, & computers*, vol. 20, no. 1, pp. 6–10, 1988.

[29] J. W. Pennebaker and L. A. King, "Linguistic styles: language use as an individual difference." *Journal of personality and social psychology*, vol. 77, no. 6, p. 1296, 1999.

[30] S. Štajner and S. Yenikent, "A survey of automatic personality detection from texts," in *Proceedings of the 28th international conference on computational linguistics*, 2020, pp. 6284–6295.

[31] F. M. Rangel Pardo, F. Celli, P. Rosso, M. Potthast, B. Stein, and W. Daelemans, "Overview of the 3rd author profiling task at pan 2015," in *CLEF 2015 evaluation labs and workshop working notes papers*, 2015, pp. 1–8.

[32] G. Attanasio, F. M. Plaza-del Arco, D. Nozza, and A. Lauscher, "A tale of pronouns: Interpretability informs gender bias mitigation for fairer instruction-tuned machine translation," *arXiv preprint arXiv:2310.12127*, 2023.

[33] W. R. Wright and D. N. Chin, "Personality profiling from text: Introducing part-of-speech n-grams," in *User Modeling, Adaptation, and Personalization: 22nd International Conference, UMAP 2014, Aalborg, Denmark, July 7-11, 2014. Proceedings 22*. Springer, 2014, pp. 243–253.

[34] B. Verhoeven, W. Daelemans, and B. Plank, "Twisty: a multilingual twitter stylometry corpus for gender and personality profiling," in *Proceedings of the Tenth international conference on language resources and evaluation (LREC'16)*, 2016, pp. 1632–1637.

[35] B. Verhoeven and W. Daelemans, "Clips stylometry investigation (csi) corpus: a dutch corpus for the detection of age, gender, personality, sentiment and deception in text," in *LREC 2014-NINTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION*, 2014, pp. 3081–3085.

[36] D. E. Losada, F. Crestani, and J. Parapar, "erisk 2017: Clef lab on early risk prediction on the internet: Experimental foundations," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, G. J. Jones, S. Lawless, J. Gonzalo, L. Kelly, L. Goeuriot, T. Mandl, L. Cappellato, and N. Ferro, Eds. Cham: Springer International Publishing, 2017, pp. 346–360.

[37] ——, "Overview of erisk: Early risk prediction on the internet," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France*, 2018, pp. 343–361.

[38] ——, "Overview of erisk 2019 early risk prediction on the internet," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Cham: Springer International Publishing, 2019, pp. 340–357.

[39] ——, "Overview of erisk 2020: Early risk prediction on the internet," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Cham: Springer International Publishing, 2020, pp. 272–287.

[40] J. Parapar, P. Martín-Rodilla, D. E. Losada, and F. Crestani, "Overview of erisk 2021: Early risk prediction on the internet," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Cham: Springer International Publishing, 2021, pp. 324–344.

[41] ——, "Overview of erisk 2022: Early risk prediction on the internet," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Cham: Springer International Publishing, 2022, pp. 233–256.

[42] ——, "Overview of erisk 2023: Early risk prediction on the internet," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Cham: Springer Nature Switzerland, 2023, pp. 294–315.

[43] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[44] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, "Fine-tuning a masked language model," 2022. [Online]. Available: https://huggingface.co/course/chapter7/3?fw=pt

[45] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, 2019, pp. 8024–8035.

[46] M. Aragon, A. P. Lopez Monroy, L. Gonzalez, D. E. Losada, and M. Montes, "DisorBERT: A double domain adaptation model for detecting signs of mental disorders in social media," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 15 305–15 318. [Online]. Available: https://aclanthology.org/2023.acl-long.853

[47] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "Mpnet: Masked and permuted pre-training for language understanding," 2020. [Online]. Available: https://arxiv.org/abs/2004.09297

[48] A. Ibáñez, C. Blanco, P. Moreryra, and J. Sáiz-Ruiz, "Gender differences in pathological gambling," *J Clin Psychiatry*, vol. 64, no. 3, pp. 295–301, Mar. 2003.

