

> REPLACE THIS LINE WITH YOUR MANUSCRIPT ID NUMBER (DOUBLE-CLICK HERE TO EDIT) <

Revealing Emotional Insights from Mental Health Discussions on Instagram and TikTok Using BERT Models

N. Merayo, A. Ayuso-Lanchares, and C. González-Sanguino

Abstract— The research addresses challenges related to mental health issues in social media by integrating natural language processing. First, the study extends a previous corpus labelled with emotions and polarity by including new Instagram and TikTok posts related to celebrity and influencer disclosures about mental health. This corpus is the first Spanish corpus designed to analyse the impact of social responses to mental health narratives on two of the most widely used social networks.

Secondly, the research integrates BERT (Bidirectional Encoder Representations) classification models to improve emotion and polarity detection. One of the modelled algorithms, MenTaiBERT, leveraging a specialised classification layer demonstrates superiority over the other BERT algorithms, achieving 99% accuracy in emotion detection and 98% accuracy in polarity. Indeed, MenTaiBERT significantly outperforms the accuracy of the other algorithms by up to 13 percentage points.

Third, A user-friendly graphical tool has been designed, based on the previous corpus and classification models, to help practitioners identify emotional patterns in social media posts related to mental health.

In summary, analysing through innovative artificial intelligence strategies the emotional impact of celebrity posts on social networks is crucial, especially among young people, as these platforms significantly influence their self-esteem, perception of reality and emotional well-being.

Index Terms— Bidirectional encoder representations models, Emotional response, mental health, natural language processing, sentiment analysis, social networks.

I. INTRODUCTION

THE strong impact of social networks as one of the main means of communication has transformed the way we interact. Also there is a strong concern about the growing presence of mental health related problems in these environments, especially among young people, a more influential population in these environments. In 2024, it is estimated that there are around 5 billion monthly active social media users worldwide [1]. By 2025, this number is expected to rise to approximately 5.4 billion. In Spain, the number of social media users in January 2024 was 39.70 million, with a

projection to exceed 43 million by 2025 [1]. Indeed, in 2023, 90% of Spaniards between the ages of 16 and 24 were already using some type of social media. The second largest age group of users was those between 25 and 34, with more than 85%. In this social context, Instagram is the second most used social network in Spain, with 24 million active users at the beginning of 2024, while TikTok ranked fourth with 16.74 million users over the age of 18 [1]. These data are consistent with global levels, where Instagram is in first position and TikTok is in fifth position.

On the other hand, there has been a noticeable rise in the significance of mental health, particularly in advanced societies [2]. The World Health Organization states that mental health issues are increasing globally, affecting approximately 20% of children and adolescents worldwide [3]. In Spain, a recent report shows that four out of ten Spaniards (39.3%) rate their current mental health negatively, and of the more than 2000 people surveyed in this study, 42.1% had suffered from depression in their lifetime; 47.6% had experienced anxiety or panic attacks and 36.9% had experienced prolonged anxiety over time [4]. In addition, it also shows increasing trends of suicide attempts, especially among women and younger people [5].

This increase in mental health problems is also reflected on social media, with celebrities and influencers increasingly talking about their private lives [6, 7], with mental health problems being present in many of their posts. These posts are highly relevant due to their significant impact, both in quantitative terms, as they reach thousands of people, and in qualitative terms, given that they are shared by highly influential public figures. However, the extent of the influence of this phenomenon is difficult to manage in social networks [8, 9], not knowing whether the impact on society is positive or negative. Some studies have found that these social media posts can have a positive impact, fostering empathy and encouraging discussions about mental health [8, 10]. However, other research highlights negative associations between celebrity admiration and mental health [11], as well as concerns about the commercialization of mental health and the promotion of an idealized 'celebration of self-care' in celebrity posts [12]. These

This research was supported by these projects: POPTEC Equisam project (ref. 0298_EQUISAM_2_E). All authors have contributed equally to the research.

N. Merayo is with the Communications and Telematic Engineering Department, E.T.S.I. Telecomunicación, Universidad de Valladolid, Paseo de Belén 15, Valladolid, Spain. noemer@tel.uva.es.

A. Ayuso is with the Department of Pedagogy, Faculty of Medicine, Universidad de Valladolid, Av. Ramón y Cajal, 7, 47005 Valladolid, Spain. alba.ayuso@uva.es.

C. González-Sanguino is with the Department of Psychology, Education and Social Work Faculty, Universidad de Valladolid. P.º de Belén, 1, 47011 Valladolid, España. Clara.gonzalez.sanguino@uva.es

> REPLACE THIS LINE WITH YOUR MANUSCRIPT ID NUMBER (DOUBLE-CLICK HERE TO EDIT) <

portrayals are often unattainable for the average social media user, making them unrepresentative and potentially limiting their positive influence on improving public mental health. Understanding the emotional responses of social media users to these posts by famous individuals could provide valuable insights into their effects and determine whether these social media posts have a beneficial impact. What emotional responses do these celebrity posts evoke in society? Are public reactions positive or negative? How do social media users' emotional responses provide insights into the acceptance of mental health in society? This knowledge could also inform the design of more effective mental health promotion campaigns. However, society's response to this phenomenon remains unknown for now.

In this social concerning environment, Artificial Intelligence (AI), especially Natural Language Processing (NLP) and Bidirectional Encoder Representations (BERT) models [13, 14], have revolutionised data analysis by enabling deeper and more accurate understanding of large volumes of text. These technologies are crucial for analyzing emotional responses, such as polarity levels and emotions, by detecting nuances and patterns that were previously difficult to identify. The ability to analyze these emotional reactions is essential for understanding how social media users respond to content related to mental health, particularly regarding the impact of celebrity posts. By identifying the emotional tone of these responses, we can gain insights into public attitudes, mental health acceptance, and how these emotions influence the effectiveness of mental health campaigns and messages. A key technique in this process is sentiment analysis, a subcategory of NLP that enables computers to understand and interpret written or spoken human language [15]. By applying sentiment analysis, researchers can better assess emotional trends and their potential societal impacts. Their integration into the field of mental health is groundbreaking, particularly in social media applications, where they can provide valuable insights into users' emotional well-being and the detection of behavioral patterns.

