IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 27 October 2023, accepted 9 November 2023, date of publication 13 November 2023,
date of current version 20 November 2023.

*Digital Object Identifier 10.1109/ACCESS.2023.3332289*

RESEARCH ARTICLE

# Extracting Mental Health Indicators From English and Spanish Social Media: A Machine Learning Approach

**MIRYAM ELIZABETH VILLA-PÉREZ[1], LUIS A. TREJO[1], MAISHA BINTE MOIN[2], AND ELENI STROULIA[2], (Member, IEEE)**
[1]School of Engineering and Sciences, Tecnologico de Monterrey, Atizapán de Zaragoza 52926, Mexico
[2]Department of Computing Science, University of Alberta, Edmonton, AB T6G2R3, Canada

Corresponding author: Luis A. Trejo (ltrejo@tec.mx)

The work of Miryam Elizabeth Villa-Pérez was supported in part by Consejo Nacional de Humanidades, Ciencias y Tecnología (CONAHCYT) of Mexico under the Scholarship CVU under Grant 637273.

**ABSTRACT** This study examines the communications of English- and Spanish-speaking Twitter users through traditional and deep learning algorithms to automatically recognize whether they live with one of nine mental health conditions. We created two datasets in English and Spanish. The ''diagnosed'' set comprises the timeline of 1,500 users who explicitly reported in one or more of their posts having been diagnosed with one of the following: ADHD, Anxiety, Autism, Bipolar, Depression, Eating disorders, OCD, PTSD, and Schizophrenia. The ''control'' set comprises the timeline of 1,700 randomly selected users who had not disclosed a diagnosis. We extracted a variety of text features from the collected data, such as n-grams, q-grams, Part-of-speech (POS) tags, topic modeling, Linguistic Inquiry and Word Count (LIWC), and word embeddings, and trained traditional machine-learning and deep learning classifiers for two tasks: binary classification, to distinguish between diagnosed and non-diagnosed users, and multiclass classification, to identify the specific diagnosis. Overall, XGBoost and convolutional neural network (CNN) performed the best in the two classification tasks. Moreover, lexical attributes based on n-grams and q-grams are the ones that performed well in both datasets. Using our collected datasets, for binary classification, we achieved an AUC of 0.835 on the Spanish Twitter dataset using n-grams of words from one to three (UBT) and 0.846 on the English Twitter dataset with a 5-gram characters (C5) model. In multiclass classification, we obtained an AUC of 0.712 and 0.697 in the Spanish and English Twitter datasets, respectively.

**INDEX TERMS** Binary classification, machine learning, mental health disorders, multiclass classification, social media, Twitter.

## I. INTRODUCTION

In 2020, the World Health Organization (WHO) estimates that mental disorders affect close to 1 billion people [1], or slightly more than one in ten of the world's population. In terms of disability, morbidity and mortality, and overall quality of life, people who suffer from mental disorders experience a significant negative influence on their lives [2]. Given the stigma still attached to mental health challenges, people tend not to seek help early; therefore, automated methods that might signal an individual the need to do so could potentially significantly impact their long-term health.

With the rise of social media, research has turned to these platforms to gain insights into users' well-being and to explore how people's behaviors on these platforms relate to their mental health status. Topics such as depression [3], [4], [5], self-harm [6], suicide risk [7], among others, have recently been investigated in the literature by integrating data from social media alongside natural language processing (NLP) and several machine learning techniques.

For example, the workshop on Computational Linguistics and Clinical Psychology (CLPsych) was established as

The associate editor coordinating the review of this manuscript and approving it for publication was Larbi Boubchir.

an interdisciplinary forum bringing together researchers and clinicians interested in improving mental health [8]. Similarly, the shared task on early risk prediction (eRisk) on the Internet, organized as part of the Conference and Evaluation Forum Laboratories (CLEF), explores processes related to the timely detection of health and safety problems, for example, signs associated with psychological disorders such as depression and anorexia [9], [10], [11].

Most of the studies reported in the literature concentrate on English-language data [12], although Spanish-speaking countries such as Mexico, Spain, and Colombia are among the top 20 most active Twitter users in the world, with more than 13 million, 8 million, and 4 million users, respectively [13]. The United States has 76.9 million Twitter users, making it the country with the most users on the social media [13]. Therefore, including English and Spanish-speaking users could provide coverage for a large population segment. Furthermore, most past studies focus on each condition in isolation (i.e., they are approached as a binary classification, such as depression vs. non-depression) [14]. In this paper, we address the problem of collecting and categorizing large volumes of Spanish and English language data to automatically detect the nine mental health disorders described in Cohan et al. [15]. Our long-term objective is to develop a robust and general methodology for extracting indicators of a broad range of mental health conditions from the communications of social media users in a variety of languages in order to support earlier diagnosis and timely treatment.

To that end, we constructed two datasets in English and Spanish, following the same data-collection methodology. Next, we extracted the same linguistic features from English and Spanish user profiles: *n*-grams, *q*-grams, Part-of-speech (POS) tags, topic modeling, Linguistic Inquiry and Word Count (LIWC), and word embeddings. We then trained several traditional machine learning and deep learning models for binary and multiclass classification.

We use *n*-grams of words and *q*-grams of characters as features to capture a broad range of textual information. While *n*-grams refer to a sequence of "N" words that appear together in a text, *q*-grams are a collection of successive characters in a document that may include letters, numbers, and other symbols. The goal of topic modeling is to discover the main themes in a large collection of texts by grouping words that often appear in the same context and finding the underlying patterns in these groups of words. Word embeddings generate a unique vector of numbers for each word in a corpus based on the context in which they appear so that words with similar contexts have similar vector representations. As a result, patterns in the connections between words can be discovered. LIWC uses a word dictionary to categorize texts based on psychological or linguistic dimensions, such as positive or negative emotion, cognitive processing, or social relationships. The aim is to extract insights to understand human behavior and psychology better.

Our work makes three contributions to the state of the art. We have curated two new publicly available datasets (in English and Spanish) collected with the same methodology. We systematically compared traditional and deep-learning algorithms for learning binary and multiclass classifiers to distinguish between diagnosed and non-diagnosed user profiles. Finally, we have demonstrated that our proposed models can accurately detect users who may suffer from mental disorders in both datasets; for example, the model based on *n*-grams of one to three words (UBT) achieves the highest AUC and F1-score on the diagnosed class (AUC 0.835; F1 0.824) for the Spanish Twitter dataset, while the 5-gram characters (C5) perform the best (AUC 0.846; F1 0.819) on the English Twitter dataset.

The rest of this paper is organized as follows. Section II reviews work related to our own. Next, Section III describes the methodology for Twitter data collection and the development of the English and Spanish datasets, the extracted text attributes and the classification methods used in our experiments. In Section IV we present the results of our experiments for the two classification tasks (binary and multiclass) on the extracted datasets, and on the Self-reported Mental Health Diagnoses (SMHD) dataset. Subsequently, in Section V we discuss the results obtained. In Section VI we address the current limitations of the work. Finally, in Section VII, we outline the conclusion and future lines of research.

## II. RELATED WORK

Online social media platforms, such as Facebook, Reddit, and Twitter, are pathways for people to stay connected and share information. Over the past decade, several studies have utilized the rich data on social media platforms to explore the complex relationship between mental health and language usage. De Choudhury et al. [3] first explored the potential of using social media data to detect major depressive disorder in individuals. Studies in [15], [16], [17], and [18] computationally showed that distinctive linguistic patterns grounded by psycholinguistic theories can be crucial in identifying users suffering from mental disorders. As words that people use are often indicative of their psychological states, previous work [19] has compiled a test collection of corpora on depression and language use to encourage research on the differences in language usage between depressed and non-depressed users and also to explore the evolution of language used by depressed individuals. Aside from textual data, another significant component for proper classification is the study of behavior. These behavioral attributes include engagement, emotions, and ego networks [3].

Twitter and Reddit are the most studied platforms for mental health research [20]. In the recent past, shared tasks like CLPsych 2015 [8] using Twitter data and CLEF eRisk (2017-20) [9], [10], [11], [21] using Reddit data introduced datasets for mental health analysis, for conditions like depression, anorexia, post-traumatic stress disorders (PTSD) and self-harm.

The CLPsych 2015 shared task [8] used data from Twitter users who state a diagnosis of depression or PTSD and demographically-matched community controls to build a group that is not entirely unrelated to the diagnosed group. The train partition consisted of 327 depression users, 246 PTSD users, and 572 age and gender-matched control users, for a total of 1,145 users. The test data follows the same methodology, resulting in 600 users distributed as follows: 150 depression users, 150 PTSD users, and 300 control-matched users. For each user in the dataset, approximately 3,200 most recent posts were collected.

The eRisk forum each year organizes a new task around a specific disorder; for instance, in 2017 and 2018, the shared task focused on depression [9], [10], and in 2019 and 2020, there were tasks on predicting anorexia and self-harm tendencies [11], [21]. For each task, a dataset is created using Reddit posts selected from the authors' self-report of having a mental health disorder. As a result, the positive class comprises users who explicitly mentioned in at least one of their posts that they had been diagnosed with depression or anorexia. The negative class contains users from other Reddit groups and users active in the depression or anorexia groups without a self-declaration of having a mental disorder, ensuring that the gap between healthy and diagnosed users is not trivially detectable [16]. For both classes, a long history of publications is collected from the users included in the dataset.

In order to produce the dataset SMHD, Cohan et al. [15] searched for high-precision patterns to identify self-reported diagnoses of nine different mental health conditions. Some studies have used these pre-defined datasets for their research, while others focused on collecting their own data. In general, some form of regular expressions (regex) matching is used to construct these datasets, and common annotation mechanisms include community participation, clinical surveys and platform activity [20].

Jamil et al. [4] have trained classifiers on CLPsych 2015 and their newly introduced BellLetsTalk dataset, constructed using tweets collected from the #BellLetsTalk campaign, designed to support mental health across Canada. The results show that due to the increase in granularity and data imbalance, achieving satisfactory performance on tweet-level analysis that monitors individual tweets for signs of mental disorders (specifically depression) is more challenging in comparison to a user-level analysis, which involves looking at the tweets history of a user over a period of time. Aguilera et al. [22] used the eRisk datasets for their experimentation and proposed a one-class classification, which considers only one mental disorder for training and testing. They obtained competitive results against traditional binary classification methods.

Though most early studies focused on depression detection, Guntuku et al. [23] analyzed the language of social media users diagnosed with attention-deficit/hyperactivity disorder (ADHD) to see how their language is correlated with users' characteristics, such as personality and temporal orientation. Mitchell et al. [24] explored potential linguistic markers such as topic distribution to increase classification accuracy for schizophrenia, which has traditionally been challenging to identify given the low prevalence of the condition.

Most existing mental health disorder classification methods use features engineered via natural language processing techniques (NLP) and various machine learning (ML) classifiers for evaluating text data from social media platforms. Coppersmith et al. [18], [25] examined a range of mental health disorders, and in their studies, *n*-grams provided superior performance compared to other analytic methods. In several studies [5], [6], [16], [26], broader textual features, such as LIWC and Latent Dirichlet Allocation (LDA) based topic allocation of posts, have been found to be useful. An ensemble of approaches combining linguistic information with user metadata achieved the best results in some studies [17], [27]. Tadesse et al. [5] showed the power of proper feature selection, especially multiple feature combinations, for depression detection. Their experimental data was collected from Reddit, where bigrams showed the best performance as a single feature. The top performance was the combination of LIWC, LDA and bigram features with a multilayer perceptron (MLP) classifier. Additionally, Skaik and Inkpen [14] have summarized the features used in various mental health studies in their review.

