*Expert Systems*

WILEY

Expert Systems

ORIGINAL ARTICLE OPEN ACCESS

# Machine Learning Algorithms to Address the Polarity and Stigma of Mental Health Disclosures on Instagram

Noemí Merayo[1] | Alba Ayuso-Lanchares[2] | Clara González-Sanguino[3]

[1]Signal and Communications Theory and Telematics Engineering, School of Telecommunications Engineering, Universidad de Valladolid, Valladolid, Spain | [2]Department of Pedagogy, Faculty of Medicine, Universidad de Valladolid, Valladolid, Spain | [3]Department of Psychology, Faculty Education and Social, Universidad de Valladolid, Valladolid, Spain

**Correspondence:** Noemí Merayo (noemer@tel.uva.es)

**Received:** 9 July 2024 | **Revised:** 16 December 2024 | **Accepted:** 23 December 2024

**Funding:** This work was supported by Universidad de Valladolid.

**Keywords:** Instagram | machine learning | mental health | natural language processing | sentiment analysis | social networks | stigma

## ABSTRACT

This research explores the social response to disclosures and conversations about mental health on social media, which is a pioneering and innovative approach. Unlike previous studies, which focused predominantly on psychopathological aspects, this study explores how communities react to conversations about mental health on Instagram, one of the favourite social media platforms among young people, breaking new ground not only in the Spanish context, but also on a global scale, filling a gap in international research. The study created a novel corpus by collecting and labelling comments on Instagram posts related to celebrity mental health disclosures, categorising them by polarity (positive, negative, neutral) and stigma. Additionally, the research implements machine learning algorithms to detect stigma and polarity in mental health disclosures on Instagram. While traditional techniques like Support Vector Machine (SVM) and RF (Random Forest) displayed decent performance with lower computational loads, advanced deep learning and BERT (Bidirectional Encoder Representation from Transformers) algorithms achieved outstanding results. In fact, BERT models achieve around 96% accuracy in polarity and stigma detection, while deep learning models achieve 80% for polarity and 87% for stigma, very high accuracy metrics. This research contributes significantly to understanding the impact of mental health discussions on social media, offering insights that can reduce stigma and raise awareness. Artificial intelligence can be used for more responsible use of social media and effective management of mental health problems in digital environments.

## 1 | Introduction

Social networks have become one of the most widespread communication channels today and allow for a constant flow of information that reflects the attitudes, trends and opinions of our society in real time. In fact, there are currently 4.76 billion social network users worldwide (Datareportal 2023), equivalent to approximately 60% of the world's population. Looking at social networks in terms of monthly active users, the latest data suggests that Facebook remains the world's number one social network with nearly 3000 million users. In this context, Instagram has also consolidated its position among the top social media platforms, ranking fourth with 2 million users behind Facebook, Youtube and Whatsapp, with an average time per user per month of 12 h. When examining social media preferences based on age and gender, individuals aged 16–24 and young women aged 25–34 prefer Instagram as their top social platform. Indeed, in January 2023, nearly two-thirds of Instagram's total audience were 34 years old or younger (51% of the total audience were between 13 and 17 years old; 33.7% between 18 and 24 years old, and 31.3% between 25 and 34 years old) (Statist 2023). Furthermore, some recent studies show that

Instagram is the SN most used by young people (Oden and Porter 2023).

Similarly, the significance of mental health in society has grown in recent years, as the World Health Organisation reports a global increase in mental health issues. In 2019, 970 millions of people worldwide were living with a mental disorder, with anxiety and depression being the most prevalent conditions (World Health Organization 2024). Moreover, an European study from October 2023 revealed that 46% of EU citizens had experienced an emotional or psychosocial issue in the last 12 months, such as feelings of depression or anxiety (European Commission 2023). When analysing the mental health landscape of younger populations, the situation becomes particularly concerning. According to UNICEF, in 2019, approximately one in seven adolescents globally, representing 166 million individuals (89 million boys and 77 million girls), were estimated to be affected by mental illness (UNICEF 2021). Half of all mental disorders in young people develop before age 14, and 75% by their mid-twenties. In fact, 3% of 12- to 17-year-olds affirm experiencing depression and 32% report anxiety. This issue extends to young adults, with 33.7% of those aged 18 to 25 reporting some form of mental illness. Particularly, in 2023, approximately 1 in 5 children and young people aged 8 to 25 in the UK were estimated to have a probable mental disorder, with prevalences of 20.3% in the 8 to 16 age group, 23.3% among those aged 17 to 19, and 21.7% in the 20 to 25 age group (NHS Digital 2023). In the same year (2023), 20.17% of 12–17 year olds in the USA reported at least one depressive episode during that year (Mental Health America 2023). A recent study conducted in Spain claims that 42.1% of individuals have experienced depression, and 14.5% have had suicidal ideation or attempted suicide, with an average age of diagnosis at 26 years (Fundación Mutua 2023).

However, despite the commonality and importance of mental health, it is possible to say that mental health stigma still exists in our society, and despite the progress still is an issue necessary to address (Gronholm and Thornicroft 2022). This construct refers to thoughts (beliefs, myths, attributions), emotions (such as reactions of fear or pity) and negative behaviours (usually discrimination, or desire to distance oneself), shared by the society towards a specific group, in this case people with mental health problems (Corrigan and Watson 2002). In Spain, stigma towards mental health is present in the general population (González Sanguino et al. 2023), and certain stigmatising beliefs and attributions have been similarly found in adolescents (González Sanguino et al. 2024).

On the face of it, leveraging social media to talk about mental health should promote acceptance and reduce stigma, as some studies show how high-impact posts by celebrities can promote awareness and help-seeking by reaching large audiences (Gronholm and Thornicroft 2022; Jain, Pandey, and Roy 2017; Lee 2019; Lee, Yuan, and Wohn 2021). However, due to the sheer volume and immediacy of opinions expressed on these platforms, it has become a complex phenomenon. As a result, it is unclear whether these social media posts are actually fostering acceptance and good quality knowledge or inadvertently perpetuating further stigmatisation and/or trivialization of mental health (Pavlova and Berkers 2022; Robinson et al. 2018). In this social environment of widespread adoption of social networks, together with the constant increase of information and opinions in real time, the application of automated techniques becomes highly relevant. Thus, Artificial Intelligence (AI) allows these actions to be carried out jointly, specifically the branch of Natural Language Processing (NLP) (Mäntylä, Graziotin, and Kuutila 2018) through what is known as Sentiment Analysis. Sentiment analysis combines natural language processing and computational linguistics to explore the meanings of words and their context, with the aim of understanding the underlying emotional tones. This technology can be applied to discern emotional reactions in comments posted on social networks, enabling real-time trend tracking, understanding current or future behaviours. However, the linguistic nature of comments posted on social media platforms exhibits significant disparities compared to conventional language use (Martínez-Cámara et al. 2014), often posing substantial challenges. These unique features, including brief messages, missing context, grammatical errors, and a casual writing style, complicate the application of effective sentiment analysis techniques. Furthermore, in sentiment analysis, two main approaches are primarily used: lexicon-based techniques and supervised learning-based techniques. Lexicon-based techniques require dictionaries where words are labelled with emotional responses, such as polarity or associated emotions. However, in Spanish, the number of dictionaries is limited, and it is necessary to develop specific dictionaries for each area of application (Redondo et al. 2007). Additionally, this method should be supplemented with techniques for identifying negation or language ambiguity, which adds complexity to these strategies, especially on social networks (Taboada et al. 2011). In contrast, supervised learning techniques require labelled corpora, which are examples of opinions or comments that have been previously annotated. This allows machine learning algorithms to learn from this data and make efficient predictions (Arco et al. 2021; Shah et al. 2023). This method provides the benefit of allowing various machine learning algorithms to be flexibly used on the same dataset. Furthermore, certain studies, like the one mentioned in (Srivastava, Bharti, and Verma 2022), have shown that supervised methods can outperform lexical-based techniques in specific scenarios, so our research will follow the machine learning approach.

Despite the multiple advantages that AI could provide, research integrating AI to address mental health and stigma in social networks is limited. NLP brings important benefits to mental health research by autonomously uncovering important insights and patterns in the data that might go unnoticed or unavailable to mental health experts, and might even be overlooked in manual reviews. In addition, expressions related to mental health often exhibit greater emotional complexity and subtlety, as individuals in these contexts tend to use ambivalent or metaphorical language to describe their emotional state. This characteristic requires the development of more advanced NLP models to accurately interpret these expressions. In addition, discussions about mental health are often marked by social stigma. This influence can lead people to hide their true feelings or use language that does not accurately reflect their experience, adding an additional level of complexity to the analysis of these conversations. Even more, the existing

studies that assess stigma and polarity in social networks comments relate to specific diagnoses such as schizophrenia, bipolar disorder and Alzheimer's disease (Budenz et al. 2019; Jilka et al. 2022; Oscar et al. 2017). Other studies have focused on reactions to antipsychotic medication (Mon et al. 2021), and other studies are indirectly related to mental health issues such as obesity (Bograd, Chen, and Kavuluru 2022) or Covid-19 (Xue et al. 2020). In this way, other research (Budenz et al. 2019) uses supervised learning to analyse the existence of stigma in social networks towards bipolar disorder, one of the most stigmatised mental illnesses. Specifically, the study allows us to characterise stigma from supportive messages about bipolar disorder and their repercussion and impact on Twitter. Regarding psychosis, the conducted research of (Jilka et al. 2022) proposes to identify stigmatising tweets on Twitter related to schizophrenia by applying algorithms such as SVM (Support Vector Machine), RF (Random Forest) among others. Finally, the authors of (Oscar et al. 2017) use machine learning to analyse the existence of stigma on Twitter regarding Alzheimer. All these studies have been carried out on the Twitter social network and only one of them was developed in the Spanish context (Mon et al. 2021). In that Spanish research, the authors aimed to investigate Twitter conversations concerning antipsychotic drugs in order to gain insights into public reactions and identify the most frequently discussed areas of clinical interest related to their usage. Consequently, this context depicts a scenario where the mental health and social networks are revealed as an exceptional environment, in terms of their relationship and impact, and allows us to understand public opinion of mental health, the emotional responses it elicits and the presence of stigma. Thus, the main contributions of our research are: (a) Design a novel corpus labelled with polarity (positive, negative, neutral) and stigma from Instagram post comments on celebrity mental health disclosures. This dataset can be accessed on GitHub (Merayo, Ayuso, and González-Sanguino 2023), allowing researchers to use it. (b) Modelling machine learning algorithms to predict polarity and stigma in social networks in case of disclosure of mental health problems, specifically anxiety or depression. Thus, this research is innovative on several levels. It proposes machine learning analysis of mental health on Instagram for which there is barely any precedent, since Studies on AI have mainly concentrated on Twitter. Secondly, this is a novel approach, as previous research has focused primarily on identifying psychopathology rather than examining the community response to it (Ahmed et al. 2022; Birnbaum et al. 2017; Fodeh et al. 2019; Guntuku et al. 2017; Joshi and Kanoongo 2022; Lejeune et al. 2022). Thirdly, this research holds significant social relevance as it examines the emerging trend of public reactions to celebrities discussing their mental health issues, influencing millions of individuals, rather than focusing on hashtags or general comments on a topic. This can facilitate more effective campaigns and actions, leading to increased awareness and favourable effect on society as a whole, along with particular groups, including mental health professionals, and organisations, promoting responsible use of social media and making informed decisions to address mental health on these platforms.