Although Malgaroli et al. [16] highlighted a significant increase in NLP research in mental health since 2019, there are a limited number of studies, especially incorporating BERT models, for examining mental health issues on social networks, particularly within the Spanish context and on widely used platforms among youth, like Instagram or TikTok. On the one hand, the application of these techniques in these contexts has mainly focused on the detection of mental disorders [17, 18], rather than assessing its impact on society, using clinical samples of diagnosed individuals or social network users [19]. On the other hand, the main social network that has been researched is Twitter (now X) [20], although Instagram and TikTok have become the most used social networks worldwide, especially among young people, and X is aimed at other audiences. In this way, TikTok content is primarily shared through short, vertical videos with interactive elements like dubbing, visual effects, and text, while Instagram provides a more diverse range of formats, including photos, stories, and

short videos (Reels). However, messages reacting to posts on both social networks have a significant impact, as they reflect opinions, preferences, beliefs or behaviours, especially made by young people. This diversity of content and interaction styles highlights the unique way in which each social networking platform influences the expression of feelings of its users. Furthermore, there are hardly any studies that analyse the impact of emotional response in the field of mental health on social media, especially when it comes from celebrities and influencers, who wield great power over millions of users, which is very difficult to monitor and analyse. In this framework, in a recent research [21], we proposed in a novel approach the creation of an emotion-tagged corpus from responses to posts made by influencers about disclosures of mental health problems on Instagram, together with the modelling and application of a set of machine learning algorithms. However, there is ample room for improvement and extension of this initial proposal, both in terms of the corpus and its linguistic diversity, as well as the machine learning algorithms implemented. Indeed, in our initial research [21], we conducted a comparison with classical algorithms, including Random Forest (RF) and Deep Learning. The results were relatively poor, with accuracy metrics of approximately 48% for RF and around 72% for the Deep Learning algorithm. Accordingly, this new research makes three crucial contributions that respond to new challenges related to this area of mental health and social networks:

- 1) To extend the initial corpus labelled with emotions and polarity with new comments from Instagram and TikTok posts related to celebrity and influencer disclosures about mental health issues. This corpus and a categorization decalogue are freely available in a GitHub repository [22].
- 2) To model BERT-based classification models to identify emotions and polarities expressed on Instagram and TikTok linked to mental health disclosures made by celebrities and influencers. Models are accessible in GitHub [22].
- 3) To design an intuitive and easy-to-use graphical software tool for practitioners and researchers to identify emotional patterns in social media posts dealing with mental health issues by integrating the corpus and classification models.

This paper is structured as follows. First, Section II outlines the state of the art. Then, Section III details the methodology followed to extend the corpus, and Section IV describes the corpus itself. Section V explains the BERT-based classification models, Section VI reports the main results and Section VII describes the graphical interface. Finally, Section VIII and IX summarizes the main limitations and conclusions.

II. STATE OF THE ART

Few studies incorporate artificial intelligence to analyse and detect the impact of mental health on social networks and online communities [23]. In fact, most research focuses on the detection of suicide [24, 25, 26, 27, 28, 29, 30], depression and anxiety [31, 32, 33, 34, 35, 36, 37, 38, 39], schizophrenia [40], or anorexia [41]. Some of these studies focus on detecting these mental illnesses and symptoms through facial analysis, rather

> REPLACE THIS LINE WITH YOUR MANUSCRIPT ID NUMBER (DOUBLE-CLICK HERE TO EDIT) <

than relying on text analysis [42, 43]. Besides, it is important to note that most of these studies have focused on the social network X (formerly Twitter), neglecting much more popular platforms today, especially among young people, such as TikTok or Instagram.

In the context of social networks and the social interactions they generate, emotions play a crucial role not only in individual survival but also in regulating social interactions. According to Kavaklı [44], emotions help individuals maintain interpersonal relationships and group cohesion, as well as influence social behaviors. For example, emotions like happiness facilitate cooperative interactions, while anger or contempt allow for setting boundaries and signaling transgressions. These social functions are especially relevant in the context of social networks, where emotional responses mediate the relationship between content characteristics and user engagement, amplifying their impact on interactions [45]. On platforms such as Instagram and TikTok, positive emotions, such as admiration and gratitude, can foster a sense of community and validation, while negative emotions, such as contempt, can reinforce social boundaries by signalling disapproval. Moreover, empathetic responses play a crucial role in normalising mental health discussions and encouraging openness, making the study of emotions essential to understanding users' behaviour on social media [46, 47].

On the other hand, celebrities and influencers are increasingly sharing and publicizing their mental health challenges on social media [9, 48]. Notably, some studies using artificial intelligence have been conducted to assess the emotional responses to social media posts, providing insights into their social impact. For example, Alvarez-Mon et al. [49] explored Twitter posts about antipsychotic medications to understand responses and clinical interests. Jilka et al. [50] identified stigmatizing tweets about schizophrenia. Oscar et al. [51] modeled stigmatization related to Alzheimer. Bograd et al. [52] examined feelings towards obesity, and Xue et al. [53] analyzed the response to COVID-19. However, all these research studies have been conducted exclusively by analyzing the social network Twitter, which comes with certain limitations. Twitter is not the most popular platform among young people, has fewer users compared to other more 'trendy' social networks, and is significantly more politicized. To begin to fill this gap, in a research, Merayo et al. have recently proposed the application of machine learning algorithms to understand society's emotional reactions to mental health-related posts on Instagram [21], a much more popular social networking site, especially among young people.