On the SMHD dataset, Cohan et al. [15] applied term frequency-inverse document frequency (TF-IDF) vectorization of documents and performed classification using Logistic Regression, XGBoost, Support Vector Machines (SVM) and Convolutional Neural Networks (CNN) with fastText embeddings. They found that when working with several mental health conditions, the performance of the models is strongly affected by the number of diagnosed users in the training set. The use of fastText word embeddings obtained the best performance in their study in terms of F1-score. Regarding deep learning approaches, CNNs have shown the best performance in several studies [6], [28], [29]. Husseini et al. [28] have investigated the most effective deep neural network architecture for CLPsych 2015 and BellLetsTalk datasets. In most recent studies [16], [30], transformer-based architectures like bidirectional encoder representations from transformers (BERT) and robustly optimized BERT pre-training approach (RoBERTa) are also being explored for the task.

Most studies in social media mental health research only looked at posts written in English, though not as extensively; some studies have also explored data written in other languages. Tsugawa et al. [31] looked at the activity history of users to recognize the presence of depression; for that, they used ML classifiers on tweets written in Japanese and found topic models to be useful as features. Almouzini et al. [32] explored depression-related behaviors of Arabic words; in their experiments, the Liblinear classifier outperformed the

other ML models. Leis et al. [33] aimed to identify users' linguistic features and behavioral patterns for tweets posted in Spanish. Their visualization-based approach considered distribution over time, parts of speech, word count, use of negation, emotion, and polarity analysis. Uddin et al. [34] applied the Long Short Term Memory (LSTM) deep recurrent model and showed the effect of hyperparameter tuning for analyzing Bangla tweets for depression.

Table 1 briefly overviews some of the research papers discussed in the related work section.

## III. METHODOLOGY

### A. DATA COLLECTION

Twitter has more than 550 million registered users [35], with the population between 25 and 34 years of age being the most active (38.5%) [36]. Our dataset comprises tweets of users who reported in one of their tweets (in English or Spanish) a diagnosis of a mental health condition (diagnosed users) and users who have not reported any such condition (control users). Users and posts were extracted from Twitter through its application programming interface (API) [37]. All data obtained is public and collected between September 1st, 2020, and August 31st, 2021. The selection of users and their tweets was computed using the full-archive endpoint provided by the Twitter API, which allows developers to request historical tweets by searching them via a set of rules. An overview of the data-collection pipeline is shown in Fig. 1.

#### 1) DIAGNOSED GROUP

Following the work of Cohan et al. [15], the selected mental disorders correspond to the subdivisions of the Diagnostic and Statistical Manual of Mental Disorders of the American Psychiatric Association (DSM-5) [38]. They are the six first-level disorders, i.e., Anxiety disorders (Anx), Bipolar disorders (Bipo), Depressive disorders (Depr), Eating disorders (Eat), Obsessive-compulsive disorders (OCD), and Schizophrenia spectrum disorders (Schizophrenia); Post-traumatic stress (PTSD), classified under trauma-and stress-related disorders; lastly Autism spectrum disorder (Auts) and Attention-deficit hyperactivity disorder (ADHD), classified under neurodevelopmental disorders.

We leverage self-reported diagnostic statements in which a user posts on Twitter about being diagnosed with a mental health disorder such as depression, anxiety, bipolar, among others. It is essential to note that medical experts have not validated or confirmed these self-reports. While we have taken steps to filter and clean the data, the possibility of misreporting, misunderstanding, or inaccuracies in self-diagnoses exists. However, previous research indicates that this type of data collection has high inter-rater reliability [6], [15], [19], [23], [24], [25].

For this reason, we selected users who have publicly stated that they have been diagnosed with a mental condition. Tweets were obtained using regular expressions, e.g., "Me diagnosticaron", "He sido diagnosticado(a)", "I was diagnosed", and "I've been diagnosed". Tweets with matching diagnoses were reviewed to find out whether the tweet included a genuine statement of a mental health diagnosis (text patterns used are included in Section I in the Supplementary Material).

We searched for the nine conditions explained above. Following the methodology of Losada et al. [19], the regular expressions were strict. Expressions like "Tengo…", "Creo que tengo…", or "Nunca he sido diagnosticado(a), pero…", "I have…", "I think I have…", "I've never been diagnosed, but…" did not qualify as explicit expressions of a diagnosis. We only included a tweet in the diagnosed group when there was a clear and explicit mention of a diagnosis and the name of the mental health condition.

Table 2 shows examples of false (often jokes or non-clinical diagnoses) and authentic diagnostic posts. As shown in Table 2, the collected tweets do not usually include a definitive date, which might range from days to months or even years. There are also cases in which a description of the stage of recovery follows the statement. Using this method makes it possible to retrieve posts with some degree of uncertainty about the specific date of diagnosis [19]; however, these data are still valuable for the development of automatic methods for detecting mental disorders.

#### 2) CONTROL GROUP

In addition to the diagnosed users, we collected a random sample of users from the general population, constituting the control group or non-diagnosed class. These users were chosen from a pool of candidate users based on their similarity to diagnosed users as measured by the number of messages they posted [15]. This was done to avoid biases between the control and diagnosed groups in the dataset.

We follow a similar process for both languages: 1,500 users' usernames were randomly selected from a list of Twitter users who tweeted in a two-week window in early 2021 and did not have a self-reported mental disorder diagnosis. We also included in the control group 500 usernames whose tweets were followed by different hashtags related to mental health (e.g., #SaludMental / #MentalHealth, #NoEstasSolo / #YouAreNotAlone, or #ApoyoEmocional / #EmotionalSupport).

Hashtags are keywords people use to mark the topic of their social network content. Since on Twitter there are no forums where people frequently share their experiences about a common topic (as in other social media such as Reddit), it is possible to filter searches around a specific topic through hashtags. Therefore, as these people are talking about mental health topics, including them in the control group helps to make the collection more realistic.

Given this pool of 2,000 users, we identified the most similar control user (in terms of number of posts) for each user in the mental health group. In more detail, we matched each diagnosed user with a candidate control user based on

**TABLE 1. Overview of selected studies of mental disorder detection in social media. TF-IDF: Term frequency - Inverse document frequency; NRC: National Research Council Canada; HAN: Hierarchical Attention Network; ocSVM: one-class Support Vector Machine; ocKNN: one-class k-Nearest Neighbor; kNN: k-Nearest Neighbor; P: Precision; R: Recall; Acc: Accuracy.**

| Year | Reference | Data Source | Features | Algorithms | Best Results |
|---|---|---|---|---|---|
| 2017 | Jamil et al. [4] | Twitter (CLPsych & BellLetsTalk) | Bow, polarity word counts, depression word count, pronoun counts, LIWC, NRC sentiment lexicon, emoticon frequency, and readability | Linear SVM | F1: 0.21<br>P: 0.12<br>R: 0.80<br>Acc: 0.61<br>(Depression Tweet-level)<br><br>F1: 0.66<br>P: 0.58<br>R: 0.77<br>Acc: 0.61<br>(Depression User-level) |
| 2018 | Cohan et al. [15] | Reddit (SMHD) | TF-IDF vectorization | Logistic Regression, XGBoost, SVM, Shallow neural network, and CNN with fastText embeddings | F1: 0.27<br>P: 0.23<br>R: 0.48<br>(Multi-label) |
| 2018 | Trotzek et al. [17] | Reddit (eRisk) | 27-dimensional vector of metadata features | Logistic Regression, CNN with word-embeddings (GloVe, fastText, and own corpora) | F1: 0.73<br>P: 0.77<br>R: 0.85<br>(Depression) |
| 2018 | Husseini et al. [28] | Twitter (CLPsych & BellLetsTalk) | Word embedding pre-trained on the CLPsych 2015 Shared task data | CNN, BiLSTM with attention | AUC: 0.95<br>F1: 0. 87<br>P: 0.87<br>R: 0.87<br>Acc: 0.88<br>(CLPsych 2015)<br><br>AUC: 0.92<br>F1: 0.82<br>P: 0.82<br>R: 0.84<br>Acc: 0.83<br>(BellLetsTalk) |
| 2019 | Tadesse et al. [5] | Reddit | LIWC, LDA, *n*-grams (TF-IDF) | Logistic Regression, SVM, Random Forest, Adaptive Boosting, and Multilayer Perceptron | F1: 0.93<br>P: 0.90<br>R: 0.92<br>Acc: 0.91<br>(Depression) |
| 2020 | Aguilera et al. [22] | Reddit (eRisk) | TF-IDF and word-embeddings (GloVe, fastText, Word2vec) | Baseline: ocSVM, ocKNN, SVM, kNN. Proposal: gSC and OCC-kSS | F1: 0.63<br>(Depression)<br><br>F1: 0.81<br>(Anorexia) |
| 2021 | Uban et al. [16] | Reddit (eRisk) | GloVe embeddings, content, style features, LIWC, emotions, and sentiment | Logistic Regression, BiLSTM with attention, HAN, CNN, and Transformers (RoBERTa and AlBERT) | AUC: 0.87<br>F1: 0.65<br>(Self-harm)<br><br>AUC: 0.96<br>F1: 0.61<br>(Anorexia)<br><br>AUC: 0.83<br>F1: 0.45<br>(Depression) |
| 2021 | Malviya et al. [30] | Reddit | Word2vec, TF-IDF | Linear and Boosting classifiers, and Transformers (RoBERTa, BERT, Electra) | F1: 0.98<br>Acc: 0.98<br>(Depression) |

the similarity of the number of texts they had. We selected controls without replacement, so a control user could only be included once.

We are aware that there is a possibility that the control group includes users with some of the conditions investigated, for example, users who do not disclose depression, ADHD,

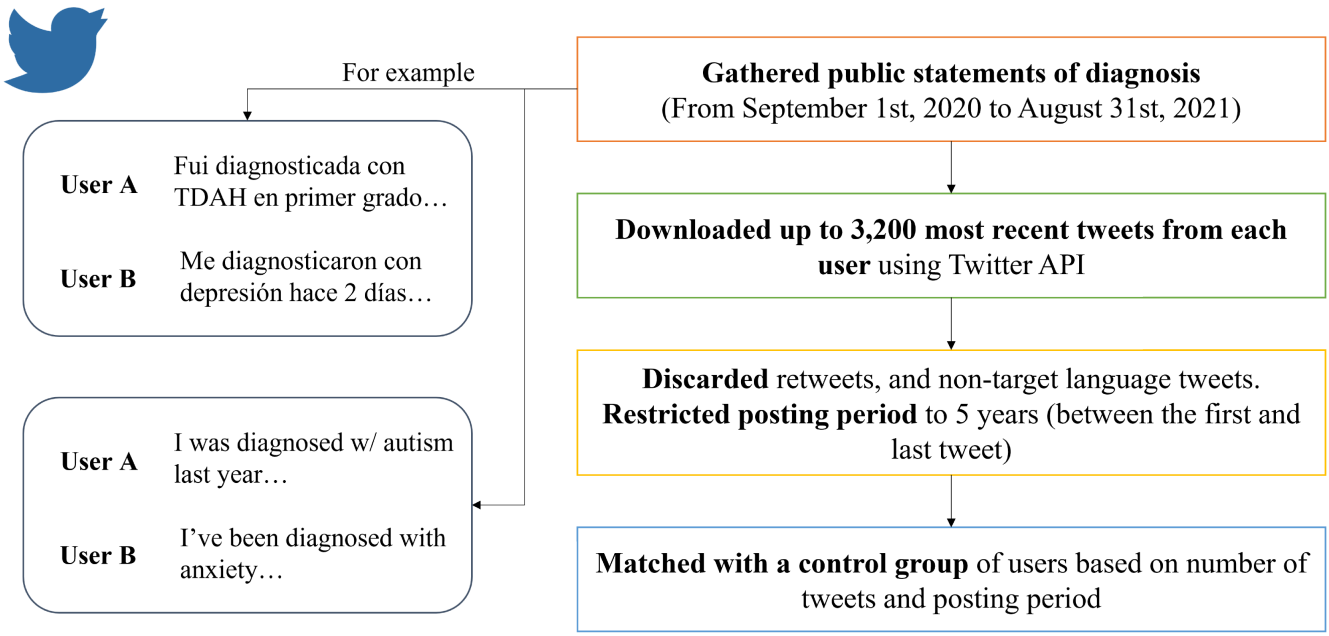


**FIGURE 1.** Overview of the data collection pipeline.

**TABLE 2.** Examples of tweets with genuine (left) and false (right) statements of self-reported diagnoses.

| Language | Positive tweets (genuine) | Negative tweets false) |
|---|---|---|
| Spanish | Me diagnosticaron recientemente depresión, hasta el día de hoy no he logrado salir adelante. Es difícil vivir diariamente con esta enfermedad. | No me han diagnosticado nada, pero según TikTok tengo depresión, ansiedad y diversos trastornos mentales :( |
| | A casi un mes que fui diagnosticada con trastorno de ansiedad, a veces ni las pastillas ayudan... | @usuario tengo síntomas de esquizofrenia, pero no me han diagnosticado algo concreto. |
| English | I was diagnosed with bulimia this year. In March, I started to get better thanks to my family <3, who has taken care of me a lot. | I've been diagnosed with OCDL: Obsessive Compulsive Disorder about Loki (referring to the Thor movie character). It is serious, and there is no cure D: |
| | Years ago, I was diagnosed with PTSD. I was on medication, and it is something I don't want to go through it again... | I feel sad even though I have never been clinically diagnosed with depression. |

or other mental health problems in social media; however, the impact is expected to be minimal since alternative approaches (e.g., screening tests, crowd-sourced surveys, etc.) are not noise-free either [19], [25].