This document is structured as follows. Section 2 describes the methodology for creating the dataset from Instagram social media posts. Section 3 details the classification models implemented. Section 4 reveals the results of these models. Finally, Section 5 explains the main conclusions.

## 2 | Design of the Mental Health Corpus in Social Networks

### 2.1 | Selection of Posts on Instagram

A search was conducted for primary posts (made directly by the author) containing disclosures or conversations about their mental health problems by Spanish influencers on Instagram (with over 100,000 followers). To carry out this process, we have searched for publications from different profiles of the people with the most followers in Spain, as well as reviewing press and television news that usually announce this type of publication. This search covered publications from September 2020 to December 2022. After an initial review, most of the posts were made by women, and as we were unable to have a gender balance in the posts, we decided to include the male gender in the Instagram posts as an exclusion criterion to avoid possible bias in the analysis. We found around 20 posts by high-impact female influencers on Instagram and selected the 10 with more responses or comments. All posts had a similar format, with one or more photos accompanied by text talking about mental health problems, such as depression or anxiety. A couple of posts announced that they were withdrawing from the social network due to their mental health problems, and in another couple of posts, the image showed the person crying. Unusually, one of the posts consisted of a promotional video in which an influencer talked about her mental health issues to promote a product. Once the Instagram posts were selected, all comments in response to them were collected using IGCommentExport ("One Click Comment Extractor for IG") (Chrome Web Store 2023), a tool to export Instagram comments to CSV (Comma Separated Values) format. A total of $N = 21{,}151$ comments were collected. Regarding ethical considerations and data privacy, the study has been approved by the ethics and deontology committee of the University of Valladolid (PI 23-3365) and we anonymised all Instagram comments (discarding @mentions, usernames and URLs). The final selection of Instagram posts, together with the name of the influencers, the number of followers and the responses associated with each post are in the Supporting Information.

### 2.2 | Description of the Labels in the Corpus: Polarity and Stigma

Following a manual observation of the dataset, and in line with previous literature on manual labelling of comments (Budenz et al. 2019; Mon et al. 2021; Bograd, Chen, and Kavuluru 2022; Tomar, Mathur, and Suman 2022; Delanys et al. 2020) we set up guidelines for the different labelling categories: polarity and stigma. Regarding the polarity associated with a comment, it consists of giving a positive, negative or neutral/undefined value to the comments in response to the disclosure or description of the symptomatology in the post. Positive polarity reflects understanding, encouragement or

even admiration of the post. For example, "Cheer up, we love you". Negative polarity is assigned when the person expresses negative opinions, usually by questioning the post with ironic, sarcastic or even derisive and derogatory comments. For example, "how you show that you don't know what depression or anxiety is, shame on you!". Neutral or undefined polarity is assigned in cases where no clear opinion is detected or can be interpreted in both directions. For example, "take medication, it will help you" "and your partner?"

About stigmatisation, stigmatising responses to comments are behaviours in which negative beliefs and emotions towards mental health problems are expressed. Stigma manifests in a variety of forms including rejection and anger against the person, which may extend to contempt or mockery, belittling their problem. For example, "What a desire to draw attention to yourself"; "you're so inconsistent and seeking the limelight". Because socially we know that "stigma is wrong" many rejection comments are made in an ironic or sarcastic way. For example, and how do you write on insta?. Additionally, anger is shown by arguing that such posts "trivialise or commercialise" mental health. For example, "don't come and tell me your false stories of overcoming, without even knowing what it is to work...". Other times the stigma manifests itself as pity or sorrow for the person. For example, "It breaks my heart".

### 2.3 | Process of Categorising the Corpus

The labelling process was divided into three phases: an initial phase with a pilot corpus ($N = 787$ comments), a second phase focused on the development of the corpus with all the comments of the selected posts ($N = 21{,}151$) and a third phase with the final corpus ($N = 2287$). The same methodology was followed in the first two phases: once the comments were collected, the corpus was cleaned, and then two independent experts were responsible for labelling each category. A third expert then reviewed the comments to resolve discrepancies. In the third and final phase, a final corpus is built from the large corpus to apply machine learning algorithms ($N = 2287$).

In this way, the labelling process in the pilot corpus represented a first stage carried out on a random subset of comments ($N = 787$, including emoticons). The process was divided into the following stages:

(a). Initial data cleaning: comments in other languages, with acronyms only (e.g., "TQ", "I love you" in English) and those lacking coherence, e.g., "cui-de-se-BR" or "gusta ver tu" ("like see you") in English, were deleted to maintain the sample's relevance. Additionally, comments labelling other people who have replied to the same post have been removed, except when the author of the post is labelled and relevant information is give (e.g., "@dulceida I hope all is well"). This results in a final sample of $N = 573$ comments.

(b). Handling Emoticons: emoticons have been excluded to focus solely on the linguistic effects.

(c). Expert labelling: two specialists independently labelled the clean sample without knowing each other's assessments, while a third expert examined the inconsistencies that emerged in the clean sample.

In this pilot corpus, we identified discrepancies in only 2.43% ($N = 14$) of the labelled comments, which showed a very satisfactory inter-rate reliability among experts who categorised (Hallgren 2012). These discrepancies occurred predominantly in the neutral polarity categories, as some comments contained messages with both positive and negative polarities ("What a pity! I'm so sorry about what happened to you, lots of encouragement"). The third reviewer found that the message of sympathy prevailed in these cases, so it was categorised as positive polarity. Satisfactory results from this initial process (pilot corpus) provide a solid basis for replicating the results in a larger sample, ensuring consistency of labelling and maintaining data quality for future analysis. Regarding the second phase, which corresponds to the whole corpus, the corpus was labelled with all comments ($N = 21{,}151$). The same procedure as in the pilot study was followed:

a. Initial data cleaning: this process reduced the whole corpus to a final sample of $N = 15{,}213$ comments. To guarantee the suitability of the sample, comments that consisted solely of acronyms (e.g., LOL), that were not written in English or that were incomprehensible were discarded in this process. In addition, comments that only served to name other users without further content were also removed.

b. Handling Emoticons: emoticons have been removed.

c. Expert labelling: the final sample was labelled by two experts separately, while a third expert evaluated any inconsistencies. If the two initial experts could not agree on the assigned label, a third reviewer was required. If this third reviewer also could not resolve the tie, the comment was excluded from the corpus to avoid possible errors. To carry out the labelling process, the experts were psychologists or trained persons who used a labelling guide, elaborated with examples and descriptions of the different categories of our corpus (polarity, stigma). This labelling guide is also freely accessible and available in a Github repository (Merayo, Ayuso, and González-Sanguino 2023).

The percentage of discrepancy in this second phase was around 2.3% (489 comments). The greatest disparity was observed between neutral and positive polarity in those comments where advice was given ("Ayyy, take as much time as you need, health comes first"). In these cases, the third reviewer categorise these comments with neutral polarity. Finally, the third phase was associated with the final corpus. Once the categorisation of the full corpus was completed, a thorough selection process was undertaken to develop a representative corpus that would be appropriate for our IA algorithms. A key consideration in data corpora is class balance; when addressing a classification problem, an imbalance where one class has significantly more data than another can lead classification models to favour predictions for the majority class. This imbalance adversely affects the algorithm's performance and predictive capability. As a result, the original set of 15,213 comments from the extensive corpus was systematically condensed into 2287 comments in the following steps:

a. Removal of redundant comments: messages with identical content were removed, keeping only a single example for each recurring topic.

b. Distribution: we evaluated the distribution of comments among different categories to ensure a more equitable representation of polarity and stigma in the dataset, ensuring that models would make accurate predictions in all classes. Although there are more comments of positive polarity than negative or neutral, and more non-stigmatising than stigmatising comments, the distribution is significantly more balanced compared to the whole corpus.

c. Comments randomly chosen: after determining the target distribution percentages, comments from all posts were randomly picked to be part of the final dataset.

## 2.4 | Corpus Statistics

Table 1 displays the frequency-based descriptive statistics for each category of the corpus, and Figure 1 presents the descriptive statistics by percentage. The most predominant polarity is positive (66.7%), as well as non-stigmatising responses (80.1%). Besides, it is observed that there are more negative polarity comments than stigmatising ones. This is because many comments that include disclosures of mental health problems are not stigmatising but their emotional polarity is negative (e.g., "I cried with you when I saw your tears... even though I don't know you in person, tell you that my hand will always be with you"). Likewise, there are also neutral messages that include advice or reactions without a specific emotional tone (e.g., "Good bless you"; "real life") that also do not meet the requirements to be categorised as stigmatising.