This analysis reveals a limited number of studies using AI for sentiment analysis and social response to mental health issues on social networks, especially in the Spanish context and on platforms like Instagram and TikTok, popular among young people. Furthermore, BERT models have represented a significant breakthrough in NLP, as their transformer-based architecture allows them to capture contextual information in both directions, leading to substantial improvements in tasks such as sentiment analysis and providing much more accurate

results. Due to these advantages, an increasing number of studies and research in the literature on mental health issues are adopting the integration of BERT models, demonstrating improved performance. However, most of this research primarily focuses on detecting symptoms of mental health disorders, such as anxiety and depression [33, 36, 37] or suicide [25, 26], rather than analyzing the societal emotional response to mental health disclosures on social networks. Consequently, the integration of AI to analyse emotional responses to mental health-related posts on social networks - particularly those based on BERT models - holds significant potential due to their considerable social and global impact. This technology can provide valuable information on the social acceptability of mental health, identify which messages are most likely to resonate with audiences, and guide the design of more effective promotion and prevention campaigns. In addition, AI can help detect emerging trends, uncover hidden emotional patterns and assess the influence of different types of influencers in shaping public perceptions of mental health. Moreover, it is essential to develop new linguistic corpora tailored to this area of research to facilitate the integration of AI techniques in the analysis of mental health disorders. Consequently, further research is needed in these areas. Moreover, it is essential to develop new linguistic corpora tailored to this area of research to facilitate the integration of AI techniques in the analysis of mental health disorders. Consequently, further research is needed in these areas.

III. METHODOLOGY

A. Selection and integration of new Instagram and TikTok posts

In this research, profiles of celebrities and influencers with Instagram and TikTok accounts were analysed, reviewing news (in press, television and social media) related to statements about their mental health problems. The tracking of these new social media posts was conducted between September 2023 and March 2024. It is important to note that, while our initial corpus only included posts by women [21], a subsequent process of searching and identifying posts revealed that most of the content was still authored by women. However, this new research also identified and incorporated some posts by men. In addition to Instagram, TikTok posts have been added to complement the corpus with data from the two social networks most used by the youngest and therefore most vulnerable and influential population. This emphasis on gender and social network diversity aims to mitigate data bias. If the data used to train artificial intelligence is significantly biased, it could result in misleading or inaccurate information, posing potential risks to health, equity, and inclusivity. This introduces substantial ethical concerns, as biased algorithms may reinforce stigma or marginalize vulnerable communities. Thus, the selection of the new Instagram and TikTok posts followed the criteria below:

- a) The posts had to be from male or female influencers and celebrities and related to their mental health problem.
- b) The posts had to be published by Spanish celebrities

> REPLACE THIS LINE WITH YOUR MANUSCRIPT ID NUMBER (DOUBLE-CLICK HERE TO EDIT) <

and influencers and written in Spanish.

- c) The selection of the celebrities and influencers was based on a significant impact, so a minimum of 100000 followers was set for their profiles.

This methodology is consistent with the guidelines followed in our initial corpus [21]. According to these requirements, a total of 22 new posts were selected, 8 posts on TikTok by women, 8 posts on Instagram by women and 6 posts on Instagram by men. All these new posts are in addition to the 11 women's Instagram posts we already had in our initial corpus [21]. Notably, no high-impact posts were found on TikTok by male celebrities, and very few on Instagram. Furthermore, these new 22 posts show the next characteristics:

- a) The format of the posts remained uniform, consisting of a photo (9 posts) or a video (13 posts) accompanied by relevant text.
- b) The content referred to the manifestation of mental health symptoms or highlighted the importance of specialised care in these situations.
- c) The most frequently discussed mental health concerns were anxiety and depression.

These posts were downloaded using the software One Click Comment Extractor for IG (2024) [54].

On the other hand, the ethical dimensions of artificial intelligence in social sciences are becoming a crucial issue and working with mental health issues in social networks raises ethical concerns of privacy and ethical protocols [55]. As a consequence, our study has been conducted following the guidance of various scholars on the application of AI in social sciences, emphasizing the importance of result interpretability and ensuring transparency in the design of algorithms and datasets [55]. Therefore, only public posts and comments were chosen, and comments were carefully anonymized by removing @mentions, names, usernames, and URLs. Furthermore, the dataset was processed to exclude any identifiable information or sensitive content that could compromise user privacy, adhering to the principle of data minimization by collecting only the data strictly necessary for our research purpose. These data will be used exclusively for the purposes of this research and for no other purpose. The social media accounts used were strictly public, ensuring that no private data was accessed or utilized. This approach allows us to ensure that data collection and usage are conducted ethically, respecting user privacy, avoiding bias and discrimination, maintaining transparency, and adhering to data protection regulations such as GDPR (General Data Protection Regulation). In this way, our collected data complies with the principles of purpose limitation, transparency, data minimization, and the prevention of discrimination and harm. Finally, our research received formal approval from the Ethics and Deontology Committee of the University of Valladolid (PI 23-3365), ensuring its adherence to the highest standards of ethical conduct and academic integrity. Supplementary Table I (in the supplementary files) describes the selected Instagram and TikTok posts, the number of responses to each post, the usernames and account names of the celebrities/influencers (instead of using full names to

maintain anonymity and privacy), and their number of followers, ensuring a balance between data transparency and anonymization as discussed.