### 3) TEXTS EXTRACTED

For both groups (diagnosed and control) in each dataset (English and Spanish), users were filtered based on verification of the tweets -removing ads and spam. Then, we retrieved the most recent tweets (up to 3,200, which is the maximum allowed by Twitter API for each user). Retweets and tweets that were not in the target language were excluded.

Next, to ensure that we have sufficient data for the analysis, users were filtered to eliminate those with fewer than 25 tweets or a total number of words less than 500 and those whose tweets are more than five years apart between the first and last tweet.

### 4) DATA STATISTICS

The statistics of the resulting collection are reported in Table 3. Following our strategy, we collected a substantial number of subjects and a large number of posts. In both datasets, the total number of users is slightly more than 3,200, with a distribution of approximately 1,500 users in the diagnosed group and 1,700 in the control group.

The total number of tweets per user depends on the language. For instance, an average of 700 tweets were collected per user in the Spanish Twitter dataset, while in the English Twitter dataset, approximately 1,500 tweets were collected. Also, the difference in the average period between the first and last tweet is notable. In the first case, the tweets cover between three and six months, and in the second case, between seven months and a year or more.

Regarding the distribution of the number of users by disorder, in the Spanish-language dataset, the mental disorder with the most users is Depr, followed by Anx, and the class

**TABLE 3.** Overview of the collected datasets.

| Class | Users | | Total tweets | | Tweets per user, mean (SD | | Time interval (days)[a], mean (SD | |
|---|---|---|---|---|---|---|---|---|
| | Spanish | English | Spanish | English | Spanish | English | Spanish | English |
| ADHD | 206 | 622 | 145,498 | 1,009,002 | 706.3 (247.4) | 1,622.2 (979.1) | 323.0 (376.3) | 539.3 (633.8) |
| Anx | 351 | 124 | 256,657 | 195,944 | 731.2 (226.7) | 1,580.2 (987.9) | 310.2 (332.4) | 360.9 (423.4) |
| Auts | 119 | 170 | 84,488 | 273,435 | 710.0 (245.4) | 1,608.4 (983.6) | 295.4 (310.2) | 561.4 (757.9) |
| Bipo | 72 | 136 | 54,562 | 200,425 | 757.8 (227.9) | 1,473.7 (986.3) | 345.4 (328.3) | 523.8 (585.4) |
| Depr | 478 | 249 | 339,927 | 336,997 | 711.1 (237.1) | 1,353.4 (971.3) | 314.4 (349.8) | 509.0 (643.3) |
| Eat | 141 | 26 | 103,778 | 37,893 | 736.0 (257.6) | 1,457.4 (904.6) | 263.1 (276.7) | 509.0 (485.3) |
| OCD | 71 | 65 | 53,076 | 119,545 | 747.6 (201.1) | 1,839.2 (929.4) | 356.6 (359.8) | 508.2 (590.3) |
| PTSD | 65 | 127 | 48,218 | 195,519 | 741.8 (240.8) | 1,539.5 (1,014.2) | 327.5 (367.3) | 445.8 (627.3) |
| Schizophrenia | 48 | 24 | 27,516 | 40,178 | 573.3 (299.4) | 1,674.1 (1,143.8) | 174.4 (204.4) | 273.0 (236.2) |
| Control | 1,704 | 1,703 | 1,484,651 | 2,761,555 | 871.3 (528.9) | 1,621.6 (1,013.8) | 325.1 (351.9) | 698.1 (818.8) |

a Between the first and last tweet.

with the fewest users is Schizophrenia and PTSD. In the English-language dataset, ADHD and Depr are the classes with most users, and Schizophrenia and Eat are the classes with the fewest users.

The imbalance ratio (IR) indicates the fraction between the number of objects in the positive and negative classes (also called minority and majority class, respectively) [39]. Consequently, a higher IR value indicates more imbalance between the classes, which means that the number of objects in the positive class is low compared to the negative class.

Table 4 shows the IR for each mental health condition with respect to the control class (which is the majority class). Based on the IR, in the Spanish-language dataset, the Depr class is about one-third the size of the control class, while the Schizophrenia class is thirty-five times smaller than the control class. In the English-language dataset, the ADHD class has the lowest imbalance, being slightly more than half smaller than the control class. Once again, the Schizophrenia class is the most imbalanced because the control class is seventy-one times larger than the control class.

**TABLE 4.** IR of each class with respect to the control class.

| Class | IR value | |
|---|---|---|
| | Spanish | English |
| ADHD | 8.3 | 2.7 |
| Anx | 4.9 | 13.7 |
| Auts | 14.3 | 10.0 |
| Bipo | 23.7 | 12.5 |
| Depr | 3.6 | 6.8 |
| Eat | 12.1 | 65.5 |
| OCD | 24.0 | 26.2 |
| PTSD | 26.2 | 13.4 |
| Schizophrenia | 35.5 | 71.0 |

### B. FEATURE EXTRACTION

The first step towards classifying users was to preprocess the corpus. This step removes unnecessary punctuation marks and white spaces for each post. Then, we use the pysentiment toolkit [40] to preprocess tweets from each user, removing URLs and hashtags symbols, replacing mentions (@user), and eliminating continuously repeated symbols or characters (such as "aaaa") and stop words defined in the NLTK (Natural Language Toolkit) library [41]. Finally, we use tokenization to divide the posts into individual tokens and apply lemmatization to reduce the words to their root form using the spaCy library [42].

We consider all the tweets posted by a user as their "profile" document, and we extract the following features from each document.

#### 1) LEXICAL

Words are a powerful tool of expression, and they are one way of communicating with each other; they can be used to express ideas, concepts, and emotions. Therefore, we use word $n$-grams ($n = 1, 2, 3$) and char $q$-grams ($q = 3, 5, 7$) as features. We weigh each term $t$ with its TF-IDF, as follows:

$$TF - IDF(t, d, D) = TF(t, d) \cdot IDF(t, D), \quad \text{where:}$$

$$TF(t) = \frac{\text{Number of times term t appears in a document}}{\text{Total number of terms in a document d}}$$

$$IDF(t, D) = log_e\left(\frac{\text{Total number of documents in the corpus}}{\text{Number of documents with term t in them}}\right)$$

The TF-IDF value increases with the number of times a word appears in a document but decreases with the number of documents in which the word appears. This helps to filter out words that often appear across the document collection. The higher the TF-IDF score, the more important or relevant the term is.

In our study, we use TF-IDF vectorizer from the scikit-learn Python library [43], and we restrict the term-document matrix to 10,000 most frequent $n$-grams and $q$-grams. We also consider only the words that appear at least in more than 15 documents and ignore terms that appear in more than 75% of the documents. From our results, the model performs better when the term matrix is limited to these numbers.

#### 2) PART-OF-SPEECH (POS) N-GRAMS

We use the n-gram ($n = 1, 2, 3$) with their POS tags to understand the importance of the role played in the text. We used the POS Tagger provided by the spaCy library [42] to

identify lexical and grammatical properties of words such as pronouns, adjectives, verbs, adverbs, nouns, etc. We restricted the vocabulary size to 10,000, as we did for the lexical attributes because the model performed better with this vocabulary size.

#### 3) TOPIC MODELING

This category aims to capture the topic of each document by detecting patterns and automatically clustering groups of similar words and expressions that best represent a set of documents. Topic modeling is an unsupervised machine learning technique, so this clustering is performed without the use of any dictionary. We employ two of the best-known topic modeling algorithms, Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF). For creating the feature set, we compute the probability of the appearance of each topic in the document. Based on our results, both models perform best when limited to 50 topics. For both algorithms, we use the implementation of the scikit-learn library [43].

#### 4) LIWC

The LIWC dictionary is widely used in computational linguistics as a source of features for psychological and psycholinguistic analysis [44]. The Spanish LIWC2007 dictionary includes around 70-word categories to analyze different language dimensions like emotions (e.g., sadness, anger, etc.), self-references, and words for perceptual, cognitive, or biological processes in each text. The tool returns the estimated percentage of words in each category, so we transform the feature set by scaling each numerical attribute within the interval [0, 1]. We use the most recent version of the dictionary, LIWC2015, for English-language datasets, which contains about 98 categories.

#### 5) EMBEDDINGS

Recent advances in language models extracted from large corpora have led to the development of a new style of text analysis, where each word is located in a multidimensional vector space based on its adjacent words in the underlying corpus. Word embeddings can be trained using the input corpus itself or can be generated using pre-trained word embeddings such as Global Vectors for Word Representation (GloVe) [45], fastText [46], and Word2Vec [47]. These three models have advantages and disadvantages and depend on the specific use case. However, fastText is more suitable for handling unknown or new words and language with typos or abbreviations, such as Twitter data. Furthermore, GloVe can handle large corpora efficiently due to its co-occurrence matrix-based approach, which allows training to be scalable and fast. Hence, in this study fastText and GloVe were tested as word embeddings. In the case of the Spanish Twitter dataset, we use pre-trained word embeddings vectors from the Spanish Billion Word Corpus (SBWC) [48]. For the English Twitter dataset, we used word vectors trained on Common Crawl from the fastText library [49] and the Stanford project [50]. The description of all utilized word embeddings is shown in Table 5.

### C. CLASSIFICATION OF MENTAL HEALTH DISORDERS

We address two different classification problems. The first concerns binary classification to recognize users diagnosed with a mental health condition versus non-diagnosed individuals (or control users). The second classification problem deals with multiclass classification of the nine mental health conditions, i.e., recognizing the type of diagnosis of users. We applied the same methodology and evaluation approach for both tasks, including training eight different classifiers with a diverse set of linguistic features (explained above).

We experimented with several algorithms, including traditional machine learning classifiers and deep learning-based classifiers.

The first four classifiers (Naïve Bayes, SVM, Bagging with Decision Trees, and XGBoost) are employed in our analysis with the lexical, POS tag, topic modeling, and LIWC features explained in Section III-B. We used the algorithms implemented in the scikit-learn and XGBoost libraries of Python, and all the selected classifiers were executed using the parameters by default, which are shown in Table 6. This allows us to establish a baseline performance on the classification problems we addressed. In the case of SVM, we use three types of kernel functions: linear, radial basis function (RBF), and polynomial with degree 3 to explore different possibilities of data classification; since SVM is a powerful algorithm, but it can be sensitive to the shape of the data.