**TABLE 1** | Descriptive statistics by frequencies.

| Variable | Frequency |
|---|---|
| Polarity | |
| Negative (P) | 588 |
| Positive (N) | 1526 |
| Neutral (NEU) | 173 |
| Total | 2287 |
| Estigma | |
| No | 1833 |
| Yes | 454 |
| Total | 2287 |

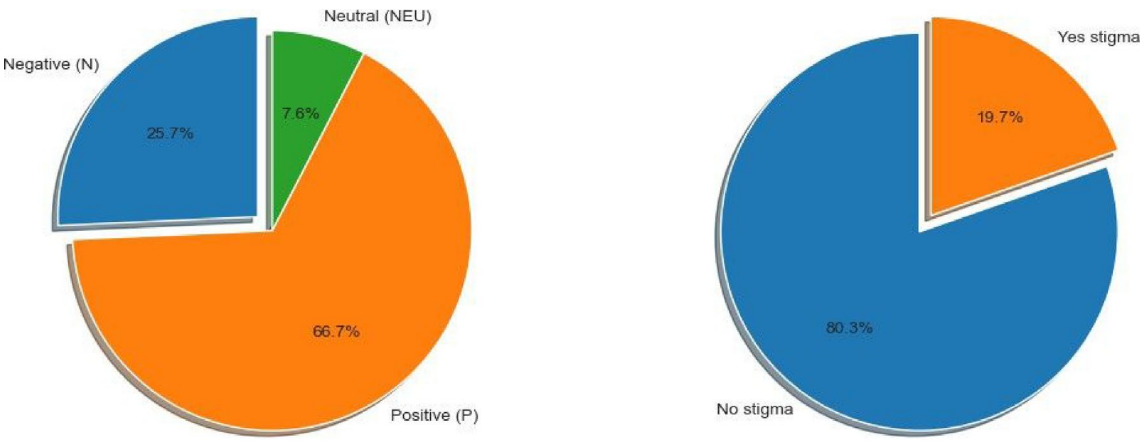
## 3 | Classification Models

### 3.1 | Support Vector Machine

The Support Vector Machines (SVM) algorithm is used in both classification and regression problems. Its goal is to find an optimal separation hyperplane that maximises the distance between data classes, in our case polarity and stigma categories, in a feature space (Chollet 2021; Noble 2006). The most important configurable hyperparameters in SVM include the Kernel, which determines the type of transformation used to separate data in a high-dimensional space, and the regularisation parameter C, which controls the trade-off between achieving a wider margin and minimising errors in classification. Proper tuning of these hyperparameters is essential to achieve optimal performance in SVM models and ensure accurate classification of the data.

### 3.2 | Random Forest

The Random Forest (RF) algorithm relies on building multiple decision trees during training and combining their results to make more accurate and robust decisions (Probst, Wright, and Boulesteix 2019). Each tree in the forest is trained on a random sample of the data and uses a random selection of features to make decisions. Typical configurable hyperparameters in Random Forest include the number of trees and the maximum number of features to consider in each split (max_features). Proper tuning of these hyperparameters can significantly influence the performance and generalisability of the model.




(a) Polarity (b) Stigma

**FIGURE 1** | Descriptive statistics by percentage of the dataset for polarity and stigma categories (a) Polarity (b) Stigma.

### 3.3 | Hybrid CNN-LSTM Model

The proposed hybrid deep learning architecture consists of the following layers (Figure 2): 1 Embedding layer ⟶ 1 Dimension convolutional layer (Conv1D) ⟶ 1 MaxPooling Layer ⟶ 1 LSTM (Long Short-Term Memory) layer ⟶ 1 Dense Layer. Our model benefits from the potential offered by the combination of recurrent (Recurrent Neural Network, RRN) and convolutional (CNN, Convolutional Neural Network) layers. As the convolutional layer assumes certain tasks, it reduces the processing load on the LSTM (recurrent layer), improving the effectiveness of that layer. This model will classify the mental health comments into three polarities (P, N, NEU) and in two stigma categories (Yes, No). In the following, each layer of the model will be described in detail:

1. Embedding layer: This layer transforms texts, Instagram comments in our case, into numerical vectors so that they can be interpreted by neural networks. This process is carried out using word embedding, in which each word is depicted as a vector. The goal is to assign similar values to words that share a certain semantic relationship. Thus, two techniques can be applied: learning word embeddings in conjunction with the problem to be solved (using the problem corpus) or loading embedding vectors from precomputed databases of word embeddings. The first option was chosen because pre-calculated dictionaries are designed in a general way and their effectiveness depends to a large extent on their similarity to the words in our specific corpus. In contrast, learning directly from our corpus will be optimal as it provides a relevant source of information. The embedding layer functions like a dictionary that maps integer indices, which correspond to specific words, to dense vectors. It accepts integers as input, searches them in its internal dictionary, and returns the associated vectors, much like a dictionary lookup. Initially, when the embedding layer is set up, its internal word vectors, or weights, are randomised. Throughout the training process, these vectors are adjusted using backpropagation, resulting in a distinct structure that is tailored to the specific task by the end of training. Therefore, the embedding layer takes a two-dimensional tensor as input and returns a three-dimensional tensor that can be further processed by a convolutional layer.
2. 1D (Dimension) Convolution Layer (Conv1D): This layer employs filters on the data to identify local features within the input. Its function is to identify important patterns using convolution operations, reducing the workload for the subsequent RNN layer. In essence, it streamlines the processing for the RNN by removing intermediate stages through text pattern detection. In addition, the ReLU (Rectified Linear Unit) function shall be used as an activation function, which shall be applied after convolution. Additional key parameters include the number of filters (with each filter designed to capture a specific pattern in the input data) and the filter size (kernel size), which determines the dimensions of the filters based on the length of the window.
3. MaxPooling layer: This layer is used after the convolution layer to reduce the dimensionality of the features extracted by the previous layers of the network, while preserving the most relevant information. The MaxPooling layer transforms a data matrix into a smaller matrix, retaining the key elements of the original.
4. LSTM layer: This RNN layer is employed to identify long-range patterns in the input data. In particular, LSTM layers improve the functionality of traditional recurrent networks by combining both long-term and short-term memory capabilities. The parameters to be adjusted are the number of neurons, a dropout rate and a recurrent dropout rate. Both parameters are adjusted to prevent overfitting and enhance the model's generalisation during training. These parameters indicate the proportion of neuron units that are randomly "deactivated" during each training step, to prevent them from becoming too dependent on neighbouring neurons.
5. Dense Layer: This layer takes the features learned by the previous layers and produces the final output of the network, crucial for producing the final model predictions. Since we use the model for classification, the number of neurons in the output layer corresponds to the categories into which we want to classify the data, that is in multi-class classification problems there will be one neuron per class. Therefore, in our case we will put three neurons to categorise polarity (P, N, NEU) or two neurons to categorise stigma (Yes, No). Finally, the softmax activation function will be utilised to transform

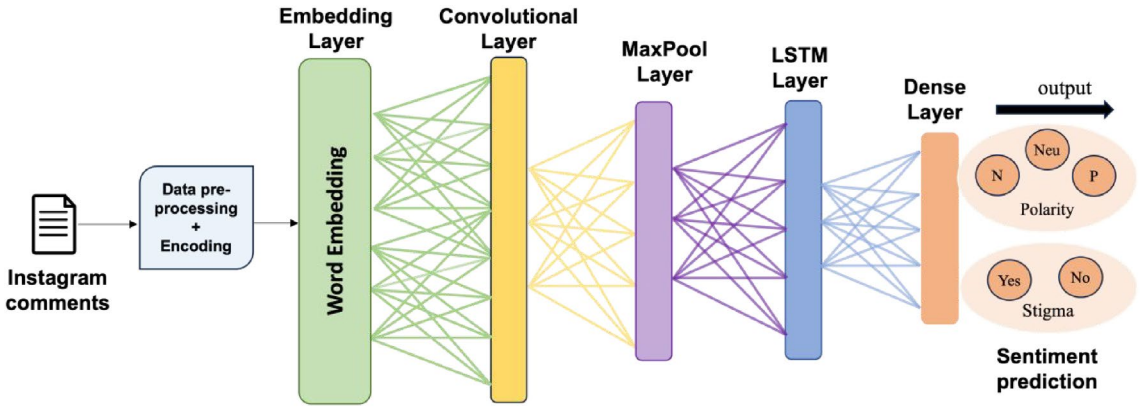



**FIGURE 2** | Configuration of layers in the proposed hybrid CNN-LSTM network model.

the outputs into a probability distribution, as it is frequently used in multi-class classification tasks.

Finally, in the context of classification problems, there are other crucial hyperparameters influencing model convergence, including optimizers and loss functions. In our particular scenario, we chose the categorical cross-entropy loss function, which is a commonly used option for classification tasks, especially when dealing with three or more labels as in our scenario. Additionally, the ADAM optimizer was selected due to its versatility, strong performance, and widespread utilisation in similar problem domains. As mentioned above, the word embedding process can be carried out in two ways: learning word embeddings jointly with the problem you want to solve (using the corpus of the problem) or by loading embedding vectors from precomputed databases. The first option was chosen because of the advantages that this technique provides. To accomplish this, a vocabulary is generated from the corpus by employing the Information Gain (IG) technique, which emphasises terms that occur most frequently. IG is favoured over absolute frequency since it evaluates how often a word appears in a particular class compared to its occurrence in other classes, while absolute frequency only quantifies total occurrences without considering class differences. To determine the IG of a word, we first compute its entropy (Larose and Larose 2014; Witten, Frank, and Hall 2011), shown in 1:

$$H = \sum_{i=1}^{N-1} p_i \log_i(p_i). \tag{1}$$

where $p_i$ is the probability that the word $w_i$ appears in the class $c_i$. Then, we calculate the IG following 2:

$$\mathrm{IG}(C, X) = H(C) - H(C, X). \tag{2}$$

In this equation, $C$ represents the set of classes and $X$ is the subset of texts in which the word $w_i$ is found. To calculate $H(C)$, the probabilities of each category in the corpus are determined, while to calculate $H(C, X)$ the likelihood of a word occurring or not occurring in the corpus needs to be calculated, along with its probabilities of occurrence and non-occurrence in each category.

### 3.4 | BERT Model

BERT is a transformer-based language model that understands the context of words in both directions, enhancing tasks like natural language processing and comprehension. Employing BERT for sentiment analysis requires initially training the model on a substantial dataset before fine-tuning it on a dedicated dataset. This process enables the model to gain a broad understanding of language, which it can then refine to capture the nuances of sentiment analysis within a specific field, such as mental health in social media contexts. In our case, we have used a pre-trained linguistic model for social networking text in Spanish, called RoBERTuito, trained following the RoBERTa guidelines on 500 million tweets (Pérez et al. 2021). This model surpasses other pre-trained language models for Spanish. Moreover, the Transformers library offers the Trainer class to fine-tune any of the pre-trained models on a particular dataset. This approach allows us to identify the best training parameters, such as the learning rate (which accelerates model convergence), batch size (the number of samples before updating weights), and epochs (which indicate the number of iterations over the training dataset).

## 4 | Experiments and Results

This section describes the main results of the experiments. To identify the ideal configuration, we will search for the most suitable hyperparameter values for each model in order to optimise various performance metrics. Although accuracy reflects the overall proportion of instances that the model has correctly classified, it is essential to consider additional metrics that assess the model's performance for each individual class. Metrics like precision, recall, and F1-score are particularly valuable when dealing with imbalanced datasets. A key aspect of model evaluation is having two separate datasets: the training set, used for training the model, and the validation set, used for evaluating its performance. In our case this has been split in a typical ratio of 70%–30%, respectively (Oneiros 2023). On the other hand, we used the cross-validation technique to identify the optimal parameter sets for the model being trained. Specifically, We utilised k-fold cross-validation, a method that entails performing k iterations in which the model is trained and assessed k times. Additionally, we implemented the EarlyStopping technique to maintain the model's generalisation ability. This method halts training once the validation loss reaches its minimum, preventing further overfitting. We implemented our classification models in Python (version 3.1) using Keras (Oneiros 2023) and TensorFlow (Google Brain Team 2023), and executed all models on the Google Colab platform.