B. Description of dataset labelling

To carry out manual labelling of the corpus we followed the rules of previous literature [49, 52, 56] and developed a labelling guide with guidelines and examples of the different categories, in our case polarity and emotions [21]. For polarity, we adopted the standard binary classification, designating labels as Positive (P) and Negative (N). We included a Neutral (NEU) label exclusively for messages that could not be categorized with a specific polarity. On the other hand, the labelling of emotions was more complex. Initially, the classification was based on Ekman's model of basic emotions [57], which includes fear, anger, sadness, disgust, happiness, and surprise. However, this approach proved insufficient to capture the range of positive emotions observed in many comments. This aligns with previous findings that suggest basic emotion models may oversimplify complex emotional phenomena, leading to reduced agreement among annotators when analyzing emotive language [58]. To address this, the categories for positive emotions were expanded, based on Plutchik's [59] emotional framework and Fredrickson's perspective on positive emotions [60]. The final categories included love/admiration, gratitude, understanding/empathy/identification, sadness, and anger. Although a more detailed description of this process can be found in [21], the emotions considered in our research along with a brief description are shown below:

- Love (Love/Admiration): messages expressing approval, admiration and love, such as *"Wowwww, you're a role model absolute admiration!"*
- Gratitude: messages containing appreciation for sharing content or experiences related to mental health, such as *"Thank you, you have helped me a lot."*
- Empathy (Comprehension/Empathy/Identification): messages that transmit interest, understanding or identification with the situation or context, such as *"I agree with you. I recently started following you, and I really identify a lot with you."*
- Sadness: messages that express pity for the person or for their situation, such as *"Aww, what a pity! Everything passes... it's good to cry too."*
- Anger (Anger/Contempt/Mockery): messages conveying irritation, hatred and attacks on the person as ridiculous and shallow, such as *"But if everything is fake"*
- Neutral: messages that do not convey any clear emotion or lack sufficient context to be labelled appropriately, such as *"What's wrong?"*

C Labelling methodology

The same labelling methodology has been followed as for the initial corpus and is broadly summarised as follows, although a detailed description can be found in [21]:

- Data cleaning: This process consists of discarding

> REPLACE THIS LINE WITH YOUR MANUSCRIPT ID NUMBER (DOUBLE-CLICK HERE TO EDIT) <

comments written in languages other than Spanish, those that only contain acronyms, those that lack linguistic coherence and those that only contain mentions of other users.

- Exclusion of emoticons: To focus solely on the linguistic impact.
- Expert labelling: Two independent experts (psychologists, trained people) labelled the corpus separately, and to ensure accuracy and consistency, they were provided with a guide containing detailed instructions and real examples. Subsequently, a third expert reviewed the results to resolve discrepancies. If disagreements arose, they were discussed until a consensus was reached; otherwise, the message was discarded. This method guaranteed accuracy and prevented errors or biases in the final labeled corpus.

After completing this labelling process, 2086 new Instagram and TikTok comments were selected for integration into the initial corpus, making a total of 4373 messages in the final corpus. The disagreement rate was around 7.82%, showing strong agreement among the evaluators. This result reinforces our confidence in the consistency of the categorisation process.

IV. CORPUS STATISTICS

The complete corpus, integrating the extension of comments made in this research, consists of 4373 comments, in which the distribution of posts is as follows: 2751 (63%) Positive (P), 1281 (29%) Negative (N) and 340 (8%) Neutral (Neu), and their distribution is shown in Fig. 1. The messages are also classified according to the corresponding emotion (Fig. 2), with the most frequent being: Love with 1267 samples (29%) and Empathy with 1148 samples (26%), followed by Anger with 830 samples (19%). In contrast, the less common emotions in the corpus are Gratitude with 548 samples (12.7%), Sadness with 293 samples (7%), and the Neutral class with 285 samples (6.7%). As can be seen in both graphs, there is a certain imbalance between some classes in the corpus, especially in the Gratitude, Sadness and Neutral classes.

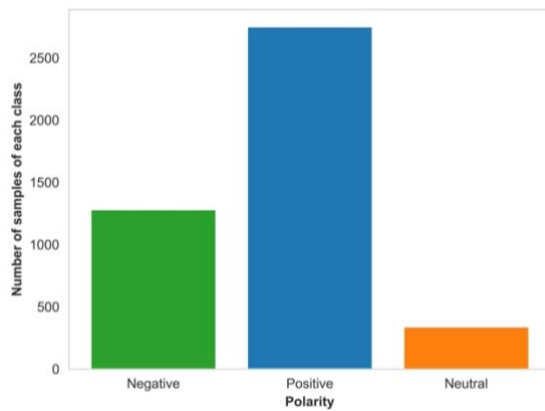


Fig. 1. Distribution of the corpus by polarity.

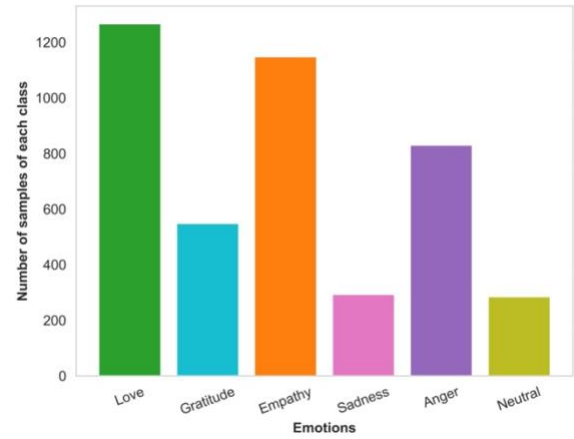


Fig. 2. Distribution of the corpus by emotions.