[49] F. Moloney, J. Amini, M. Sinyor, A. Schaffer, K. L. Lanctôt, and R. H. B. Mitchell, "Sex differences in the global prevalence of nonsuicidal self-injury in adolescents: A meta-analysis," *JAMA Netw. Open*, vol. 7, no. 6, p. e2415436, Jun. 2024.

[50] M. Solmi, J. Radua, M. Olivola, E. Croce, L. Soardo, G. Salazar de Pablo, J. Il Shin, J. B. Kirkbride, P. Jones, J. H. Kim, J. Y. Kim, A. F. Carvalho, M. V. Seeman, C. U. Correll, and P. Fusar-Poli, "Age at onset of mental disorders worldwide: large-scale meta-analysis of 192 epidemiological studies," *Molecular Psychiatry*, vol. 27, no. 1, pp. 281–295, Jan. 2022.

[51] S. Ronzitti, V. Lutri, N. Smith, M. Clerici, and H. Bowden-Jones, "Gender differences in treatment-seeking british pathological gamblers," *Journal of Behavioral Addictions*, vol. 5, no. 2, pp. 231 – 238, 2016. [Online]. Available: https://akjournals.com/view/journals/2006/5/2/article-p231.xml

[52] H. S. Friedman, "The multiple linkages of personality and disease," *Brain Behav Immun*, vol. 22, no. 5, pp. 668–675, Oct. 2007.

[53] H. S. Friedman, M. L. Kern, S. E. Hampson, and A. L. Duckworth, "A new life-span approach to conscientiousness and health: combining the pieces of the causal puzzle," *Dev Psychol*, vol. 50, no. 5, pp. 1377–1389, Oct. 2012.

[54] P. S. Schaefer, C. C. Williams, A. S. Goodie, and W. Campbell, "Overconfidence and the big five," *Journal of Research in Personality*, vol. 38, no. 5, pp. 473–480, 2004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0092656603001089

[55] M. Wolfenstein and T. J. Trull, "Depression and openness to experience," *J Pers Assess*, vol. 69, no. 3, pp. 614–632, Dec. 1997.

[56] S. Assari and M. Dejman, "Gender, depressive symptoms, chronic medical conditions, and time to first psychiatric diagnosis among american older adults," *Int J Prev Med*, vol. 10, p. 182, Oct. 2019.

IX. Biography Section

Mario Ezra Aragón is an Associate Researcher at Centro Singular de Investigación en Tecnoloxias Intelixentes, a research center associated with the University of Santiago de Compostela (USC, Spain). He obtained his PhD in Computer Science from the National Institute for Astrophysics, Optics and Electronics (INAOE, Mexico) in 2022. His current research interests include, but are not limited to, natural language processing, pattern recognition, mental health analysis, and risk predictions on the internet.

Marcos Fernandez-Pichel is an Assistant Professor at the University of Santiago de Compostela and a research collaborator at Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS). He obtained his PhD in Computer Science (with honours) in 2023. His research focuses on assessing the credibility of online health information, IR, and applying NLP to mental health. In 2022, the Royal Galician Academy of Sciences awarded him the Best Young Researcher Paper Award.

David E. Losada is a Full Professor of Computer Science and Artificial Intelligence at the University of Santiago de Compostela (Spain). He received his BS in Computer Science (with honors) in 1997 and his PhD (with honors) in 2001 from the University of A Coruña. He was a lecturer at San Pablo-CEU University (2001–2002) and joined the University of Santiago de Compostela in 2003 as a Ramón y Cajal Senior Research Fellow. His research interests cover Information Retrieval and related areas, including IR evaluation, early risk detection, misinformation detection, IR models, summarization, novelty detection, sentence retrieval, and opinion mining. He is an active member of the IR community, regularly serving on the Program Committees of leading international conferences such as SIGIR and ECIR, has led several R&D projects and contracts in search technologies, and was recognized as an ACM Senior Member in 2011.