### 4.1 | Data Pre-Processing and Encoding

Text pre-processing consists of cleaning and preparing textual data to obtain a semantically richer representation that facilitates its computational representation. Thus, the following series of pre-processing techniques have been implemented:

- To convert to lower case: to reduce duplicate words.
- To eliminate mentions (), hashtags (#), and URLs: as they do not contribute any valuable information.
- To delete punctuation marks.
- To minimise the repetition of characters: for example, change “Siiiiii” to “Si” (“yes” in English).
- To standardise slang/jargon: for example, change “tb” to “también” (“furthermore” in English).

The following step involves narrowing down the vocabulary used by the classification models (feature reduction) by applying the techniques outlined below:

- Remove stopwords: words that carry no significant meaning on their own, such as articles, adverbs, prepositions,

conjunctions, and certain verbs. They are usually very frequent words in natural language and depend on the language.

- Apply stemming: a text normalisation technique that reduces words to their root. This technique removes affixes from words, which can lead to invalid words. For instance, the Spanish words "pensando" (meaning "thinking" in English) and "pensamiento" (meaning "thought" in English) will be shortened to "pensa." We will employ the SnowballStemmer from NLTK (Python Software Foundation 2021) for this purpose in Spanish.

In addition, the BERT model (RoBERTuito) (Pérez et al. 2021) includes a specific data pre-processing consisting of: character repetitions are capped at three, usernames are replaced with a designated token, hashtags are substituted with another token, and emojis are converted into their textual descriptions using a specialised library. However, RoBERTuito was evaluated using both data preprocessing methods, and the performance results were quite similar. The subsequent step involves tokenization, a fundamental step in text processing, that consists of breaking text into discrete units called "tokens", in our case words. Tokens provide discrete units that computers can work with to understand and parse text more effectively. In this case we use the TweetTokenizer tokenizer (NLTK Project 2023) for the SVM, RF and hybrid CNN-LSTM models. On the contrary, BERT algorithms use their own tokenizer. The goal is to find the most meaningful but smallest representation. The next step is to transform the texts into number (feature extraction process), since machine learning models and their inputs have to be numbers. In our case we have to transform two parts: on the one hand the tokenised and normalised messages (Instagram posts), and on the other hand the labels that correspond to the categories (polarity and stigma in this case). To convert the messages into numerical format, a dictionary has been established where each word is assigned a specific index vector (as described in the previous section on word embedding). For the labels, One Hot Encoding was used, which encodes various classes as a matrix. In this matrix, a "1" is placed in the column corresponding to the class of the text (Instagram message), while "0" is used for all other classes. Therefore, for the polarities (P, N, and NEU), we will create a matrix with three columns. A similar approach is used for categorising stigma labels (Yes and No), resulting in a two-column matrix. For the SVM and RF models, instead of using individual columns for each variable with binary values (0 or 1), we use a single global variable to represent one output, as these models generate only a single result. Here, polarity is indicated as P, N, or NEU, unlike the earlier binary encoding of "0" or "1".

## 4.2 | Polarity Results in the Mental Health Corpus

### 4.2.1 | Results SVM Model

The optimal hyperparameters in SVM will be searched in the next order: kernel type and regularisation parameter ($C$). The optimisation process begins with a kernel scan, which reveals that the RBF kernel achieves the highest accuracy at 63%. In contrast, the Poly and Sigmoid kernel types yield lower results of 62%. We then proceeded to optimise the $C$ regularisation parameter, where the best value is 1.4, achieving an accuracy of 65%. However, as we increase the value of $C$, starting at 10, the model exhibits the worst accuracies, reaching values of 59% and 57% for 100 and 200, respectively. In contrast, if we continue decreasing the value of $C$ below 1, the model's accuracies remain more stable but lower, reaching levels of 63% from $C=0.01$ to $C=0.001$.

In summary, the values of the hyperparameters that optimise the SVM model are RBF for the kernel type and 1.4 for $C$ (regularisation parameter), reaching a final accuracy of 65%. In addition to accuracy, it is important to evaluate the model in each class separately through other metrics such as precision, recall and F1 score (Table 2). The results show that the P class is the best predicted, reaching a precision of 68%. In contrast, the NEU and N classes show lower performance in all metrics.

### 4.2.2 | Results of the RF Model

In the case of RF, the hyperparameter search will focus on determining the ideal number of trees to employ. Additionally, RF selects random feature subsets to optimise splits, making the hyperparameter max_features crucial in deciding how many features should be considered. Therefore, the hyperparameters will be searched in the following order: max_features and the number of trees. First, the accuracy of the model was evaluated with different values of max_features. Using Log2, the accuracy achieved was 66%. Using the square root (Sqrt), the accuracy improved to 69%, so Sqrt was selected. Then, we proceed to find the best number of decision trees to address the problem. Figure 3 shows that the optimal value is 600 decision trees, obtaining an accuracy of 71%.

In addition to global accuracy, Table 3 shows the results of precision, recall and F1-score for each class separately. It is observed that the P class is the class best predicted by the RF model, as all metrics show very good results. As for the N class, a precision close to 60% is achieved (higher than the SVM model). Finally, the NEU class has a low precision of 40%, which is reasonable since it is quite difficult to detect neutral comments when we express opinions on social networks. As expected, the RF classification model improves SVM performance for all polarities.

### 4.2.3 | Results of the Hybrid CNN-LSTM Model

To train and evaluate the performance of the model, different types of tests have been performed to see the variations in the accuracy metrics involved. These tests are: adjustment of the number of filters and neurons, the dropout rates and the learning rate for the Adam optimizer. Next, tests were made with the reduction of the total number of unique words chosen

**TABLE 2** | Summary of the SVM model results for precision, recall and F1-score metrics considering three polarity classes (P, N, NEU).

| Label | Precision | Recall | F1-score |
|---|---|---|---|
| P | 68% | 90% | 77% |
| N | 42% | 19% | 26% |
| NEU | 50% | 2% | 4% |

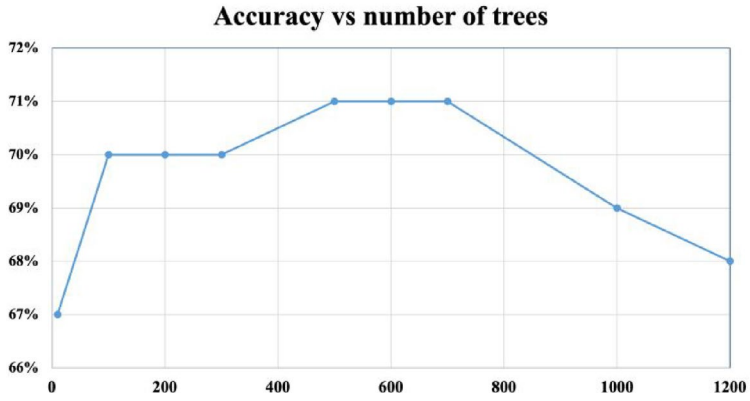



**FIGURE 3** | Optimisation of the number of trees.

**TABLE 3** | Summary of the RF model results for precision, recall and F1-score metrics considering three polarity classes (P, N, NEU).

| Label | Precision | Recall | F1-score |
|---|---|---|---|
| P | 74% | 92% | 82% |
| N | 59% | 37% | 46% |
| NEU | 40% | 8% | 14% |

from the corpus and finally an adjustment of the batch size parameter. All this to adjust the hyperparameters and avoid overfitting the developed model. The testing phase was carried out using several examples of performance metrics: precision, recall and F1-score.

As a first stage of the training phase, the training process will be repeated by varying the number of neurons and filters in the layers to find the combination that gives the best accuracy results, the values of which are shown in Table 4. For these tests, the kernel size has been set to 8, the dropout parameter to 0.2 and the recurrent parameter to 0.3. The results in Table 4 show that the model is quite sensitive to variations in these hyperparameters, achieving a higher accuracy rate for a number of 180 in the convolution layer and 256 neurons in the LSTM layer.

The second step in training the model consists of varying the dropout rates (dropout and recurrent dropout rate) of the LSTM layer in the range of 0.2 to 0.8. According to the Keras documentation (Oneiros 2023), the LSTM layers have two different types of dropout rates, which are represented as a floating point number between 0 and 1. In addition, for these tests, the kernel size was set to 8, the number of filters in the convolutional layer to 180 and the number of neurons in the LSTM layer to 256. Table 5 shows that varying the values generates significant changes in the accuracy of the model, and the best performance, 85.02%, is achieved with values 0.2 and 0.3, for the dropout and recurrent dropout rate parameters respectively.

The next step in the model is a sweep for different values of the learning rate. This is a fundamental factor when training machine learning models, since it will determine the degree of magnitude with which adjustments to the different parameters

**TABLE 4** | Results of the evaluation and optimisation of the hyperparameters number of neurons and filters in the hybrid CNN-LSTM model.

| Number filters convolutional layer | Number neurons LSMT layer | Cross-validation accuracy |
|---|---|---|
| 192 | 256 | 76.97% |
| 192 | 128 | 65.74% |
| 192 | 96 | 76.82% |
| 192 | 64 | 77.11% |
| 180 | 256 | 79.15% |
| 180 | 96 | 65.74% |
| 160 | 128 | 77.84% |
| 160 | 64 | 77.55% |
| 150 | 256 | 76.24% |
| 128 | 128 | 65.74% |
| 128 | 64 | 65.74% |
| 96 | 64 | 65.74% |
| 192 | 150 | 76.8% |

of the model will be made, which in turn affects the convergence of the model. There are two main reasons why it is interesting to control the learning rate: to control the speed of convergence and to overcome local minima that may occur. It can be seen in Table 6 that depending on the value, there is no considerable variation in the accuracy metric of the model when changing the value of this parameter, with the best value, 79.01% achieved for a learning rate of 0.01.