V. CLASSIFICATION MODELS BASED ON BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS (BERT)

Large Language Models (LLMs) have become a fundamental building block for natural language generation and understanding. In this context, BERT stands out for its ability to enhance deep text understanding in tasks such as classification, information extraction, and sentiment analysis. BERT is a pre-trained language model developed by Google that uses the Transformer architecture. One of the key strengths of BERT models is its highly effective ability to capture the bidirectional context of words in a sentence [14, 61]. This means that BERT can understand the meaning of a word based on the surrounding context, both preceding and following it in the sequence, making it a powerful and efficient model that outperforms classical techniques. While BERT excels in this capability, it is important to note that other models, particularly those based on transformer architectures or recurrent networks, can also capture bidirectional context. However, BERT may perform better due to its pretraining approach using masked language modeling, which allows for more efficient contextual understanding compared to other alternatives. Furthermore, other alternative advanced models such as GPT (Generative Pre-trained Transformer) technologies offer significant advantages for text classification by generating rich and coherent contextual representations, easily adapting to specific tasks through prompts, which enhances flexibility and efficiency in classifying complex texts. Unlike these powerful architectures, our initial research focused on traditional classification models such as Random Forest (RF) and Deep Learning approaches. Specifically, our Deep Learning model featured a hybrid architecture that combined a Convolutional Neural Network (CNN) with a Long Short-Term Memory (LSTM) network [21]. This architecture consisted of an Embedding layer, followed by a one-dimensional convolutional layer, a MaxPooling layer, an LSTM layer, and finally, a Dense layer. However, both the RF model and the hybrid CNN-LSTM model delivered relatively modest performance, with accuracy metrics ranging between 48% and 72%. Consequently, in this new research, we transitioned to more advanced architectures based on BERT and GPT, specifically the following classification models:

> REPLACE THIS LINE WITH YOUR MANUSCRIPT ID NUMBER (DOUBLE-CLICK HERE TO EDIT) <

- A. RoBERTuito: A linguistic model for social networking texts in Spanish, based on RoBERTa [62] and trained on 500 million tweets [63]. It outperforms other models for Spanish such as BETO, BERTin and RoBERTa-BNE.
- B. GPT-3.5 Turbo (gpt-3.5-turbo-0125): It is an advanced language model developed by OpenAI, designed to generate human-like text based on the input it receives. Its advantages include improved performance and efficiency, making it faster and more cost-effective for various applications compared to its predecessors [64].
- C. MenTaiBERT (with a classification layer): A new model based on the pre-trained “bert-large-uncased” BERT model [65]. This pre-trained BERT model was trained on large text data to understand the context of words bidirectionally and has the ability to be tuned for specific tasks with relatively small amounts of data, making it very versatile and effective for a wide range of language-related applications. To achieve this, a classification layer is added on top of the pre-trained “bert-large-uncased” model to adapt it to our specific context. The number of neurons in this classification layer corresponds to the number of classes of the classification problem, emotions or polarities in our case. After incorporating this new classification layer, the model will undergo training using the labelled data that fits the specific classification task, in our case the corpus of mental health disclosures on Instagram and TikTok. Throughout training, the model weights are fine-tuned using optimization methods (backpropagation, stochastic gradient descent, among others). When a specific classification layer is added to a pretrained BERT model, its ability to adapt and perform tasks related to language improves. This enhancement is due to the new layer allowing the model's weights to be adjusted to optimize its performance in the specific classification task, using relevant information extracted during the BERT model's pretraining.

Thus, RoBERTuito and GPT-3.5 Turbo use transfer learning, a process in which a model is trained on a large-scale dataset and then reused to learn a new task. Transfer learning with GPT-3.5 Turbo and similar occurs primarily during the training process, which is divided into two key phases: pretraining and fine-tuning [64]. In the first phase, the model acquires broad linguistic and factual knowledge from general-purpose datasets that can be applied across a wide variety of tasks, such as summarization, translation, and question answering (without task-specific tuning), and in the second phase, it specializes in instruction-following and ethical alignment through fine-tuning with human feedback and task-specific datasets. In contrast, MenTaiBERT employs a pre-trained BERT model solely to encode and contextualize text through the embeddings provided by the model. However, a new classification layer is added, which will be trained with the specific labeled data for the classification task, thereby optimizing the performance of the initial model [66]. The workflow of a BERT model involves several key steps:

1. **Tokenization and Encoding:** The input text is tokenized and encoded using BERT's tokenizer, converting it into numerical tokens.
2. **BERT Processing:** The encoded tokens are passed through a pre-trained BERT model, which produces contextualized

representations of the text. These representations capture both semantic meaning and contextual information in high-dimensional feature vectors.

3. **Classification:** The contextualized representations are inputted into a classification layer. It may consist of multiple neurons generating a probability distribution across potential labels.
4. **Prediction:** The label with the highest probability is selected as the final prediction.

Moreover, it is necessary to adjust the hyperparameters of classifications models to optimise its performance for the specific task by means of the fine-tuning process [67]. This may include adjustments to the learning rate, batch size, number of training epochs and other settings related to the model architecture. In RoBERTuito and GPT-3.5 Turbo, hyperparameter optimization is carried out using Weights and Biases (wandb) and the `trainer.hyperparameter_search` function, which maximizes accuracy using a search space defined by `wandb_hp_space`. This space includes the hyperparameters that need to be adjusted. The entire process is logged in wandb to facilitate analysis and result tracking. Indeed, as the parameters of GPT-3.5-turbo are not open source, the fine tuning process in GPT-3.5-turbo involves tuning the model to optimise its performance for specific tasks. The process begins with the preparation of the data, which must be in a compatible format—specifically, JSON (JavaScript Object Notation) files containing input-output examples (with chat-style format). This JSON file includes specific roles: the System, which guides the model's response with detailed instructions. For instance, in the case of emotion classification, the prompt was: “What is the emotion of the following text? Respond with ‘Love/Admiration’, ‘Sadness’, ‘Neutral’, ‘Gratitude’, ‘Anger/Contempt/Mockery’, or ‘Comprehension/Empathy/Identification’.” For polarity classification, the instruction was: “What is the polarity of the following text? Answer with ‘Positive’, ‘Negative’, or ‘Neutral’”; The User role corresponds to the input message—in this case, an Instagram or TikTok comment—while the Assistant role contains the model's generated response, i.e., the predicted emotion or polarity. This structure enables the model to understand and generate contextual responses based on the provided data. Subsequently, the OpenAI API is used to load this data and launch the training process, where the model adapts its parameters based on the examples provided. During the fine-tuning process (supervised fine-tuning or instruction-based fine-tuning), GPT-3.5 Turbo's hyperparameters—particularly the learning rate, batch size, and number of epochs—are optimized to balance the model's performance and prevent overfitting. This allows GPT-3.5 Turbo to specialize in specific tasks without compromising its generalist capabilities. In short, this is supervised fine-tuning to adapt the conversational behaviour of the model to a specific task (such as emotion or polarity analysis), using the chat-like message flow and integrating wandb for tracking and monitoring the process. More detailed information on this process can be found in our GitHub repository [22].