Based on neural network architecture, the other four classifiers include the same number of layers and are trained with a similar approach. Thus, a Convolutional Neural Network (CNN), a Gated Recurrent Unit (GRU) network - that is, an improved version of standard Recurrent Neural Networks (RNNs)-, a Long Short-Term Memory (LSTM) network, and a CNN-LSTM were chosen for the word embedding features. The CNN-LSTM architecture uses a CNN layer for feature extraction on the input data combined with LSTM to support sequence prediction. The parameters of the neural network architectures are detailed in Section II of the Supplementary Material. It is essential to highlight that all the architectures include a Dropout layer that penalizes the model weights to prevent them from growing too large. Similarly, each neural network was trained with early stopping to interrupt the training when the performance on the validation set is no longer improving; this prevents the model from being overfitted.

The source code for our experiments is available in a GitHub repository,[1] and the Twitter dataset for mental disorder detection is available for download through the IEEE DataPort platform.[2]

[1] https://github.com/miryamelizabeth/Twitter-Mental-Health
[2] https://ieee-dataport.org/documents/twitter-dataset-mental-disorders-detection

**TABLE 5.** Characteristics of the fastText and GloVe word embeddings used in this work.

| Model | Dimensions | Corpus tokens (in billion) | Word vectors (in million) | % Words in the dataset |
|---|---|---|---|---|
| **Spanish Twitter dataset** | | | | |
| fastText SBWC | 300 | 1.4 | 0.85 | 86.2% |
| GloVe SBWC | 300 | 1.4 | 0.85 | 86.1% |
| **English Twitter dataset** | | | | |
| fastText Crawl | 300 | 600 | 2.0 | 83.9% |
| GloVe Crawl | 300 | 840 | 2.2 | 81.3% |

**TABLE 6.** Parameter specification for the algorithms tested in our experimentation.

| Classifier | Parameters |
|---|---|
| Decision Tree (DT) | criterion: 'gini'; splitter: 'best'; max_depth: None (nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples); min_samples_split: 2; min_samples_leaf: 1; min_weight_fraction_leaf: 0.0; max_features: None (all features are considered); max_leaf_nodes: None; min_impurity_decrease: 0.0 |
| Naïve Bayes (NB) | The algorithm has no specific adjustable hyperparameters |
| Bagging with Decision Tree (BDT) | estimator: DecisionTreeClassifier; n_estimators: 10; max_samples: 1.0 (all samples are used to fit each base estimator); max_features: 1.0 (all features are used to fit each base estimator); bootstrap: True (samples are drawn with replacement) |
| SVM lineal | C: 1.0; kernel: 'linear'; gamma: 'scale'; coef0: 0.0; shrinking: True; class_weight: 'balanced'; probability: True; max_iter: -1 (no iteration limit) |
| SVM rbf | C: 1.0; kernel: 'rbf'; gamma: 'scale'; coef0: 0.0; shrinking: True; class_weight: 'balanced'; probability: True; max_iter: -1 (no iteration limit) |
| SVM poly | C: 1.0; kernel: 'poly'; degree: 3; gamma: 'scale'; coef0: 0.0; shrinking: True; class_weight: 'balanced'; probability: True; max_iter: -1 (no iteration limit) |
| XGBoost (XGB) | n_estimators: 100; max_depth: 3; learning_rate: 0.1; subsample: 1.0; gamma: 0; reg_alpha: 0; reg_lambda: 1; min_child_weight: 1; objective: 'binary:logistic'; booster: 'gbtree' |

## IV. RESULTS

In this section, we present our experiment results on the collected Twitter dataset for both classification tasks, binary and multiclass.

### A. EVALUATION METRICS

We randomly split the Twitter datasets using stratified 5-fold cross-validation (5FCV), so we preserved the distribution of the classes in all data partitions. By partitioning the data into K different folds to train and test the model, a more accurate indication of the model's generalizability to unobserved data can be obtained. To evaluate the different sets of features, we use the following performance indicators, as traditionally done in supervised classification:

- **Area under the ROC curve (AUC):** Area under the curve of true positive detection rate (TPR) versus false positive detection rate (FPR). AUC values above 0.5 are preferable since a value equal to 0.5 indicates that the classifier's performance is random guessing.
- **F1-score:** The harmonic mean of the precision and recall.
- **Precision:** It calculates how many positively identified samples are correct; it measures the proportion of users the classifier considers to have a mental health disorder that actually had the disorder. If precision=1, that means that all users diagnosed by the classifier really had the disorder.
- **Recall:** Also known as sensitivity (True Positive Rate). It estimates what proportion of positive samples was correctly identified. That measures how successfully the classifier recognizes users with the condition. If recall=1, that means that all users with the disorder are detected.

Because AUC gives an idea of the amount of work done by the classifier and is less sensitive to imbalance [51], it will be the main metric we consider when comparing performance between different models and datasets.

Based on the results, the Friedman non-parametric test [52], [53], [54] and the Finner post-hoc procedure [54] with $\alpha = 0.05$ were applied to determine whether the models' differences are statistically significant.

### B. BINARY CLASSIFICATION OF DIAGNOSED AND NON-DIAGNOSED USERS

We first evaluate the performance of our models in a binary context, that is, whether we can recognize diagnosed users versus non-diagnosed individuals (control users) based on the texts of their publications. By considering the users from the nine mental disorders as a single group (for instance, the diagnosed class), the balance between this class and the control class is preserved.

In both the Spanish and English Twitter datasets, XGBoost (XGB) and CNN showed marginally superior performance

**TABLE 7. Evaluation results for the binary classification for both datasets (Spanish and English) in each set of attributes. The values for the best AUC, F1-score, Precision, and Recall are in bold.**

| Feature type | | Spanish | | | | English | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | AUC | F1 | Precision | Recall | AUC | F1 | Precision | Recall |
| Word *n*-grams | Unigram (U) | 0.819 | 0.810 | 0.823 | 0.800 | 0.837 | 0.810 | 0.853 | 0.789 |
| | Bigram (B) | 0.830 | 0.817 | ***0.853*** | 0.784 | 0.771 | 0.748 | 0.780 | 0.723 |
| | Trigram (T) | 0.768 | 0.754 | 0.768 | 0.742 | 0.699 | 0.675 | 0.701 | 0.655 |
| | Word 1, 2, 3gram (UBT) | ***0.835*** | ***0.824*** | 0.849 | 0.803 | 0.835 | 0.805 | 0.855 | 0.781 |
| Char *q*-grams | Char 3gram (C3) | 0.802 | 0.793 | 0.796 | 0.792 | 0.807 | 0.788 | 0.816 | 0.769 |
| | Char 5gram (C5) | 0.819 | 0.811 | 0.820 | 0.806 | ***0.846*** | ***0.819*** | ***0.867*** | ***0.795*** |
| | Char 7gram (C7) | 0.822 | 0.814 | 0.820 | ***0.810*** | 0.836 | 0.809 | 0.854 | 0.786 |
| POS *n*-grams | POS unigram (POS_U) | 0.822 | 0.813 | 0.832 | 0.798 | 0.820 | 0.795 | 0.843 | 0.764 |
| | POS bigram (POS_B) | 0.821 | 0.807 | 0.844 | 0.772 | 0.740 | 0.725 | 0.736 | 0.717 |
| | POS trigram (POS_T) | 0.762 | 0.750 | 0.756 | 0.745 | 0.701 | 0.683 | 0.697 | 0.671 |
| | POS 1,2,3gram (POS_UBT) | 0.824 | 0.814 | 0.830 | 0.801 | 0.820 | 0.799 | 0.838 | 0.771 |
| Topic | LDA | 0.720 | 0.718 | 0.691 | 0.749 | 0.754 | 0.742 | 0.741 | 0.747 |
| | NMF | 0.724 | 0.725 | 0.693 | 0.762 | 0.774 | 0.759 | 0.773 | 0.749 |
| Psycholinguistic | LIWC | 0.754 | 0.749 | 0.729 | 0.770 | 0.746 | 0.734 | 0.737 | 0.734 |
| Embeddings | fastText | 0.782 | 0.779 | 0.752 | ***0.810*** | 0.808 | 0.787 | 0.823 | 0.760 |
| | GloVe | 0.759 | 0.761 | 0.744 | 0.803 | 0.809 | 0.796 | 0.825 | 0.769 |

compared to the other classifiers; hence, we report results for these two classifiers only (the results for the remaining five algorithms are detailed in Section III in the Supplementary Material). As explained in feature extraction, we used CNN with word embeddings such as fastText and GloVe, and XGBoost with the remaining attributes.

Table 7 summarizes the performance of the best two binary classification models for both datasets. AUC shows the overall performance of the classification model, while F1, precision, and recall show the score for the positive class.

For the Spanish Twitter dataset, UBT showed the highest AUC and F1-score on the diagnosed class (AUC 0.835; F1 0.824), while in the English Twitter dataset, C5 had the highest performance (AUC 0.846; F1 0.819).

Table 7 shows that models based on lexical attributes (*n*-grams and *q*-grams) and POS *n*-grams perform well for the Spanish Twitter dataset, except when $n = 3$ (for instance, Trigrams and POS_T). Top AUCs range from 0.802 to 0.835. Meanwhile, the attributes related to topic modeling, NMF, and LDA showed the lowest AUC, 0.720 and 0.724, respectively.

Regarding AUC and F1, fastText is better than GloVe in identifying users with mental health disorders. However, both word embedding methods do not improve their performance with respect to lexical attributes. The performance of LIWC is comparable to GloVe, with an AUC score of 0.754 and an F1-score of 0.749 in the diagnosed class.

Based on the AUC values, Friedman's test showed significant differences between the models ($\chi^2_{15}$=51.3, $p$-value $< 0.001$). The post-hoc revealed a significantly better score for the UBT model than LDA (R=14.8, $p$-value=0.002), NMF (R=14.2, $p$-value=0.002), LIWC (R=12.6, $p$-value=0.009), GloVe (R=12.2, $p$-value=0.01), POS-T (R=11.8, $p$-value=0.01), T (R=10.8, $p$-value=0.03), and fastText (R=10.6, $p$-value=0.03).

For the Twitter English dataset, lexical attributes based on *q*-grams ($q = 5$ and $q = 7$) and POS *n*-grams ($n = 1$ and $n = 1 - 3$) show the highest performance in terms of AUC and F1-score on the diagnosed class (Table 7). Trigrams and POS_T had the lowest performance of all attributes in terms of AUC, 0.699 and 0.701, respectively. A similar trend is also observed in the Spanish Twitter dataset.

The performance between both word embedding methods is very similar, but once again fastText (AUC 0.808; F1 0.787) and GloVe (AUC 0.809; F1 0.796) do not improve their performance with respect to lexical attributes, mainly C5, C7, POS_U, and POS_UBT. Regarding topic modeling attributes, LDA and NMF also showed low AUC, 0.754 and 0.774, respectively, and their F1-scores in classifying users with mental illness are about 0.75 on average.

The performance of LIWC is comparable to LDA and POS_B, with an AUC score of 0.746 and F1 of 0.734 in the diagnosed class. Based on the AUC values, Friedman's test revealed significant differences between the models ($\chi^2_{15}$=49.7, $p$-value $< 0.001$). The post-hoc revealed a significantly better score for the C5 model than T (R=15.6, $p$-value=0.001), POS_T (R=15.4, $p$-value=0.001), POS_B (R=13, $p$-value=0.007), LDA (R=11, $p$-value=0.04), B (R=11, $p$-value=0.04), and LIWC (R=10.8, $p$-value=0.04).

A Receiver Operator Characteristic (ROC) curve is constructed by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR). It shows the trade-off between sensitivity (or TPR) and specificity (1 - FPR). Classifiers that give curves closer to the top-left corner indicate a better performance. In contrast, a random classifier is anticipated to produce data points lying along the diagonal line, where the FPR equals the TPR. The closer a curve approaches the

45-degree diagonal within the ROC space, the less reliable the test's accuracy becomes.