The next training step is to reduce the total number of unique words in the corpus used in the model to increase computational efficiency. However, there is a trade-off between the reduction of the vocabulary used and the possibility of losing important information, so it is necessary to perform the reductions in a stepwise manner (in this case in 100-word jumps). Using all words, the model achieved an accuracy of

**TABLE 5** | Results of the evaluation and optimisation of the hyperparameters dropout rate and recurrent dropout variation in the hybrid CNN-LSTM model.

| Dropout parameter | Recurrent parameter | Cross-validation accuracy |
|---|---|---|
| 0.2 | 0.3 | 79.15% |
| 0.3 | 0.3 | 79.01% |
| 0.4 | 0.3 | 77.11% |
| 0.5 | 0.3 | 65.74% |
| 0.6 | 0.3 | 65.74% |
| 0.7 | 0.3 | 65.74% |
| 0.8 | 0.3 | 65.74% |
| 0.2 | 0.4 | 75.95% |
| 0.2 | 0.5 | 76.24% |
| 0.2 | 0.6 | 77.70% |
| 0.2 | 0.7 | 78.57% |
| 0.2 | 0.8 | 77.70% |

**TABLE 6** | Results obtained for different values of learning rate.

| Learning rate | Cross-validation accuracy |
|---|---|
| 0.001 | 77.15% |
| 0.003 | 77.70% |
| 0.005 | 78.43% |
| 0.007 | 77.84% |
| 0.01 | 79.01% |

79.15%. When limiting the vocabulary to 2421 words, accuracy dropped to 78.28%. With a smaller vocabulary of 2321 words, accuracy improved to 80.61%, and with 2221 words it reached its best result of 81.05%. However, by reducing the vocabulary further to 2121 words, accuracy dropped to 77.70%. Thus, using a smaller number of words, discarding those of lesser relevance, leads to an increase in the accuracy of the model, but as the number continues to decrease, the accuracy starts to decrease, as there will come a point when important words will start to be discarded from the corpus. Therefore, the number of words that achieves the best accuracy, 81%, is 2221.

Finally, the batch size parameter defines the amount of data that the model has in each iteration, so it will be analysed how the modification of this value affects the model. The chosen values (32, 64, 128, 256, 512) are typical values used as standards in other models. The model achieved an accuracy of 80% with a batch size of 32, improving to 80.32% with a batch size of 64 and 80.61% with a batch size of 128. With a batch size of 256, the best accuracy of 81.05% was achieved, but when increased to 512, the accuracy dropped drastically to 65.74%. No major differences in accuracy results were observed; however, very high values can lead to excessive demands on memory and computing resources. Conversely, if very small values are used, more variability in the training data is observed. Therefore, the selected batch size was 256. The final parameter configuration that has allowed the model to be optimised is shown in Table 7 as a summary together with the selected values.

**TABLE 7** | Summary of the final values for the optimised hybrid cnn-lstm model for polarity.

| Parameters | Values |
|---|---|
| Number of filters convolutional layer | 180 |
| Number of neurons LSTM layer | 256 |
| Dropout rate | 0.2 |
| Recurrent dropout rate | 0.3 |
| Learning rate | 0.01 |
| Vocabulary size | 2221 |
| Batch_size | 256 |

**TABLE 8** | Summary of the hybrid CNN-LSTM model results for accuracy, recall and F1-score metrics considering three polarity classes (P, N, NEU).

| Label | Precision | Recall | F1-score |
|---|---|---|---|
| P | 85% | 94% | 89% |
| N | 78% | 67% | 72% |
| NEU | 42% | 22% | 29% |

Although accuracy measures the percentage of total cases where the model has predicted correctly, in our case 81.05% for three classes (P, N, NEU), it is important to consider other metrics to test the model in each class separately (precision, recall and F1-score). To analyse the model's performance in detecting different classes, Table 8 presents the corresponding metrics for each class. This allows us to determine which classes the model excels at detecting and which ones require improvement. The results indicate that the P class demonstrates the best performance, followed by the N class, while the NEU class exhibits the lowest performance. It is observed that the model experiences the greatest degree of difficulty when classifying messages of NEU polarity, due to the possible ambiguity that these messages represent, since messages usually have an inherent polarity and are often not totally neutral. It may also be due to the imbalance between the classes of the corpus, since the number of messages of NEU polarity is much lower than the other polarities. In contrast, the P class reaches very high values of 85% in the precision and the N class values close to 80%, which is very important, as extreme polarities play a more fundamental role in the analysis of discourses or topics on social networks as opposed to the neutral class that hardly contributes significant information.

#### 4.2.4 | Results of the BERT Model: RoBERTuito

We tested the RoBERTuito model with the following parameter settings, that allows the best performance of accuracy metrics:

Learning_rate = $8.759 \times 10^{-5}$, train batch size = 16, eval_batch_size = 32, num_train_epochs = 10.

For this configuration we obtain very good results, achieving a global accuracy of 96%. If we analyse the results for each class, we can see in Table 9 a worse result in the NEU class with respect to the extreme polarities (P, N), although the performance of the NEU class is also relatively good. Better results can also be observed in all metrics for the P class versus the N class, with differences above 5%. Even so, the results of all quality metrics for both P and N polarities are extremely good with levels equal to or above 90%, indicating the good performance of this algorithm.

#### 4.2.5 | Comparison of Results for Polarity

Figure 4 shows the comparison of accuracy for each algorithm. The best performance is achieved by the RoBERTuito model (96%) followed by the hybrid CNN-LSTM model (80.2%). On the other hand, the worst performance is obtained by the SVM model (65%), as expected. Finally, the RF model gives relatively good results, around 71%.

**TABLE 9** | Summary of RoBERTuito results for accuracy, recall and F1-score metrics considering three polarity classes (P, N, NEU).

| Label | Precision | Recall | F1-score |
|---|---|---|---|
| P | 95.7% | 97.5% | 96.5% |
| N | 90% | 91.8% | 90.8% |
| NEU | 83% | 61% | 69% |

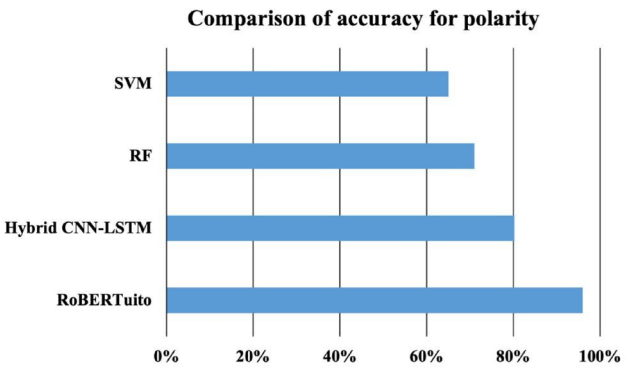



**FIGURE 4** | Comparison of all algorithms (SVM, RF, RoBERTuito, Hybrid CNN-LSTM) for the accuracy metric.

Although accuracy is a very important metric, it is important to consider other additional metrics, especially when the dataset may be unbalanced. Thus, Table 10 shows the values of the precision, recall and F1-score metrics. Once again, the best results for each metric are obtained by RoBERTuito followed by the hybrid CNN-LSTM algorithm, and the worst performance corresponds to SVM. When comparing the algorithmic performance across each class, a consistent pattern emerges. The P class demonstrates the highest predictability across all algorithms, closely followed by the N class. As anticipated, the NEU class exhibits the poorest performance, potentially attributed to the inherent ambiguity often associated with such messages, as they typically possess some degree of polarity. In addition, the low presence of this class in the data set contributes to making accurate prediction difficult.

Furthermore, the superior performance of the P class in contrast to the N class may be influenced by the larger representation of the P class within the dataset, approximately 67% of the total samples. Despite this unbalanced distribution in the dataset, the RoBERTuito and hybrid CNN-LSTM classification algorithms can predict both classes (P, N) with high quality metrics, especially the former. Indeed, this performance for the P and N classes is depicted in Figure 5, where it can be clearly observed that all algorithms perform much better for the P class than for the N class on all metrics. Moreover, RoBERTuito shows the best performance for all metrics in both classes, in contrast to RF and SVM, which obtain the worst results, especially for the N class. In this way, in contexts where the consequences of false positives or false negatives carry significant adverse implications, such as the identification of negative comments (expressing anger or hate) linked to mental health, maintaining high levels of precision is of utmost importance. Consequently, it becomes clear that both RoBERTuito and CNN-LSTM algorithms not only excel in this crucial metric in both classes (P, N), but also show strong performance in recall and F1-score metrics, which underlines the effectiveness of these algorithms in detecting extreme polarities.

Meanwhile, the confusion matrix offers insights into the model's performance. The matrix's diagonal showcases the correctly classified instances for each class, with the columns representing the model's predictions and the rows denoting the real values. This matrix facilitates the visualisation of the model's ability to differentiate between classes, helping to detect cases of confusion between them. Analysing the confusion matrix of the two best algorithms (Figure 6), RoBERTuito and the hybrid CNN-LSTM model, it is observed that RoBERTuito shows few confusions between the three classes, hence its good performance in all metrics (precision, recall, F1-score). In contrast, the

**TABLE 10** | Comparison of all algorithms for accuracy metrics, recall and F1 score considering three polarity classes.

| | N | | | P | | | NEU | | |
|---|---|---|---|---|---|---|---|---|---|
| Classification model | Precision | Recall | F1 | Precision | Recall | F1 | Precision | Recall | F1 |
| RoBERTuito | 90% | 91.8% | 90.8% | 95.7% | 97.5% | 96.5% | 83% | 61% | 69% |
| Hybrid CNN-LSTM | 78% | 67% | 72% | 85% | 94% | 89% | 42% | 22% | 29% |
| RF | 59% | 37% | 46% | 74% | 92% | 82% | 40% | 8% | 14% |
| SVM | 42% | 19% | 26% | 68% | 90% | 77% | 50% | 2% | 4% |