Furthermore, in the context of fine-tuning language models like BERT, the main difference between fine-tuning (where the

> REPLACE THIS LINE WITH YOUR MANUSCRIPT ID NUMBER (DOUBLE-CLICK HERE TO EDIT) <

model's weights are updated) and using a frozen model with an additional classifier layer (often referred to as feature-based transfer learning) lies in how the pre-trained model's weights are handled. In the latter case, BERT's weights remain frozen (MenTaiBERT in our case), and only an additional classifier layer is trained using its representations, allowing for task-specific adaptation with minimal computational cost and a lower risk of overfitting. In contrast, fine-tuning updates the weights of all or some layers of the model, enabling BERT (RoBERTuito, GPT-3.5 Turbo in our case) to adjust its internal representations based on the new task's specific data, albeit at a higher computational cost and with an increased risk of overfitting. Therefore, the K-Fold cross-validation technique is best suited when training only the classifier layer (as in the case of MenTaiBERT), since the fixed weights ensure that the base model does not change between iterations, allowing each partition of the dataset to be evaluated under consistent conditions. In contrast, in full fine-tuning, the model's weights change with each training iteration, meaning that each K-Fold iteration results in a different model, making it difficult to fairly compare across folds and increasing the variability of the results. Consequently, full fine-tuning requires multiple training sessions to find a model with good performance and generalization ability. For this reason, K-Fold is applied to MenTaiBERT, where only the classifier layer is trained, but not to RoBERTuito or GPT-3.5, which undergo full fine-tuning.

Once fine-tuning is complete, it is necessary to evaluate the model's performance on a separate test dataset. This involves calculating performance metrics such as accuracy, precision, recall and F1-score to determine how well the model generalises to unseen data [68].

VI. RESULTS AND ANALYSIS

This section shows the performance of BERT-based classification models using the metrics of accuracy, precision, recall and F1 score for both emotion and polarity detection.

A. Emotions results

BERT models provide the Trainer class to fit pre-trained models to a specific data set, i.e. to find the optimal training parameters. Therefore, the optimal values of the hyperparameters of the BERT models used in this research are:

- RoBERTuito: learning rate = 0.0003198, train_batch_size = 32, eval batch size = 8, num_train_epochs = 11.
- GPT-3.5 Turbo: learning rate Multiplier = 2, batch size = 6, seed = 976547957, epochs = 3.
- MenTaiBERT: learning rate = 10^{-5} , batch_size = 8, num_train_epochs = 8, k-fold = 6, optimizer = Adam.

The results in Table I show that MenTaiBERT outperforms both RoBERTuito and GPT-3.5 Turbo across all emotional categories. Notably, MenTaiBERT achieves an overall accuracy of 98%, significantly surpassing the 88% attained by GPT-3.5 Turbo and the 85.5% of RoBERTuito. This superior performance can be attributed to MenTaiBERT's specialized design, which incorporates an additional classifier layer. Unlike other BERT-based models that rely on generalized contextual training, MenTaiBERT's architecture emphasizes

classification-specific training. While traditional LLMs excel in text generation and generalization, MenTaiBERT effectively leverages this capability by introducing an optimized classifier layer. This refinement enhances its ability to encode and decode language patterns, improving its accuracy in text classification tasks compared to models like RoBERTuito or GPT-3.5 Turbo. Furthermore, MenTaiBERT shows the most significant improvements over the other two algorithms in Sadness, Neutral and Comprehension/Empathy. MenAI's performance suggest that it is not overfitting because its performance is consistently high in all emotion categories, including the more difficult ones such as Neutral or Sadness, where the other two models show more problems. Overfitting would likely result in disproportionately high scores on the easier categories, but significantly lower scores on the more difficult ones or on metrics such as recall. The balanced and uniformly strong results on all metrics indicate that MenAI generalises well rather than memorising patterns from the training data.

TABLE I
RESULTS OF PRECISION, RECALL AND F1-SCORE METRICS
FOR EMOTION DETECTION

Emotions	Metrics	RoBERTuito	GPT3.5 Turbo	MenTaiBERT
Love/Admiration	Precision	91%	93%	98%
	Recall	92%	93%	99%
	F1-Score	92%	93%	99%
Comprehension/ Empathy/Identification	Precision	83%	91%	99%
	Recall	83%	81%	99%
	F1-Score	83%	85%	99%
Gratitude	Precision	92%	86%	99%
	Recall	94%	98%	97%
	F1-Score	93%	91%	98%
Sadness	Precision	73%	70%	99%
	Recall	75%	88%	98%
	F1-Score	74%	78%	97%
Anger/Contempt/ Mockery	Precision	88%	90%	99%
	Recall	87%	93%	98%
	F1-Score	87%	91%	98%
Neutral	Precision	65%	73%	98%
	Recall	52%	62%	97%
	F1-Score	57%	67%	96%
Global Accuracy		86%	88%	99%

The same performance can be observed in the confusion matrixes of Fig. 3 (a), (b) and (c). As can be seen the RoBERTuito and GPT3.5 Turbo algorithms show slight confusions between the emotions Comprehension/Empathy/Identification and Love/Admiration, and between Comprehension/Empathy/Identification and Sadness. This behaviour may be due to the fact that when people show love, they often carry implicit emotions of identification and understanding towards other people and/or their situations. The same happens when a person shows empathy, which could be because they show support or empathy towards another person in a situation of sadness, in this case in the face of a mental health problem.