The ROC curves in Fig. 2 and Fig. 3 show satisfactory performance of the models with the highest result, with an AUC of 0.84 and 0.85 for the Spanish and English datasets, respectively. In both cases, the optimal cut-off point is located at a point on the ROC curve where the sensitivity and specificity are approximately 83%. This means the models can correctly identify 83% of the positive and 83% of the negative classes.

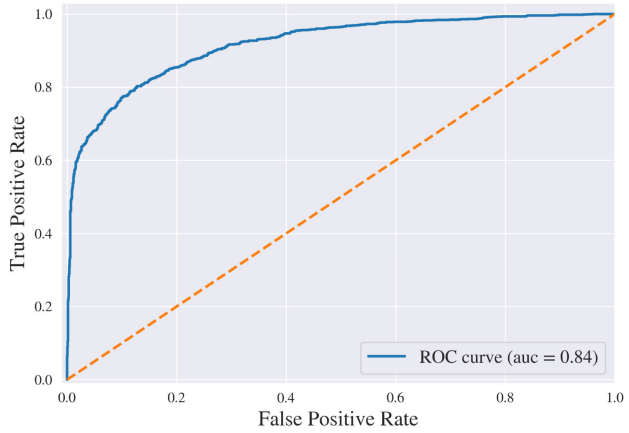


**FIGURE 2.** ROC curve for binary classification using UBT model in the Spanish Twitter dataset.

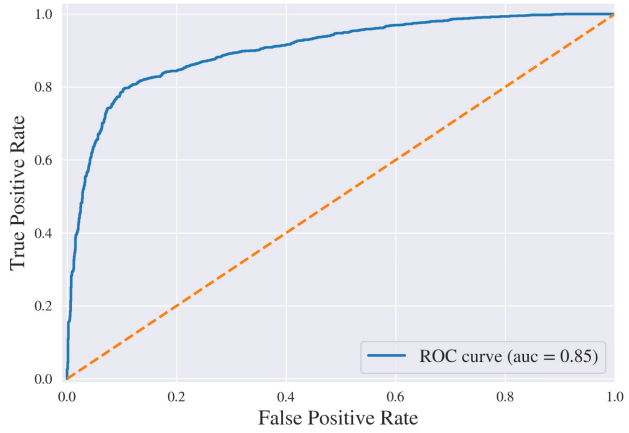


**FIGURE 3.** ROC curve for binary classification using C5 model in the English Twitter dataset.

Based on the results, in general, our proposed models can accurately detect potential users who may have psychological disorders using only their texts for both datasets.

### C. MULTICLASS CLASSIFICATION OF USERS WITH MENTAL HEALTH DISORDERS

In the second set of results, we evaluated the performance of our models when applying multiclass classification to identify the mental health condition a user was diagnosed using their timeline of tweets. In other words, we aim to recognize the type of diagnosis of each user.

Table 8 illustrates the macro-average (AUC, F1, precision, and recall) of all feature-based models used for both datasets. The macro-average computes the corresponding metric independently for each class and then takes the arithmetic mean. In general, if we are working with an imbalanced dataset where all classes are equally important (like ours), using the macro-average is a good choice, as it treats all classes equally. Once more, XGB and CNN performed better than the other classifiers, so we only report the results for these two classifiers (additional algorithm results are listed in Section IV in the Supplementary Material).

For both datasets (Spanish and English), based on the AUC, most of the classifiers perform above 0.5, thus, they are generally able to detect mental health disorders in a multiclass context.

For the Spanish Twitter dataset, UBT achieved the highest macro-average results (AUC 0.712; F1 0.501), while LIWC got the lowest results (AUC 0.538; F1 0.185). Based on the AUC values, Friedman's test showed significant differences between the models ($\chi^2_{15}$=66.8, $p$-value $< 0.001$). The post-hoc revealed a significantly better score for the UBT model than B (R=9.4, $p$-value=0.03), T (R=9.8, $p$-value=0.02), C3 (R=10.6, $p$-value=0.01), POS_B (R=11, $p$-value=0.008), POS_T (R=13.6, $p$-value<0.001), NMF (R=13.6, $p$-value<0.001), LDA (R=15, $p$-value<0.001), and LIWC (R=15.8, $p$-value<0.001).

In the English Twitter dataset, the Unigram model obtained the highest performance (AUC 0.697; F1 0.477), whereas POS_T got the lowest results (AUC 0.514; F1 0.118). Based on the AUC values, Friedman's test revealed significant differences between the models ($\chi^2_{15}$=68.6, $p$-value $< 0.001$). The post-hoc showed a significantly better score for the Unigram model than B (R=10, $p$-value=0.02), NMF (R=11.2, $p$-value=0.006), LDA (R=12.2, $p$-value=0.002), POS_B (R=12.8, $p$-value=0.001), LIWC (R=14.2, $p$-value< .001), T (R=14.8, $p$-value<.001), and POS_T (R=15.2, $p$-value<.001).

The ROC curves in Fig. 4 and Fig. 5 show a medium performance of the models with the highest result, with an AUC of 0.71 and 0.69 for the Spanish and English datasets, respectively.

According to the results, in general, our proposed models can classify users as belonging to one of nine mental disorders using only their texts for both datasets.

In multiclass classification, it is helpful to report the results of the experiments separated by classes as well. Therefore, Table 9 presents the evaluation metrics (AUC, F1, precision, and recall) detailed for each class of the best model for both datasets.

Results for individual disorders vary (Table 9). The Spanish Twitter dataset's AUC values range from 0.499 (PTSD) to 0.843 (ADHD). The PTSD class achieved a zero value in F1, precision, and recall. The cause of this may be that using UBT results in overfitting. That is, when using

**TABLE 8.** Overall evaluation results for the multiclass classification for both datasets (Spanish and English) in each set of attributes. The values for the best macro-average AUC, F1-score, Precision, and Recall are in bold.

| Feature Type | | Spanish | | | | English | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | AUC | F1 | Precision | Recall | AUC | F1 | Precision | Recall |
| Word *n*-grams | Unigram (U) | 0.702 | 0.491 | 0.573 | 0.466 | ***0.697*** | ***0.477*** | ***0.583*** | ***0.455*** |
| | Bigram (B) | 0.650 | 0.384 | 0.441 | 0.368 | 0.608 | 0.284 | 0.305 | 0.288 |
| | Trigram (T) | 0.646 | 0.389 | 0.479 | 0.361 | 0.518 | 0.128 | 0.179 | 0.141 |
| | Word 1, 2, 3gram (UBT) | ***0.712*** | ***0.501*** | 0.562 | ***0.480*** | 0.694 | 0.470 | 0.557 | 0.448 |
| Char *q*-grams | Char 3gram (C3) | 0.633 | 0.364 | 0.447 | 0.349 | 0.650 | 0.379 | 0.438 | 0.371 |
| | Char 5gram (C5) | 0.683 | 0.444 | 0.509 | 0.429 | 0.684 | 0.440 | 0.535 | 0.426 |
| | Char 7gram (C7) | 0.702 | 0.492 | 0.569 | 0.466 | 0.680 | 0.447 | 0.534 | 0.423 |
| POS *n*-grams | POS unigram (POS_U) | 0.696 | 0.485 | ***0.580*** | 0.456 | 0.663 | 0.415 | 0.496 | 0.391 |
| | POS bigram (POS_B) | 0.626 | 0.331 | 0.379 | 0.323 | 0.551 | 0.195 | 0.269 | 0.196 |
| | POS trigram (POS_T) | 0.596 | 0.273 | 0.336 | 0.270 | 0.514 | 0.118 | 0.182 | 0.134 |
| | POS 1, 2, 3gram (POS_UBT) | 0.687 | 0.442 | 0.497 | 0.433 | 0.663 | 0.410 | 0.500 | 0.392 |
| Topic | LDA | 0.558 | 0.214 | 0.250 | 0.213 | 0.550 | 0.190 | 0.214 | 0.194 |
| | NMF | 0.590 | 0.276 | 0.317 | 0.268 | 0.572 | 0.227 | 0.252 | 0.229 |
| Psycholinguistic | LIWC | 0.538 | 0.185 | 0.224 | 0.179 | 0.532 | 0.159 | 0.208 | 0.165 |
| Embeddings | fastText | 0.667 | 0.417 | 0.482 | 0.399 | 0.641 | 0.359 | 0.442 | 0.349 |
| | GloVe | 0.675 | 0.436 | 0.548 | 0.414 | 0.620 | 0.308 | 0.346 | 0.311 |

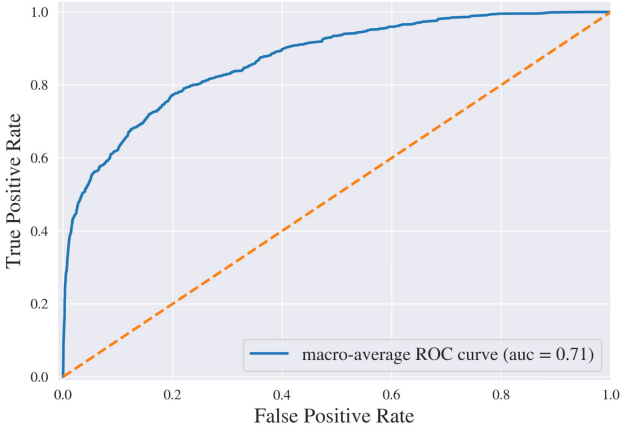


**FIGURE 4.** ROC curve for multiclass classification using UBT model in the Spanish Twitter dataset.

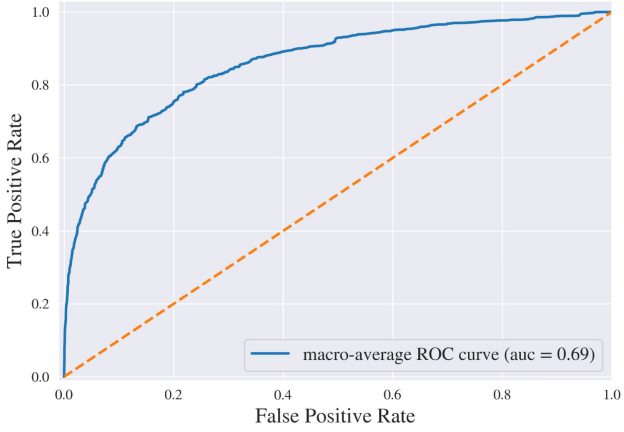


**FIGURE 5.** ROC curve for multiclass classification using Unigram model in the English Twitter dataset.

the unigram, bigram, and trigram combination, the feature dimensions are too high, degrading the performance of PTSD. Thus, discarding the PTSD class, F1 has values between 0.207 (Schizophrenia) and 0.745 (ADHD). Moreover, five out of nine disorders achieved a precision between 0.579 (OCD) and 0.776 (ADHD). Precision is generally higher (>0.50) than recall, except for Anx and Depr.

Not only does the number of users per disorder impact the performance of the classification, but also the degree to which the class is modeled by the language used by individuals of each mental health condition. For instance, the Schizophrenia class contains the lowest number of diagnosed users (48/1,551, 3.1%) scored an AUC of 0.568, compared to the PTSD class (65/1,551, 4.2%), which scored an AUC of 0.499. The neurodevelopmental-related disorders, ADHD (206/1,551, 13.3%) and Auts (119/1,551, 7.7%), achieved the highest AUC results, 0.843 and 0.806, respectively.

According to Table 9, for the English Twitter dataset, AUC values range from 0.544 (Anx) to 0.784 (ADHD). Most F1 values are lower than 0.6, except for ADHD, which is 0.751. Similarly to the Spanish Twitter dataset, overall precision is higher than recall (>0.50), but now the exception is ADHD and Bipo.