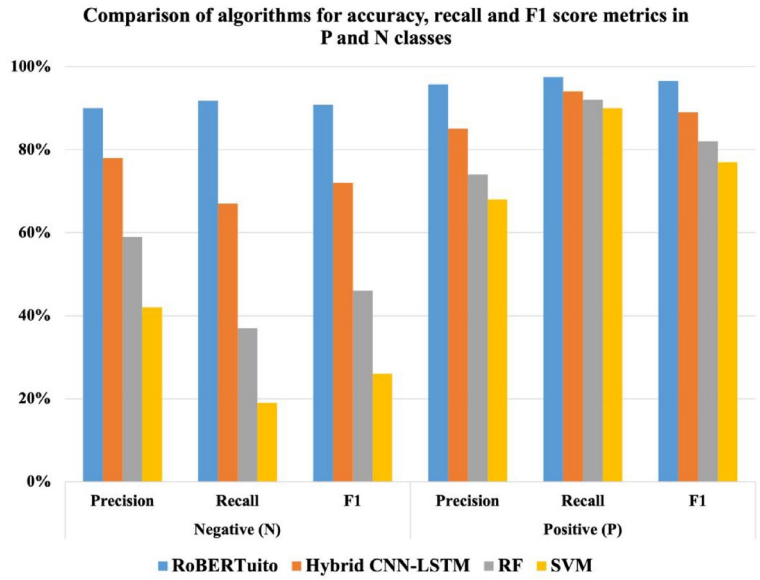
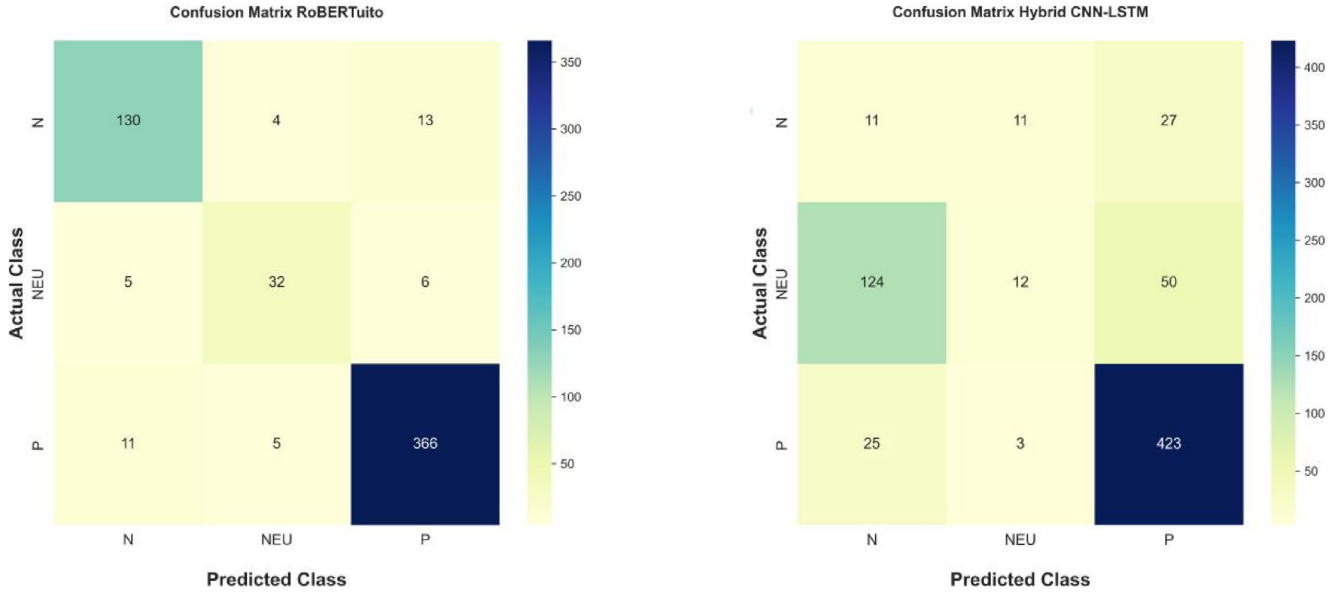



**FIGURE 5** | Comparison of all algorithms for precision, recall and F1-score metrics considering N and P polarities.




(a) RoBERTuito

(b) Hybrid CNN-LSTM

**FIGURE 6** | Confusion matrix for the three considered classes P, N, NEU (a) RoBERTuito (b) Hybrid CNN-LSTM model.

confusion matrix of the hybrid CNN-LSTM model shows that it fails more in the NEU class, as the percentage of this class in the dataset is very low. Another notable issue of the hybrid CNN-LSTM model is the fact that messages with negative polarity are being confused and predicted as positive, a fact that is also observed in other metrics such as precision. This may also be because the N class is also a minority in the dataset and certain patterns may be causing confusion.

## 4.3 | Stigma Results in the Mental Health Corpus

This section will show the results of the different algorithms, as well as a comparison of these algorithms, in predicting the detection of mental health stigma in social networks. It is important to highlight the importance of detecting and addressing stigma in mental health, not least to improve access to treatment, reduce suffering and isolation for those affected, and promote greater understanding and support in society at large. This contributes significantly to the health and well-being of people facing these challenges.

### 4.3.1 | Results of the SVM Model

The optimal hyperparameters of this model were searched for in the same order as for the previous polarity prediction case: search for the kernel type and then the regularisation parameter. The best values were obtained for Rbf and 5, respectively, reaching a global accuracy of 73%. If we analyse the performance of

the algorithm by separate classes (Table 11), we observe good performance for all metrics (accuracy, recall, F1-score) for the non-stigma class, while the algorithm fails more in predicting the existence of stigma.

#### 4.3.2 | Results of the RF Model

As in the case of polarity prediction, the optimal hyperparameters of this model will be searched for in the following order: max_features and the number of trees. The optimal values for both parameters are sqrt and 100, respectively, achieving an accuracy of 75%. Regarding the performance of separate classes (Table 12), it can be observed that RF improves the accuracy of the stigma class compared to SVM, and maintains comparable performance of the non-stigma class (No stigma), reaching high values of 75%.

#### 4.3.3 | Results of the Hybrid CNN-LSTM Model

The final parameter settings that allowed optimising the model for stigma prediction are shown in Table 13 together with the corresponding values. The process followed was the same as for polarity prediction, so we summarise all the steps in this table.

This configuration of the hybrid CNN-LSTM model achieves an overall accuracy of 87%. Analysing the results for both classes (No stigma, Yes stigma), Table 14 shows that both predictions improve on previous models (SVM, RF), especially the prediction of the stigma class, increasing performance by around 13% in the precision metric.

#### 4.3.4 | Results of the BERT Models: RoBERTuito

We tested the RoBERTuito model with the following parameter settings, that allows the best performance of accuracy metrics: Learning rate = $7.635 \times 10^{-5}$, train_batch_size = 64, eval_batch_size = 32, num_train epochs = 14.

For this configuration we have obtained very good results, achieving a global accuracy of 95.3%. If we analyse the results for each class, Table 15 shows a worse result in the class of detecting stigma (Yes stigma), with differences of around 5% between both of them.

**TABLE 11** | Summary of SVM results for accuracy, recall and F1-score metrics considering two stigma classes (No stigma, Yes stigma).

| Label | Precision | Recall | F1-score |
|---|---|---|---|
| No stigma | 74% | 96% | 84% |
| Yes stigma | 45% | 10% | 16% |

**TABLE 12** | Summary of RF results for accuracy, recall and F1-score metrics considering two stigma classes (No stigma, Yes stigma).

| Label | Precision | Recall | F1-score |
|---|---|---|---|
| No stigma | 75% | 98% | 85% |
| Yes stigma | 68% | 11% | 1% |

#### 4.3.5 | Comparison of Algorithms

Figure 7 shows the accuracy comparison for the four algorithms. Once again, the best performance corresponds to the RoBERTuito model (95.3%), followed by the hybrid CNN-LSTM

**TABLE 13** | Summary of the final values for the optimised hybrid cnn-lstm model for stigma.

| Parameters | Values |
|---|---|
| Number of filters convolutional layer | 128 |
| Number of neurons LSTM layer | 128 |
| Dropout rate | 0.2 |
| Recurrent dropout rate | 0.7 |
| Learning rate | 0.001 |
| Vocabulary size | 2421 |
| Batch_size | 256 |

**TABLE 14** | Summary of the results of the hybrid CNN-LSTM model in terms of precision, recall and F1-score metrics considering two stigma classes (No, Yes).

| Label | Precision | Recall | F1-score |
|---|---|---|---|
| No stigma | 88% | 94% | 91% |
| Yes stigma | 81% | 66% | 72% |

**TABLE 15** | Summary of the results of RoBERTuito in terms of precision, recall and F1-score metrics considering two stigma classes (No, Yes).

| Label | Precision | Recall | F1-score |
|---|---|---|---|
| No stigma | 96.4% | 97.3% | 96.9% |
| Yes stigma | 92% | 89% | 90% |

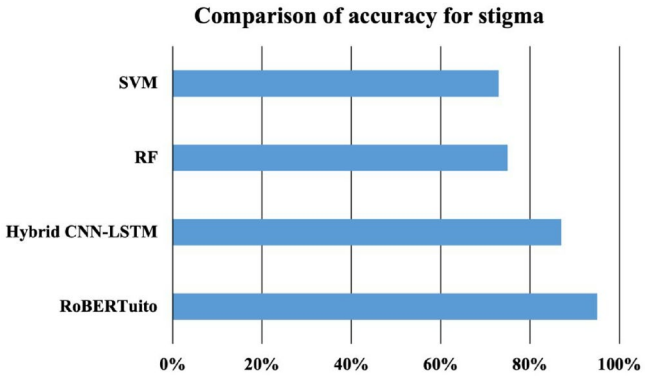



**FIGURE 7** | Comparison of all algorithms (SVM, RF, RoBERTuito, hybrid CNN-LSTM) for the accuracy metric.

**TABLE 16** | Comparison of all algorithms for precision, recall, and F1-score metrics considering stigma classes (No stigma, Yes stigma).

| | No stigma | | | Yes stigma | | |
|---|---|---|---|---|---|---|
| **Classification model** | **Precision** | **Recall** | **F1** | **Precision** | **Recall** | **F1** |
| RoBERTuito | 96% | 97.3% | 97% | 92% | 89% | 90% |
| Hybrid CNN-LSTM | 88% | 94% | 91% | 81% | 66% | 72% |
| RF | 75% | 98% | 85% | 68% | 11% | 20% |
| SVM | 74% | 96% | 84% | 45% | 10% | 16% |

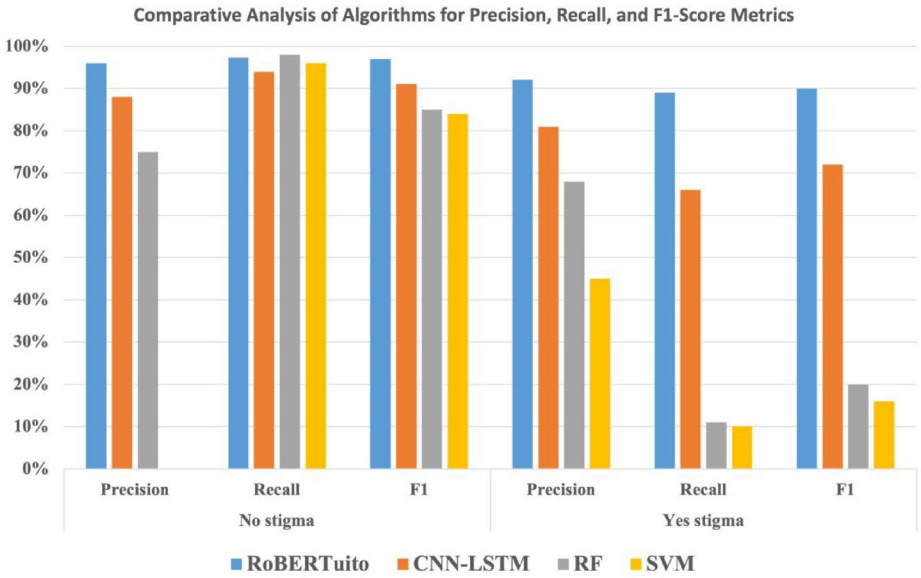



**FIGURE 8** | Comparison of all algorithms for precision, recall and F1-score metrics considering the stigma classes (No stigma, Yes stigma).

model (87%). In contrast, the SVM exhibits the worst performance with 73%, as expected, while RF offers relatively good results, reaching around 75%.