> REPLACE THIS LINE WITH YOUR MANUSCRIPT ID NUMBER (DOUBLE-CLICK HERE TO EDIT) <

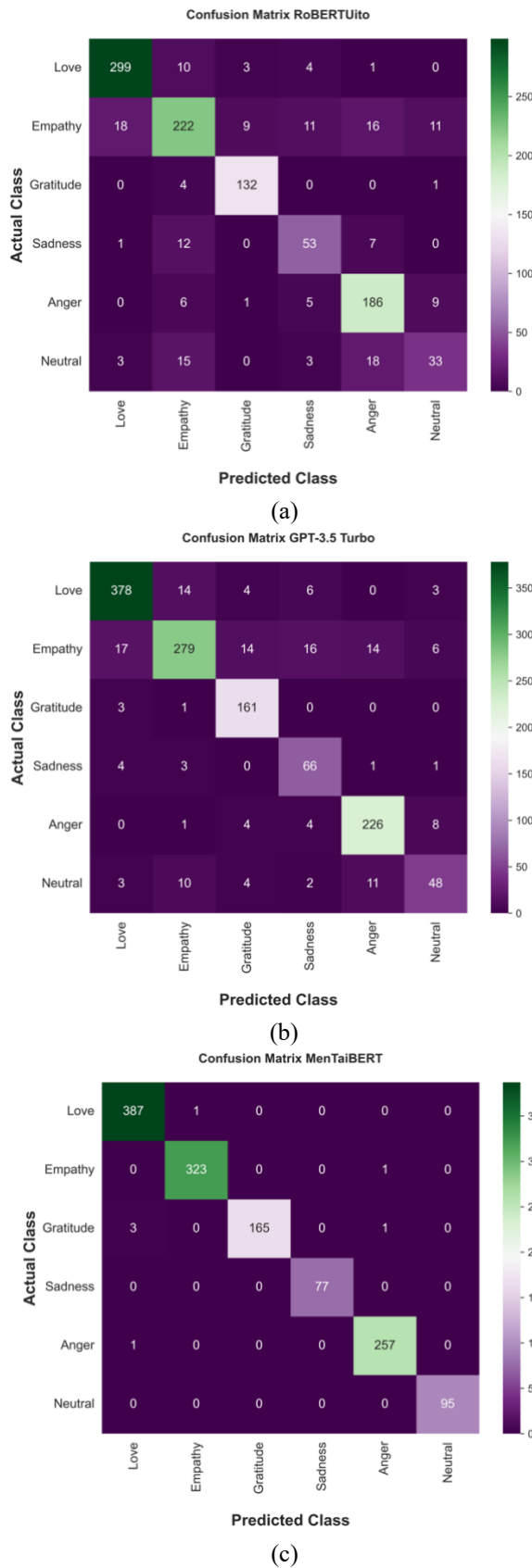


Fig. 3. Confusion matrix of algorithms: (a) RoBERTuito (b) GPT3.5-Turbo (c) MenTaiBERT on emotion detection.

B. Polarity results

In the polarity case, BERT models show the following optimal parameter settings:

- RoBERTuito: learning rate = 0.0008154, train_batch_size = 64, eval batch size = 32, num_train_epochs = 9.
- GPT-3.5 Turbo: learning rate Multiplier = 2, batch size = 6, seed = 1195075742, epochs = 3.
- MenTaiBERT: learning rate = 10^{-5} , batch_size = 8, num_train_epochs = 8, k-fold = 7, optimizer = Adam.

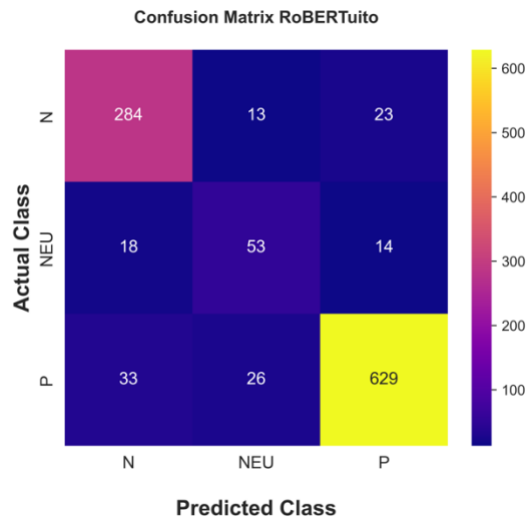
Comparing the three algorithms in Table II, it can be observed that MentaiBERT demonstrates a clear superiority over RoBERTuito and GPT3.5 Turbo across all metrics and classes. This improved performance is due to the specialised design of MentaiBERT, which includes an additional classifier layer. In the P class, MentaiBERT achieves a very good F1 score of 99%, beating RoBERTuito's 93.56% and GPT-3.5 Turbo's 94%. For the N class, MentaiBERT's F1-Score of 98% notably exceeds RoBERTuito's 87.51% and GPT-3.5 Turbo 89%. In the Neutral class, MentaiBERT shows the most pronounced advantage with an F1-Score of 91%, far outperforming RoBERTuito's 56.47% and GPT-3.5 Turbo 56%. Overall, MentaiBERT achieves the highest global accuracy at 98%, compared to RoBERTuito's 89% and GPT-3.5 Turbo 90%. Therefore, MentAI keeps high accuracy and recall consistently across all classes, even in the challenging Neutral category, indicating that the model does not simply memorise data, but captures real patterns that ensure robust performance with new data and high generalisability. In contrast, RoBERTuito and GPT-3.5 Turbo show significant drops in Recall and F1-Score, particularly in the Neutral category, revealing a lower generalisation ability in more complex classes.