For the English Twitter dataset, we can see well-defined classes such as Bipo (136/1,543, 8.8%) and Auts (170/1,543, 11.0%) that showed high AUC with the Unigrams model, 0.758 and 0.717, respectively, even though their F1-scores are slightly above 0.5. ADHD (622/1,543, 40.3%) is the class that has the most diagnosed users and presents the best AUC (0.784). Although Eat (24/1,543, 1.7%) and Schizophrenia (24/1,543, 1.6%) classes contain less than

**TABLE 9.** Multiclass classification results (breakdown by disorder) of the best models for the Spanish and English Twitter datasets.

| Dataset | Feature type | Metric | ADHD | Anx | Auts | Bipo | Depr | Eat | OCD | PTSD | Schizophrenia | Macro Avg[a] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Spanish | Word 1, 2, 3gram (UBT) | AUC | 0.843 | 0.727 | 0.806 | 0.756 | 0.775 | 0.762 | 0.676 | 0.499 | 0.568 | 0.712 |
| | | F1 | 0.745 | 0.561 | 0.679 | 0.570 | 0.683 | 0.616 | 0.447 | 0 | 0.207 | 0.501 |
| | | Precision | 0.776 | 0.503 | 0.743 | 0.621 | 0.634 | 0.711 | 0.579 | 0 | 0.490 | 0.562 |
| | | Recall | 0.719 | 0.644 | 0.630 | 0.529 | 0.745 | 0.546 | 0.366 | 0 | 0.144 | 0.480 |
| English | Unigram (U) | AUC | 0.784 | 0.544 | 0.717 | 0.758 | 0.671 | 0.692 | 0.699 | 0.733 | 0.674 | 0.697 |
| | | F1 | 0.751 | 0.158 | 0.524 | 0.534 | 0.462 | 0.429 | 0.460 | 0.533 | 0.438 | 0.477 |
| | | Precision | 0.662 | 0.533 | 0.596 | 0.516 | 0.504 | 0.633 | 0.554 | 0.580 | 0.667 | 0.583 |
| | | Recall | 0.867 | 0.104 | 0.476 | 0.565 | 0.433 | 0.387 | 0.415 | 0.497 | 0.350 | 0.455 |

a Macro average between the nine classes.

25 users, these classes achieved an AUC of 0.692 and 0.674, respectively.

Additionally, we use the normalized confusion matrices to analyze the distribution of all predicted responses with their true classes (Fig. 6 and Fig. 7). According to the matrix, each row represents an instance of the actual class, while a column represents an instance of the predicted class. Therefore, the diagonal values represent the degree of correctly predicted classes.

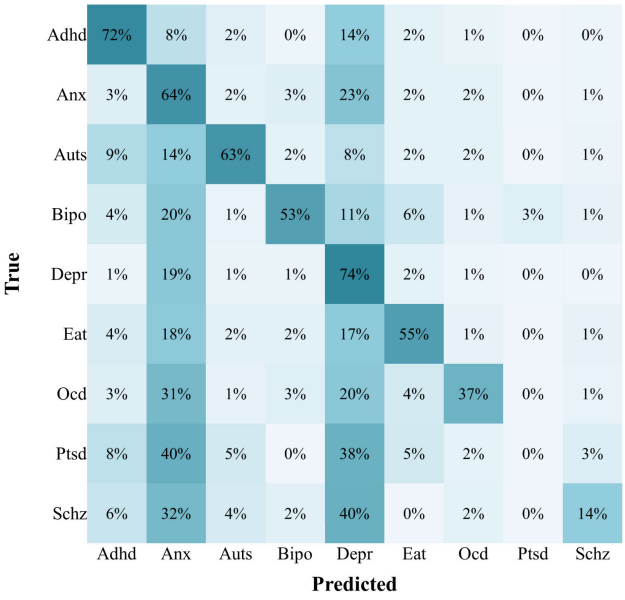


**FIGURE 6.** Normalized confusion matrix of UBT model in the Spanish Twitter dataset.

According to the diagonal values in the Spanish Twitter dataset (Fig. 6), the model correctly predicted 72% and 74% of the instances of ADHD and Depr, respectively, while for Eat, Auts, and Anx classes, between 55% and 65% of the instances were correctly predicted. From the diagonal values in the English Twitter dataset (Fig. 7), we can say that the model correctly predicts only two classes, ADHD and Bipo, 87% and 56%, respectively. All other classes are below 50%.

A trend towards mislabeled instances in the prediction phase can be observed for both datasets. Most of the incorrect predictions in the Spanish-language dataset are in the Depr

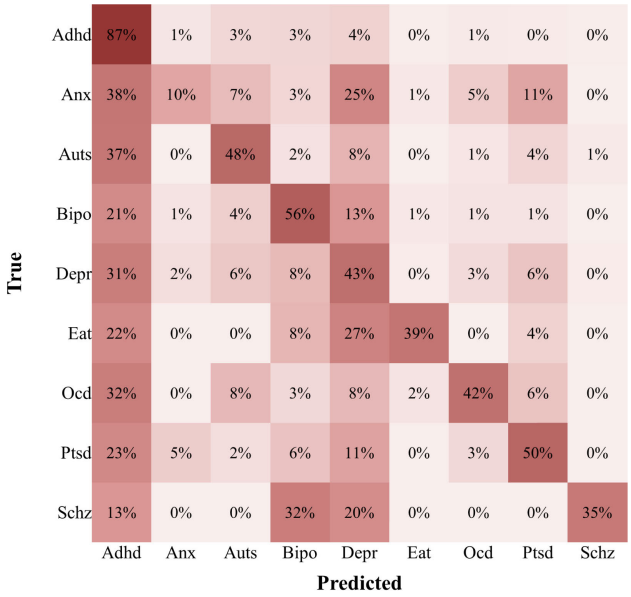


**FIGURE 7.** Normalized confusion matrix of Unigram model in the English Twitter dataset.

and Anx classes, while in the English-language dataset, they lie in the Depr and ADHD classes.

When removing the majority class for both datasets, there is a slight increase in classification performance (Section V in the Supplementary Material). For instance, the macro-average AUC in the Spanish Twitter dataset is now 0.737, compared with 0.712. In the English Twitter dataset, the macro-average AUC increased from 0.697 to 0.712.

### D. COMPARISON WITH THE SMHD DATASET

In this section, we applied our models for detecting mental disorders to another available corpus to investigate the robustness of the proposed models for both classification tasks. This will allow us to gain some insights on how feature-based models perform in a dataset with the same number of mental health conditions but in a different social media, Reddit, instead of Twitter. To the best of our knowledge, no directly comparable work assesses the performance of machine learning and deep learning methods

on the SMHD dataset for the two classification tasks we consider.

The SMHD dataset, collected by Cohan et al. [15], consists of Reddit posts from users with mental health disorders along with matched control users. The original dataset contains 20,406 diagnosed users and 335,952 control users. Due to the number of users, we selected a sample of up to 700 users for the diagnosed classes and 3,259 users for the control class. By using these values, we tried to obtain a population more or less similar to our collected dataset on Twitter. Hence, the average IR between the diagnosed and the control group ranges from 4.6 (ADHD, Anx, Bipo, and Depr) to 22.79 (Eat).

Table 10 shows the statistics of the final data subset used for our experiments. The average number of posts made per user is between 145 and 158 in the diagnosed classes and 283 posts in the control class. The mean number of posts in the SMHD dataset is lower compared to our dataset.

**TABLE 10. Main statistics of the SMHD data subset.**

| Class | Users | Total posts | Posts per user, mean (SD) |
|---|---|---|---|
| ADHD | 700 | 108,302 | 154.7 (81.1) |
| Anx | 700 | 107,927 | 154.2 (83.2) |
| Auts | 575 | 90,301 | 157.0 (83.0) |
| Bipo | 700 | 102,796 | 146.9 (79.8) |
| Depr | 700 | 109,210 | 156.0 (82.3) |
| Eat | 143 | 21,307 | 149.0 (74.7) |
| OCD | 405 | 59,105 | 145.9 (77.7) |
| PTSD | 475 | 75,102 | 158.1 (86.2) |
| Schizophrenia | 311 | 46,404 | 149.2 (77.2) |
| Control | 3,259 | 923,684 | 283.4 (157.0) |

Using the SMHD data subset, Table 11 shows the five highest results for binary classification based on the AUC values. The top five results were obtained using *n*-grams (UBT), *q*-grams (C5), LIWC, and both word-embedding methods (fastText and GloVe) as attributes.

The AUC with LIWC features, 0.823, is the highest of all the tested models. Moreover, this model performs better in this set than in the Spanish and English Twitter sets, which achieved an AUC value of 0.754 and 0.746, respectively. This is explained by the fact that the percentage of LIWC dictionary words found in SMHD is higher (mean 83.7, SD 6.6) than in the Spanish (mean 74.3, SD 7.4) and English (mean 76.7, SD 9.0) Twitter datasets.

By applying fastText and GloVe, it was possible to convert 98% of the words in this dataset. Nevertheless, while the AUC is very similar to the English Twitter dataset, in which a lower percentage of words were converted (between 81.3% and 83.9%), both F1-score and precision are higher.

Regarding multiclass classification, Table 12 shows the metric results for the best model for this task. C5 combined with XGB yields an AUC of 0.536, meaning the model has no discriminatory ability to distinguish between the nine types of disorders. Only one class, ADHD, obtained a value close to 0.6, making it barely distinguishable.

**TABLE 11. Top 5 results for the binary classification on the SMHD data subset. The highest AUC, F1-score, Precision, and Recall values are in bold.**

| Feature type | AUC | F1 | Precision | Recall |
|---|---|---|---|---|
| UBT + XGB | 0.818 | 0.865 | 0.979 | 0.775 |
| C5 + XGB | 0.814 | 0.861 | 0.978 | 0.768 |
| LIWC + XGB | ***0.823*** | 0.881 | 0.977 | 0.802 |
| fastText + CNN | 0.810 | ***0.884*** | 0.972 | ***0.810*** |
| GloVe + CNN | 0.818 | 0.848 | ***0.982*** | 0.747 |

### E. RESULTS SUMMARY

In order to ease the comparison of the best results across each dataset, Table 13 shows the AUC and F1 values for the Spanish and English Twitter datasets (our collected data) and for the SMHD data subset.

## V. DISCUSSION

This study addresses the problem of automatically classifying social media users into different mental health conditions based on their posting history.

For this purpose, we provide two datasets extracted from Twitter, in Spanish and English, and annotate each one with approximately 1,500 users diagnosed with one of nine different mental disorders and 1,700 matched-control users. For both datasets, the outcome is just over 3,000 Twitter users with their corresponding timelines (the texts retrieved from each user cover at least three months of activity on the social media), which support two user-level classification tasks, binary and multiclass.

We evaluate the datasets using Machine Learning and Deep Learning classification models, trained with different text-related attributes such as *n*-grams, *q*-grams, POS tags, topic modeling, LIWC, and word embeddings. We use binary classification to assess whether a user potentially suffers from a mental health disorder and multiclass classification to determine the condition to which the user relates.

Using our collected dataset, for binary classification (Table 7), we achieved an AUC of 0.835 on the Spanish Twitter dataset using *n*-grams of words from one to three (UBT) and 0.846 on the English Twitter dataset with a 5-gram characters (C5) model. Furthermore, in multiclass classification (Table 8), we obtained an AUC of 0.712 and 0.697 in the Spanish and English Twitter datasets, respectively, so we were able to categorize Twitter users into one of the nine categories of mental health conditions.

These findings suggest that using feature-based models in our collected datasets can detect users with mental health disorders and the type of user diagnosis, as long as there is enough data history to train the models.