On the other hand, Table 16 shows the values of the prediction, recall and F1-score metrics when predicting stigma. Once again, the best results are obtained by RoBERTuito followed by the hybrid CNN-LSTM algorithm, and the worst performance corresponds to SVM. Comparing the performance for each class, the non-stigma class is the best predicted by all algorithms, where the best algorithm (RoBERTuito) achieves 96% of precision but the worst algorithm also achieves very good results around 74% (SVM model). Moreover, the lower performance in the predict stigma class (Yes stigma class) may be due to the fact that this class has a low percentage in the dataset (around 25% of the samples). Despite this unbalanced distribution, RoBERTuito and the hybrid CNN-LSTM model classify this class with high quality metrics, especially the former, with 92% and 81% accuracy levels, respectively.

These results can be better visualised in Figure 8, where we can clearly observe the better performance of RoBERTuito in all metrics, followed by the hybrid CNN-LSTM model. In addition, it can be appreciated how the algorithms perform slightly better in non-stigma detection. Indeed, since the primary concern in our study is the detection of stigmas, it is essential that the model is highly reliable in its classifications. Therefore, precision emerges as one of the most relevant metrics to consider, and it is observed that all models achieve relatively high levels. High precision indicates that the majority of the positive (or negative) predictions made by the model are correct. This is particularly significant in applications where false positives or false negatives can lead to very negative impacts, as is the case with stigma detection in the field of mental health. Additionally, it is evident that our models also achieve satisfactory results in the metrics of recall and F1-score.

## 5 | Conclusions

This research exhibits pioneering and innovative characteristics across multiple dimensions. It is the first time that the analysis of societal reactions and stigma to mental health disclosures and conversations on social networks has been explored in depth using machine learning algorithms. On the contrary, previous research has mainly focused on pointing out psychopathological aspects, rather than investigating how the community responds. Moreover, this is the first study to focus on Instagram, one of young people's favourite social media platforms, in contrast to the dominance of Twitter in previous research. It's important to highlight that the groundbreaking aspect of this study goes beyond just the Spanish contex, but also on a global scale, as there is a scarcity of research on the same topic at an international level. Spanish is the fourth most widely spoken language

in the world and the second most widely spoken mother tongue, so the algorithms developed, taking into account cultural conditioning factors, can be generalised to other contexts at a global level. Moreover, the methodology employed in this study (corpus design, categorisation guide) could also be replicated in other linguistic contexts, where a similar approach can be adopted to develop new corpora in different languages within social media platforms, allowing for cross-cultural and cross-linguistic comparisons of social responses to mental health problems.

Thus, a corpus has been developed from Instagram posts concerning mental health disclosures by celebrities, which have been manually categorised with varying degrees of polarity and stigma. Both this corpus and the associated machine learning algorithms are accessible on GitHub, enabling researchers to apply them in various contexts related to mental health. Notably, this is the first Spanish corpus specifically created to examine the effects of social responses to discussions about mental health on virtual environments, particularly on Instagram. Furthermore, our study also concentrated on applying machine learning algorithms to predict stigma and levels of polarity to these disclosures of mental health problems on Instagram. In this sense, classical techniques such as SVM and RF were applied, which despite obtaining the worst results, offered average performance with lower computational loads. But more advanced techniques were also applied, based on hybrid CNN-LSTM and BERT algorithms, which achieved the best results in the different accuracy metrics. In terms of polarity, the best results was obtained by RoBERTuito with 96% accuracy, followed by the hybrid CNN-LSTM model, which also achieved high levels of around 80%. The same performance is observed for stigma detection, with RoBERTuito being the best with 96% accuracy, followed by the hybrid CNN-LSTM model, which also obtains high accuracy values of around 87%. In scenarios where false positives or false negatives could lead to significant negative outcomes, such as identifying negative comments (stigma, anger, sadness) associated with mental health, it is essential to achieve high accuracy in their detection. Therefore, it is evident that both RoBERTuito and the hybrid CNN-LSTM algorithms perform exceptionally well across all accuracy metrics, highlighting the effectiveness of these algorithms in identifying extreme polarities.

The proposed framework, which integrates NLP algorithms with a mental health corpus provides insight into the social response to mental health in virtual environments. In fact, it can enable the design of software tools that assist researchers or practitioners in identifying emotional patterns and assessing the influence of social media posts related to mental health issues. This approach is particularly relevant in social networks frequented by younger populations, where understanding emotional responses can inform interventions and support strategies. These software applications could offer intuitive, easy-to-navigate user interfaces, while also integrating customizable dashboards that provide interactive visualisations of emotional data, such as trend graphs and real-time sentiment analysis. Additionally, they could incorporate automated notifications to alert professionals of significant changes in detected emotional patterns, as well as features for generating detailed reports that summarise key findings, thereby facilitating informed decision-making. Therefore, the development of these software applications can offer several advantages, including: (1) Real-time monitoring of emotional and stigmatising expressions, facilitating the identification of emerging trends in mental health issues among the population; (2) Proactive intervention, where the detection of worrying patterns allows specialists to intervene before crises arise, potentially improving overall well-being; (3) Emotional response assessment, providing quantitative data on emotional and stigmatising responses that are useful for research and policy formulation; and (4) Campaign feedback, measuring the effectiveness of awareness campaigns and allowing organisations to adjust their strategies based on users' emotional responses.

However, some limitations have been identified in the course of the research, which will need to be addressed in the future. Thus, one limitation has been the number of celebrity posts selected, as although the ones used have a high impact with massive responses, a wider variety of posts would allow us to generalise the results with greater certainty. Another limitation to generalising this research globally has to do with the culture and language in which the post is made. Stigma has a strong social and cultural component, since it also depends on a social and legal structure that determines it; what is acceptable in certain places is not acceptable in others, or may even be legally punishable. Furthermore, it is necessary to bear in mind that polarity and stigma are constructs with a certain subjective component, which on certain occasions, and more so in written comments where the intentions are not explicit, can affect their categorisation. As a consequence, as a future direction for expansion, it is essential to integrate other social media platforms, particularly TikTok, as it has emerged as one of the most widely used among young people. Although TikTok content is primarily shared through short videos, text messages have a significant impact, as they reflect opinions, beliefs, and behaviours, especially among the youth. Incorporating posts from this social network will enrich the data corpus and enable a comparison of machine learning algorithms' performance in this new social context. Moreover, extending the corpus to multi-language contexts would be highly beneficial, as it would allow direct comparison of emotions expressed in different languages in the field of mental health. This approach could elucidate both similarities and differences in the perception and expression of feelings across linguistic boundaries. Additionally, this multilingual expansion improves the accessibility of mental health research, thus broadening the scope of proposed solutions.

In summary, machine learning analysis of social media posts on mental health by celebrities and influencers is socially significant. In theory, their strong impact on large audiences should help to reduce stigma and promote greater social awareness of mental health problems. In addition, the results of our research can facilitate the design of more effective campaigns and actions, which will have a greater positive impact on society in gene, institutions and individuals dealing with mental health. Consequently, our research can help to identify and promote responsible use of social networks and how to address mental health in these virtual environments.

### Author Contributions

Noemí Merayo: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Software, Data curation, Writing – original

draft, Writing – review and editing, Visualisation, Supervision, Project administration. Clara González-Sanguino: Conceptualization, Investigation, Data curation, Writing – original draft, Writing – review and editing, Visualisation. Alba Ayuso-Lanchares: Conceptualization, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review and editing, Visualisation.


### Acknowledgements

This research has been supported by the University of Valladolid.


### Conflicts of Interest

The authors declare no conflicts of interest.

### Data Availability Statement

The developed corpus, the labelling decalogue and the machine learning algorithms will be available on a Github repository, https://github.com/GCOdeveloper/Mental-Health-Dataset, for researchers to use them in contexts related to mental health in social networks. Other data and models will be made available on request.

### References


Ahmed, A., S. Aziz, C. Toro, et al. 2022. "Machine Learning Models to Detect Anxiety and Depression Through Social Media: A Scoping Review." *Computer Methods and Programs in Biomedicine Update* 2: 1–9. https://doi.org/10.1016/j.cmpbup.2022.100066.

Arco, d F M P., M. D. Molina-González, L. A. U. López, and M. T. Martín-Valdivia. 2021. "A Multi-Task Learning Approach to Hate Speech Detection Leveraging Sentiment Analysis." *IEEE Access* 9: 112478–112489. https://doi.org/10.1109/ACCESS.2021.3103697.

Birnbaum, M., S. K. Ernala, A. Rizvi, M. Choudhury, and J. Kane. 2017. "A Collaborative Approach to Identifying Social Media Markers of Schizophrenia by Employing Machine Learning and Clinical Appraisals." *Journal of Medical Internet Research* 19: e289. https://doi.org/10.2196/jmir.7956.

Bograd, S., B. Chen, and R. Kavuluru. 2022. "Tracking Sentiments Toward Fat Acceptance Over a Decade on Twitter." *Health Informatics Journal* 28, no. 1: 14604582211065702. https://doi.org/10.1177/14604582211065702.

Budenz, A., A. Klassen, J. Purtle, E. Tov, M. Yudell, and P. Massey. 2019. "Mental Illness and Bipolar Disorder on Twitter: Implications for Stigma and Social Support." *Journal of Mental Health* 29: 1–9. https://doi.org/10.1080/09638237.2019.1677878.

Chollet, F. 2021. *Deep Learning With Python*. Shelter Island, New York: Simon and Schuster.