Predictive performance can vary considerably depending on the feature type and the task. Regarding binary classification, in the Spanish Twitter dataset, when using more sparse features such as lexical features and POS tags (except for Trigrams and POS_T), the AUC and F1-score increase versus less sparse features such as topic modeling, which uses a

**TABLE 12.** Multiclass classification evaluation results (breakdown by disorder) of the best model (C5 + XGB) on the SMHD data subset.

| Metric | ADHD | Anx | Auts | Bipo | Depr | Eat | OCD | PTSD | Schizophrenia | Macro Avg[a] |
|---|---|---|---|---|---|---|---|---|---|---|
| AUC | 0.596 | 0.525 | 0.566 | 0.524 | 0.508 | 0.519 | 0.501 | 0.568 | 0.515 | 0.536 |
| F1 | 0.276 | 0.187 | 0.237 | 0.179 | 0.159 | 0.073 | 0.019 | 0.240 | 0.065 | 0.159 |
| Precision | 0.212 | 0.146 | 0.243 | 0.149 | 0.130 | 0.429 | 0.154 | 0.410 | 0.545 | 0.269 |
| Recall | 0.395 | 0.260 | 0.230 | 0.225 | 0.205 | 0.040 | 0.010 | 0.170 | 0.034 | 0.174 |

a Macro average between the nine classes.

**TABLE 13.** For both classification tasks (binary and multiclass), AUC and F1 scores of the best model for each dataset (Spanish, English and SMDH).

| Dataset | Binary Classification | | | Multiclass classification | | |
|---|---|---|---|---|---|---|
| | Best model | AUC | F1 | Best model | AUC | F1 |
| Spanish Twitter dataset | UBT + XGB | 0.835 | 0.824 | UBT + XGB | 0.712 | 0.501 |
| English Twitter dataset | C5 + XGB | 0.846 | 0.819 | U + XGB | 0.697 | 0.477 |
| SMHD data subset | LIWC + XGB | 0.823 | 0.881 | C5 + XGB | 0.536 | 0.159 |

50-dimensional space. For LIWC and word embeddings, the performance of the classifiers will depend on the percentage of words found in the dataset by the dictionaries and pre-trained words.

For the case of the English Twitter dataset, as we increase the N-word sequence, $n$-grams do not do a better job of modeling the training corpus, degrading performance when $n = 2$ and $n = 3$ (with and without attached POS tags). A similar trend is also observed in this corpus when using a less sparse feature space. For the case of word embeddings, the performance is related not only to the percentage of words found, but also to the number of words with which the model was pre-trained.

The Common Crawl corpus contains petabytes of data collected over 12 years of web crawling [55]. The corpus contains between 640 and 800 billion tokens, while the SBWC contains only 1.4 billion. For this reason, the models in the English-language obtain AUC values of 0.80 as opposed to the Spanish language, which is between 0.759 and 0.782.

Bearing in mind the number of classes in the collected dataset, the performance for multiclass classification is acceptable (Table 8), although there is some variability in class-level results (Table 9). As we stated before, the number of users per disorder and the degree to which the class is modeled by the language used by those with each mental health condition affects its performance. For the Spanish Twitter dataset, ADHD, Auts, Depr, and Eat are the more distinguishable classes of a pool of mental disorders, while for the English Twitter dataset, these are ADHD, Bipo, Auts, and PTSD.

Based on the confusion matrices (Fig. 6 and Fig. 7), we can observe that due to the number of samples, there is a bias towards the majority class, which leads to a significant proportion of the instances of the remaining disorders being classified in both the Depr (Spanish Twitter dataset) and ADHD (English Twitter dataset) classes. Furthermore, we conducted an experiment to verify the impact of removing the majority class on the performance of the two best classification models presented in multiclass classification (Section V in the Supplementary Material). Even though the increase in overall classifier performance is minimal, some noteworthy changes were observed in the distribution of predictions of the resulting confusion matrix. For example, in the Spanish Twitter dataset, correct predictions improved by 22% in Anx, while in the English Twitter dataset, correct predictions improved by 21% in Auts and Depr, 16% in Bipo, and 13% in Anx.

Our study also investigated the robustness of feature-based models on additional resources, so we used the SMHD dataset from Reddit and created a subset that better matches the dimensions of our datasets. While the number of posts per user is smaller than in our dataset, Reddit postings are not limited by the character limit; thus, we get longer texts per user.

Using this subset, we provide benchmark results for the two classification tasks to facilitate further research and conduct extensive experiments. Therefore, in the SMHD data subset, the top five binary classification results (Table 11) are very close to the results obtained in our datasets. However, in the multiclass classification task, the highest macro-average AUC is just 0.536 (Table 12), meaning there is a content overlap across all mental health disorders, making it impossible to distinguish between the nine conditions.

Finally, our goal for an automatic detection system is to design classifiers that maximize true positives (i.e., mental health disorders) and minimize false positives (i.e., false mental disorder cases). We base our analysis on the AUC mainly because it is a metric with low sensitivity to class imbalance, a problem observed in multiclass classification, and gives an idea of the classifier's overall performance.

Precision and Recall are proportional to true positives but inversely related; for instance, a classifier that maximizes precision will yield only robust positive predictions, missing positive events. On the other hand, for high recall, the classifier assigns more examples to the positive class to reduce false negatives. Whether to maximize Recall or Precision depends on the medical use case.

Untreated mental health conditions can lead to poor quality of life, unemployment, substance abuse, homelessness, and in some cases suicide [1]. In the context of detecting mental health disorders, Recall is critical. For instance, receiving multiple false alerts (false diagnosis cases) is probably preferable to missing a true one (false healthy condition).

## VI. LIMITATIONS AND THREATS TO VALIDITY

This study has several limitations. First, Twitter is well-known for spreading misinformation. Even though it has become a more open topic in recent years, given the social stigma commonly associated with mental disorders, it seems unlikely that users would intentionally mislead about their condition [19]. We leveraged self-reported diagnosis using regex matching to build the dataset, but there is no way to verify the legitimacy of the postings (e.g., confirmation from medical experts). Second, the users collected from Twitter are not representative of the entire population. We studied the part of the population related to mental disorders and who also disclosed it on social media. However, how they communicate and interact in non-traditional media is a necessary step to further understanding the mechanisms behind mental disorders. Third, for the control group, age and gender matching with the diagnosed group are desired to avoid biases originating from gender or age differences in language use [18], [25]. This strategy was discarded owing to the computational time required for calculating these features for each user in the dataset and the lack of Spanish language resources for this task. We used post-count-matching [15] combined with users identified through mental health-related hashtags to build the control group. However, automatic solutions to construct a control group that is not entirely unrelated to the diagnosed group require additional study. Fourth, in the dataset collected from Twitter, the imbalance between the diagnosed classes affects the detection performance at multiclass levels. Strategies that mitigate such majority class bias should be implemented.

## VII. CONCLUSION AND FUTURE WORK

Mental health problems do not improve on their own; the longer a disorder persists, the more challenging it can be to treat and recover from. Identifying the potential risk for a user to suffer from a mental disorder is significant for developing automated applications to recognize early signs of mental disorders and the subsequent targeted interventions.

This study presents two annotated corpora consisting of Twitter users in two of the most widely used languages worldwide. We also exploit a wide range of linguistic features to create different machine-learning models to detect mental health disorders using each user's tweet history over time.

According to the results, both datasets and the suggested models support the two proposed classification tasks: identifying users related to a mental disorder diagnosis and the type of mental health disorder they are referring to.

Regarding the algorithms, XGBoost and CNN models outperformed other classifiers. Furthermore, although there is no particular feature set that works best for both languages, we can observe that lexical attributes, such as *n*-grams (U and UBT), *q*-grams (C5 and C7), and POS *n*-grams (POS_U and POS_UBT) performed well in the Spanish and English Twitter datasets -for binary and multiclass classification-. The advantage of *n*-grams over embeddings is that they can capture more specific details in a text, especially when dealing with short texts. *N*-grams are also helpful when dealing with rare or out-of-vocabulary words since they can still be included in the analysis as part of an n-gram. These models do not depend on the percentage of words found in the dataset by the dictionaries -for example, all the input to LIWC must be grammatically correct- or the number of pre-trained words in the corpus.

For future work, we plan to evaluate the performance of PBC4cip [56], a contrast pattern-based classifier that has performed well on tasks such as detecting depressive and xenophobic tweets [57], [58]. Also, compared to popular ML algorithms, it is interpretable, resulting in a set of patterns that provide a deeper understanding of how language is manifested in these mental disorders.

In addition, more linguistic features from user's tweet history can be extracted to determine the relationship between linguistic markers associated with mental health disorders, and more advanced deep learning architectures (e.g., HAN and transformers) can also be explored to detect mental disorders from text data to improve classification performance in both tasks.

Finally, researchers can consider more comprehensive validation, comprising the robustness assessment in practical real-time scenarios. These tests may include collaboration with clinical partners, acquiring real-world data, and addressing the technical and logistical challenges that such validation entails.

## ACKNOWLEDGMENT

Miryam Elizabeth Villa-Pérez would like to thank the support given by Tecnologico de Monterrey and Consejo Nacional de Humanidades, Ciencias y Tecnología (CONAHCYT) to pursue her graduate studies.

## REFERENCES

[1] World Health Organization. *World Mental Health Day: An Opportunity to Kick-Start a Massive Scale-Up in Investment in Mental Health*. Accessed: Mar. 16, 2023. [Online]. Available: https://www.who.int/news/item/27-08-2020-world-mental-health-day-an-opportunity-to-kick-start-a-massive-scale-up-in-investment-in-mental-health

[2] J. A. Vaingankar, S. A. Chong, E. Abdin, F. D. S. Kumar, B. Y. Chua, R. Sambasivam, S. Shafie, A. Jeyagurunathan, E. Seow, and M. Subramaniam, "Understanding the relationships between mental disorders, self-reported health outcomes and positive mental health: Findings from a national survey," *Health Quality Life Outcomes*, vol. 18, no. 1, p. 55, Mar. 2020.

[3] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in *Proc. 7th Int. AAAI Conf. Web Social Media*, Aug. 2021, vol. 7, no. 1, pp. 128–137.

[4] Z. Jamil, D. Inkpen, P. Buddhitha, and K. White, "Monitoring tweets for depression to detect at-risk users," in *Proc. 4th Workshop Comput. Linguistics Clin. Psychol., From Linguistic Signal Clin. Reality*. Stroudsburg, PA, USA: Association for Computational Linguistics, Aug. 2017, pp. 32–40.

[5] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of depression-related posts in Reddit social media forum," *IEEE Access*, vol. 7, pp. 44883–44893, 2019.

[6] A. Yates, A. Cohan, and N. Goharian, "Depression and self-harm risk assessment in online forums," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 2968–2978.

[7] S. Fodeh, T. Li, K. Menczynski, T. Burgette, A. Harris, G. Ilita, S. Rao, J. Gemmell, and D. Raicu, "Using machine learning algorithms to detect suicide risk factors on Twitter," in *Proc. Int. Conf. Data Mining Workshops (ICDMW)*, Nov. 2019, pp. 941–948.

[8] G. Coppersmith, M. Dredze, C. Harman, K. Hollingshead, and M. Mitchell, "CLPsych 2015 shared task: Depression and PTSD on Twitter," in *Proc. 2nd Workshop Comput. Linguistics Clin. Psychol., From Linguistic Signal Clin. Reality*. Stroudsburg, PA, USA: Association for Computational Linguistics, Jun. 2015, pp. 31–39.

[9] 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*, vol. 10456, G. J. Jones, S. Lawless, J. Gonzalo, L. Kelly, L. Goeuriot, T. Mandl, L. Cappellato, and N. Ferro, Eds. Cham, Switzerland: Springer, 2017, pp. 346–360.

[10] D. E. Losada, F. Crestani, and J. Parapar, "Overview of eRisk: Early risk prediction on the Internet," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, vol. 11018, P. Bellot, C. Trabelsi, J. Mothe, F. Murtagh, J. Y. Nie, L. Soulier, E. SanJuan, L. Cappellato, and N. Ferro, Eds. Cham, Switzerland: Springer, 2018, pp. 343–361.

[11] D. E. Losada, F. Crestani, and J. Parapar, "Overview of eRisk 2019 early risk prediction on the Internet," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, vol. 11696, F. Crestani, M. Braschler, J. Savoy, A. Rauber, H. Müller, D. E. Losada, G. H. Bürki, L. Cappellato, and N. Ferro, Eds. Cham, Switzerland: Springer, 2019, pp. 340–357.

[12] A. Wongkoblap, M. A. Vadillo, and V. Curcin, "Researching mental health disorders in the era of social media: Systematic review," *J. Med. Internet Res.*, vol. 19, no. 6, p. e228, Jun. 2017.

[13] Statista. *Leading Countries Based on Number of Twitter Users as of January 2022 (in Millions)*. Accessed: Jul. 20, 2023. [Online]. Available: https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries

[14] R. Skaik and D. Inkpen, "Using social media for mental health surveillance," *ACM Comput. Surveys*, vol. 53, no. 6, pp. 1–31, Dec. 2020.

[15] A. Cohan, B. Desmet, A. Yates, L. Soldaini, S. MacAvaney, and N. Goharian, "SMHD: A large-scale resource for exploring online language usage for multiple mental health conditions," in *Proc. 27th Int. Conf. Comput. Linguistics*. Santa Fe, NM, USA: Association for Computational Linguistics, Aug. 2018, pp. 1485–1497. [Online]. Available: https://aclanthology.org/C18-1126

[16] A.-S. Uban, B. Chulvi, and P. Rosso, "An emotion and cognitive based analysis of mental health disorders from social media data," *Future Gener. Comput. Syst.*, vol. 124, pp. 480–494, Nov. 2021.

[17] M. Trotzek, S. Koitka, and C. M. Friedrich, "Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 3, pp. 588–601, Mar. 2020.

[18] G. Coppersmith, M. Dredze, and C. Harman, "Quantifying mental health signals in Twitter," in *Proc. Workshop Comput. Linguistics Clin. Psychol., From Linguistic Signal Clin. Reality*. Baltimore, MD, USA: Association for Computational Linguistics, Jun. 2014, pp. 51–60.

[19] D. E. Losada and F. Crestani, "A test collection for research on depression and language use," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, vol. 9822. Cham, Switzerland: Springer, 2016, pp. 28–39.

[20] K. Harrigian, C. Aguirre, and M. Dredze, "On the state of social media data for mental health research," in *Proc. 7th Workshop Comput. Linguistics Clin. Psychol., Improving Access*, Stroudsburg, PA, USA: Association for Computational Linguistics, Jun. 2021, pp. 15–24. [Online]. Available: https://aclanthology.org/2021.clpsych-1.2

[21] D. E. Losada, F. Crestani, and J. Parapar, "eRisk 2020: Self-harm and depression challenges," in *Advances in Information Retrieval* (Lecture Notes in Computer Science), J. M. Jose, E. Yilmaz, J. Magalhães, P. Castells, N. Ferro, M. J. Silva, and F. Martins, Eds. Cham, Switzerland: Springer, 2020, pp. 557–563.

[22] J. Aguilera, D. I. H. Farías, R. M. Ortega-Mendoza, and M. Montes-Y-Gómez, "Depression and anorexia detection in social media as a one-class classification problem," *Int. J. Speech Technol.*, vol. 51, no. 8, pp. 6088–6103, Aug. 2021.

[23] S. C. Guntuku, J. R. Ramsay, R. M. Merchant, and L. H. Ungar, "Language of ADHD in adults on social media," *J. Attention Disorders*, vol. 23, no. 12, pp. 1475–1485, Oct. 2019.

[24] M. Mitchell, K. Hollingshead, and G. Coppersmith, "Quantifying the language of schizophrenia in social media," in *Proc. 2nd Workshop Comput. Linguistics Clin. Psychol., From Linguistic Signal Clin. Reality*. Denver, CO, USA: Association for Computational Linguistics, Jun. 2015, pp. 11–20.

[25] G. Coppersmith, M. Dredze, C. Harman, and K. Hollingshead, "From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses," in *Proc. 2nd Workshop Comput. Linguistics Clin. Psychol., From Linguistic Signal Clin. Reality*. Denver, CO, USA: Association for Computational Linguistics, Jun. 2015, pp. 1–10.

[26] X. Chen, M. Sykora, T. Jackson, S. Elayan, and F. Munir, "Tweeting your mental health: Exploration of different classifiers and features with emotional signals in identifying mental health conditions," in *Proc. 51st Hawaii Int. Conf. Syst. Sci.*, Jan. 2018, pp. 3320–3328.

[27] D. J. Joshi, M. Makhija, Y. Nabar, N. Nehete, and M. S. Patwardhan, "Mental health analysis using deep learning for feature extraction," in *Proc. ACM India Joint Int. Conf. Data Sci. Manage. Data*. New York, NY, USA: Association for Computing Machinery, Jan. 2018, pp. 356–359.

[28] A. H. Orabi, P. Buddhitha, M. H. Orabi, and D. Inkpen, "Deep learning for depression detection of Twitter users," in *Proc. 5th Workshop Comput. Linguistics Clin. Psychol., From Keyboard Clinic*. New Orleans, LA, USA: Association for Computational Linguistics, Jun. 2018, pp. 88–97.

[29] J. Kim, J. Lee, E. Park, and J. Han, "A deep learning model for detecting mental illness from user content on social media," *Sci. Rep.*, vol. 10, no. 1, p. 11846, Jul. 2020.

[30] K. Malviya, B. Roy, and S. Saritha, "A transformers approach to detect depression in social media," in *Proc. Int. Conf. Artif. Intell. Smart Syst. (ICAIS)*, Mar. 2021, pp. 718–723.

[31] S. Tsugawa, Y. Kikuchi, F. Kishino, K. Nakajima, Y. Itoh, and H. Ohsaki, "Recognizing depression from Twitter activity," in *Proc. 33rd Annu. ACM Conf. Human Factors Comput. Syst.* Stroudsburg, PA, USA: Association for Computing Machinery, Apr. 2015, pp. 3187–3196.

[32] S. Almouzini, M. Khemakhem, and A. Alageel, "Detecting Arabic depressed users from Twitter data," *Proc. Comput. Sci.*, vol. 163, pp. 257–265, Jan. 2019.

[33] A. Leis, F. Ronzano, M. A. Mayer, L. I. Furlong, and F. Sanz, "Detecting signs of depression in tweets in Spanish: Behavioral and linguistic analysis," *J. Med. Internet Res.*, vol. 21, no. 6, Jun. 2019, Art. no. e14199.

[34] A. H. Uddin, D. Bapery, and A. S. M. Arif, "Depression analysis from social media data in Bangla language using long short term memory (LSTM) recurrent neural network technique," in *Proc. Int. Conf. Comput., Commun., Chem., Mater. Electron. Eng.*, Jul. 2019, pp. 1–4.

[35] Statista. (2023). *Most Popular Social Networks Worldwide as of January 2023, Ranked by Number of Monthly Active Users (in Millions)*. Accessed: Mar. 16, 2023. [Online]. Available: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users

[36] Statista. *Distribution of Twitter Users Worldwide as of April 2021, by Age Group*. Accessed: Jul. 20, 2023. [Online]. Available: https://www.statista.com/statistics/283119/age-distribution-of-global-twitter-users

[37] *Twitter Developer*. Accessed: Oct. 20, 2022. [Online]. Available: https://developer.twitter.com/en/docs/twitter-api

[38] *Diagnostic and Statistical Manual of Mental Disorders*, 5th ed., American Psychiatric Association, Virginia, U.K., 2013.

[39] A. Orriols-Puig and E. Bernadó-Mansilla, "Evolutionary rule-based systems for imbalanced data sets," *Soft Comput.*, vol. 13, no. 3, pp. 213–225, Feb. 2009.

[40] J. M. Pérez, J. C. Giudici, and F. Luque. (2021). *Pysentimiento: A Python Toolkit for Opinion Mining and Social NLP Tasks*. [Online]. Available: https://github.com/pysentimiento/pysentimiento

[41] S. Bird, E. Klein, and E. Loper, *Natural Language Processing With Python: Analyzing Text With The Natural Language Toolkit*. Springfield, MO, USA: O'Reilly Media, 2009.

[42] M. I. Honnibal. (2017). *Spacy: Natural Language Processing*. Accessed: Oct. 22, 2022. [Online]. Available: https://spacy.io/usage/linguistic-features

[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," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Dec. 2011.
[44] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods," *J. Lang. Social Psychol.*, vol. 29, no. 1, pp. 24–54, Mar. 2010.
[45] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. [Online]. Available: https://aclanthology.org/D14-1162
[46] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, Dec. 2017.
[47] T. Mikolov, K. Che, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. Int. Conf. Learn. Represent.*, 2013, pp. 1–12.
[48] C. Cardellino. (Aug. 2019). *Spanish Billion Words Corpus and Embeddings*. Accessed: Oct. 19, 2022. [Online]. Available: https://crscardellino.github.io/SBWCE/
[49] *English Word Vectors · FastText*. Accessed: Oct. 31, 2022. [Online]. Available: https://fasttext.cc/docs/en/english-vectors.html
[50] J. Pennington, R. Socher, and C. Manning. *Glove: Global Vectors for Word Representation*. Accessed: Oct. 31, 2022. [Online]. Available: https://nlp.stanford.edu/projects/glove/
[51] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.
[52] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Dec. 2006.
[53] S. García, A. Fernández, J. Luengo, and F. Herrera, "Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power," *Inf. Sci.*, vol. 180, no. 10, pp. 2044–2064, May 2010.
[54] J. Derrac, S. García, D. Molina, and F. Herrera, "A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms," *Swarm Evol. Comput.*, vol. 1, no. 1, pp. 3–18, Mar. 2011.
[55] *Want to Use our Data?—Common Crawl*. Accessed: Nov. 21, 2022. [Online]. Available: https://commoncrawl.org/the-data/
[56] O. Loyola-González, M. A. Medina-Pérez, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, R. Monroy, and M. García-Borroto, "PBC4cip: A new contrast pattern-based classifier for class imbalance problems," *Knowl.-Based Syst.*, vol. 115, pp. 100–109, Jan. 2017.
[57] L. M. G. Salazar, O. Loyola-González, and M. A. Medina-Pérez, "An explainable approach based on emotion and sentiment features for detecting people with mental disorders on social networks," *Appl. Sci.*, vol. 11, no. 22, p. 10932, Nov. 2021.
[58] G. I. Pérez-Landa, O. Loyola-González, and M. A. Medina-Pérez, "An explainable artificial intelligence model for detecting xenophobic tweets," *Appl. Sci.*, vol. 11, no. 22, p. 10801, Nov. 2021.

**MIRYAM ELIZABETH VILLA-PÉREZ** received the degree in computer systems engineering and the M.Sc. degree in computer science from Tecnologico de Monterrey, Mexico, in 2016 and 2019, respectively, where she is currently pursuing the Ph.D. degree in computer science. Her research interests include pattern recognition, natural language processing, machine learning, and data science.

**LUIS A. TREJO** received the Ph.D. degree in computer science (parallel processing) from Université Claude-Bernard de Lyon, France, in 1993. He is currently a full-time Professor with the School of Science and Engineering, Tecnologico de Monterrey. Since 2015, he has been a member of CONACyT's National Research System, Level 1, and a member with the GIEE-ML (Machine Learning) Research Group, Tecnologico de Monterrey. His research interests include internetworking, the Internet of Things, information security, intrusion detection and prevention systems, machine learning, data science, and parallel processing.

**MAISHA BINTE MOIN** received the M.Sc. degree in computing science from the Faculty of Science, University of Alberta, Edmonton, AB, Canada, in 2022. Her current research interests include machine learning, natural language processing, and software engineering.

**ELENI STROULIA** (Member, IEEE) is currently a Professor with the Department of Computing Science, University of Alberta. Her research interests include addressing industry-driven problems, adopting AI, and machine-learning methods to improve or automate tasks.