Chrome Web Store. 2023. "One Click Comment Extractor for IG - Chrome Web Store." https:chrome.google.com/webstore/detail/comment-exporter/cckachhlpdnncmhlhaepfcmmhadmpbgp Accessed January 26, 2023.

Corrigan, P. W., and A. C. Watson. 2002. "The Paradox of Self-Stigma and Mental Illness." *Clinical Psychology: Science and Practice* 9, no. 1: 35–53. https://doi.org/10.1093/clipsy.9.1.35.

Datareportal. 2023. Accessed May 22, 2024. "Digital 2023: Global Overview Report." https://datareportal.com/reports/digital-2023-global-overview-report.

Delanys, S., F. Benamara, V. Moriceau, F. Olivier, and J. Mothe. 2020. "Psychiatry on Twitter: A Content Analysis of the Use of Psychiatric Terms in French." *JMIR Formative Research* 6, no. 2: 1–13. https://doi.org/10.2196/18539.

European Commission. 2023. "Eurobarometer Survey 3032." Accessed July 20, 2024.

Fodeh, S., T. Li, K. Menczynski, et al. 2019. "Using Machine Learning Algorithms to Detect Suicide Risk Factors on Twitter." In *International Conference on Data Mining Workshops (ICDMW)*, 941–948.

Fundación Mutua. 2023. "Informe de Salud Mental en España 2023." Accessed July 12, 2024.

González Sanguino, C., J. Medina, J. Redondo, E. Betegón, L. Valdivieso-León, and M. Irurtia. 2024. "An Exploratory Cross-Sectional Study on Mental Health Literacy of Spanish Adolescents." *BMC Public Health* 24, no. 1: 1–10. https://doi.org/10.1186/s12889-024-18933-9.

González Sanguino, C., A. B. Santos-Olmo, S. Zamorano, I. Sánchez-Iglesias, and L. M. Muñoz. 2023. "The Stigma of Mental Health Problems: A Cross-Sectional Study in a Representative Sample of Spain." *International Journal of Social Psychiatry* 69, no. 8: 1928–1937. https://doi.org/10.1177/00207640231180124.

Google Brain Team. 2023. "Tensor Flow Library." https://www.tensorflow.org Accessed May 10, 2024.

Gronholm, P. C., and G. Thornicroft. 2022. "Impact of Celebrity Disclosure on Mental Health-Related Stigma." *Epidemiology and Psychiatric Sciences* 31, no. e62: 1–5. https://doi.org/10.1017/S2045796022000488.

Guntuku, S. C., D. B. Yaden, M. L. Kern, L. H. Ungar, and J. C. Eichstaedt. 2017. "Detecting Depression and Mental Illness on Social Media: An Integrative Review." *Current Opinion in Behavioral Sciences* 18: 43–49. https://doi.org/10.1016/j.cobeha.2017.07.005.

Hallgren, K. 2012. "Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial." *Tutorial in Quantitative Methods for Psychology* 8: 23–34. https://doi.org/10.20982/tqmp.08.1.p023.

Statista. 2023. Accessed May 13, 2024. "Instagram: distribución mundial de usuarios por edad en 2023." https://es.statista.com/estadisticas/875258/distribucion-por-edad-de-los-usuarios-mundiales-de-instagram.

Jain, P., U. Pandey, and E. Roy. 2017. "Perceived Efficacy and Intentions Regarding Seeking Mental Healthcare: Impact of Deepika Padukone, A Bollywood Celebrity's Public Announcement of Struggle With Depression." *Journal of Health Communication* 22, no. 8: 713–720. https://doi.org/10.1080/10810730.2017.1343878.

Jilka, S., C. Odoi, J. Bilsen, et al. 2022. "Identifying Schizophrenia Stigma on Twitter: A Proof of Principle Model Using Service User Supervised Machine Learning." *NPJ Schizophrenia* 8: 1. https://doi.org/10.1038/s41537-021-00197-6.

Joshi, M. L., and N. Kanoongo. 2022. "Depression Detection Using Emotional Artificial Intelligence and Machine Learning: A Closer Review." *Materials Today Proceedings* 58: 217–226. https://doi.org/10.1016/j.matpr.2022.01.467.

Larose, D. T., and C. D. Larose. 2014. *Decision Treesch*. Vol. 8, 165–186. Oxford, UK: John Wiley & Sons, Ltd.

Lee, S. Y. 2019. "Media Coverage of Celebrity Suicide Caused by Depression and Increase in the Number of People Who Seek Depression Treatment." *Psychiatry Research* 271: 598–603. https://doi.org/10.1016/j.psychres.2018.12.055.

Lee, Y., C. Yuan, and D. Wohn. 2021. "How Video Streamers' Mental Health Disclosures Affect Viewers' Risk Perceptions." *Health Communication* 36, no. 14: 1931–1941. https://doi.org/10.1080/10410236.2020.1808405.

Lejeune, A., B. M. Robaglia, M. Walter, S. Berrouiguet, and C. Lemey. 2022. "Use of Social Media Data to Diagnose and Monitor Psychotic Disorders: Systematic Review and Perspectives (Preprint)." *Journal of Medical Internet Research* 24, no. 9: 1–15. https://doi.org/10.2196/36986.

Mäntylä, M. V., D. Graziotin, and M. Kuutila. 2018. "The Evolution of Sentiment Analysis—A Review of Research Topics, Venues, and Top Cited Papers." *Computer Science Review* 27: 16–32. https://doi.org/10.1016/j.cosrev.2017.10.002.

Martínez-Cámara, E., M. T. Martín-Valdivia, L. A. Ureña-López, and A. R. Montejo-Ráez. 2014. "Sentiment Analysis in Twitter." *Natural Language Engineering* 20, no. 1: 1–28. https://doi.org/10.1017/S1351324912000332.

Mental Health America. 2023. "Mental Health America: Youth Data 2023." Accessed July 15, 2024.

Merayo, N., A. Ayuso, and C. González-Sanguino. 2023. "Repository of Corpus of Mental Health in Spanish Social Networks." https://github.com/GCOdeveloper/Mental-Health-Dataset Accessed June 19, 2024.

Mon, A. d M., C. Donat-Vargas, J. Santoma-Vilaclara, et al. 2021. "Assessment of Antipsychotic Medications on Social Media: Machine Learning Study." *Frontiers in Psychiatry* 12, no. 1: 1–13. https://doi.org/10.3389/fpsyt.2021.737684.

NHS Digital. 2023. "Mental Health of Children and Young People in England, 2023 - Wave 4 Follow Up." Accessed July 15, 2024.

NLTK Project. 2023. "Natural Language Toolkit (NLTK) Library." https://www.nltk.org/ Accessed April 1, 2024.

Noble, W. S. 2006. "What Is a Support Vector Machine?" *Nature Biotechnology* 24, no. 12: 1565–1567. https://doi.org/10.1038/NBT1206-1565.

Oden, A., and L. Porter. 2023. "The Kids Are Online: Teen Social Media Use, Civic Engagement, and Affective Polarization." *Social Media + Society* 9, no. 3: 20563051231186364. https://doi.org/10.1177/20563051231186364.

Oneiros. 2023. "Keras Library." https://keras.io/ Accessed June 26, 2024.

Oscar, N., P. A. Fox, R. Croucher, R. Wernick, J. Keune, and K. Hooker. 2017. "Machine Learning, Sentiment Analysis, and Tweets: An Examination of Alzheimer's Disease Stigma on Twitter." *Journals of Gerontology: Series B* 72, no. 5: 742–751. https://doi.org/10.1093/geronb/gbx014.

Pavlova, A., and P. Berkers. 2022. ""Mental Health" as Defined by Twitter: Frames, Emotions, Stigma." *Health Communication* 37, no. 5: 637–647. https://doi.org/10.1080/10410236.2020.1862396.

Pérez, J. M., D. A. Furman, L. A. Alemany, and F. M. Luque. 2021. *RoBERTuito: A Pre-Trained Language Model for Social Media Text in Spanish*. Preprint available at arXiv. https://arxiv.org/abs/2111.09453.

Probst, P., M. N. Wright, and A. L. Boulesteix. 2019. "Hyperparameters and Tuning Strategies for Random Forest." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9, no. 3: e1301. https://doi.org/10.1002/widm.1301.

Python Software Foundation. 2021. "Snowball Algorithms for Stemming." https://pypi.org/project/snowballstemmer Accessed April 15, 2024.

Redondo, J., I. Fraga, I. Padrón, and M. Comesaña. 2007. "The Spanish Adaptation of ANEW (Affective Norms for English Words)." *Behavior Research Methods* 39: 600–605.

Robinson, P., D. Turk, S. R. Jilka, and M. Cella. 2018. "Measuring Attitudes Towards Mental Health Using Social Media: Investigating Stigma and Trivialisation." *Social Psychiatry and Psychiatric Epidemiology* 54: 51–58. https://doi.org/10.1007/s00127-018-1571-5.

Shah, S., H. Ghomeshi, E. Vakaj, E. Cooper, and R. Mohammad. 2023. "An Ensemble-Learning-Based Technique for Bimodal Sentiment Analysis." *Big Data and Cognitive Computing* 7, no. 2: 1–20. https://doi.org/10.3390/bdcc7020085.

Srivastava, R., P. K. Bharti, and P. Verma. 2022. "Comparative Analysis of Lexicon and Machine Learning Approach for Sentiment Analysis." *International Journal of Advanced Computer Science and Applications* 13, no. 3: 71–77. https://doi.org/10.14569/IJACSA.2022.0130312.

Taboada, M., J. Brooke, M. Tofiloski, K. D. Voll, and M. Stede. 2011. "Lexicon-Based Methods for Sentiment Analysis." *Computational Linguistics* 37: 267–307. https://doi.org/10.1162/COLIa00049.

Tomar, P., K. Mathur, and U. Suman. 2022. "Unimodal Approaches for Emotion Recognition: A Systematic Review." *Cognitive Systems Research* 77: 94–109. https://doi.org/10.1016/j.cogsys.2022.10.012.

UNICEF. 2021. "Mental Health." Accessed July 20, 2024.

Witten, I. H., E. Frank, and M. A. Hall. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Boston: Morgan Kaufmann.

World Health Organization. 2024. "Mental Health." Accessed June 10, 2024.

Xue, J., J. Chen, R. Hu, et al. 2020. "Twitter Discussions and Emotions About the COVID-19 Pandemic: Machine Learning Approach." *Journal of Medical Internet Research* 22, no. 11: e20550. https://doi.org/10.2196/20550.

## Supporting Information

Additional supporting information can be found online in the Supporting Information section.