TABLE II
RESULTS OF THE PRECISION, RECALL AND F1-SCORE METRICS FOR POLARITY

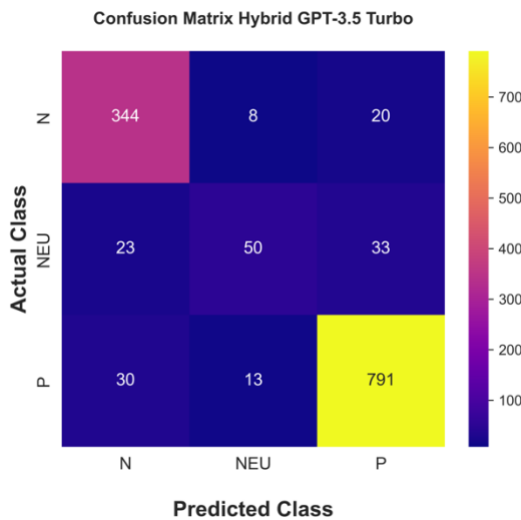
Polarity	Metrics	RoBERTuito	GPT3.5 Turbo	MentaiBERT
Positive (P)	Precision	94%	94%	99%
	Recall	93%	95%	99%
	F1-Score	93%	94%	99%
Negative (N)	Precision	86%	87%	98%
	Recall	89%	92%	99%
	F1-Score	87%	89%	99%
Neutral (NEU)	Precision	63%	70%	97%
	Recall	53%	47%	96%
	F1-Score	56%	56%	97%
Global		89%	90%	98%

The same performance can be observed in the confusion matrixes of the three algorithms in Fig. 4 (a), (b) and (c). Comparing RoBERTuito and GPT3.5 Turbo, a similar behaviour is observed when detecting some comments that are actually Positive as Negative. Finally, it can be observed that all three algorithms show the worst performance in detecting the Neutral class, which is normal because of the difficulty of clear language patterns in this more ambiguous class.

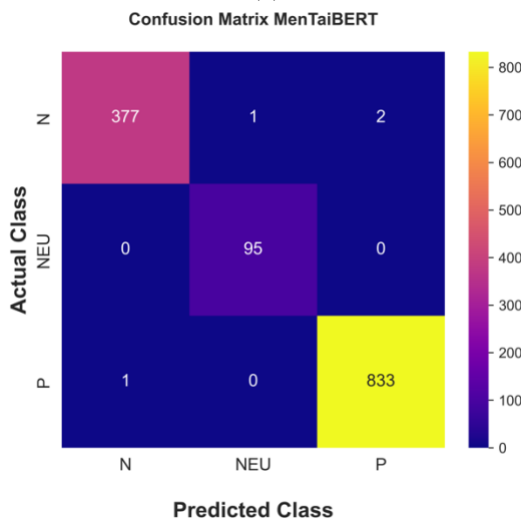
> REPLACE THIS LINE WITH YOUR MANUSCRIPT ID NUMBER (DOUBLE-CLICK HERE TO EDIT) <



(a)



(b)



(c)

Fig. 4. Confusion matrix of algorithms: (a) RoBERTuito (b) GPT3.5-Turbo (c) MenTaiBERT on polarity detection.

VII. GRAPHICAL INTERFACE

A software tool called MenT-AI has been developed in Python, integrating the previously mentioned BERT models. This tool offers an intuitive and efficient user experience, allowing users to input comments to obtain individual predictions, as well as upload data in tabular format (i.e., spreadsheet format), providing quick and easy polarity and emotion predictions in the same format.

As can be seen in Fig. 5, the screen has two clearly differentiated parts thanks to the continuous line that separates them. In the upper part you can enter a comment by keyboard and get its prediction. In the lower part there are buttons to load and process a file containing comments from social networks and to obtain their emotional prediction. If we focus on the top section, there is a text box labelled *Insert a Comment*, where the user can type a phrase. To have the model make a prediction, the user must click the *Run Algorithm* button. Immediately afterwards, the predicted emotion or polarity (depending on the tab being used) will appear in the text box labelled *Comment Prediction*.



Fig. 5. Appearance of the software application MenT-AI.

At the bottom of the screen are the buttons for loading and predicting a file (in csv/excel format). When this button is clicked, a pop-up window is displayed on the screen to allow you to select the file you want to load. When the file has been successfully loaded, the *Run Algorithm* button is enabled. By clicking on it, two graphs will be displayed in the interface, as it can be seen in Fig. 5. These graphs show both the overall distribution of emotions/polarity in a bar chart (in percentage) and the evolution of the emotional response to each comment. At this point, the *Download Excel/CSV* button would be enabled and allow us to download the prediction results obtained in a file (in csv/excel format). This file contains the *Comments* column, which corresponds to the comment of a post, and the *Prediction* column, which contains the predicted emotion or polarity for that message. It will also have as many rows as the number of comments to be predicted. This visual representation is especially useful for researchers and practitioners who want to understand the ranking patterns in the data of different social media posts.

VIII. LIMITATIONS

In this section we would like to highlight some limitations of the study. Firstly, the analysis of feelings has a subjective

> REPLACE THIS LINE WITH YOUR MANUSCRIPT ID NUMBER (DOUBLE-CLICK HERE TO EDIT) <

component that is impossible to eliminate, just as emotions are a complex construct, full of nuances. For example, some comments labelled as empathy may imply understanding and identification with the person, but may also be accompanied by sadness, by putting oneself in the other person's place. These double meanings are impossible to eliminate when studying complex human reactions, although we believe that our analysis and research does show general response trends that are valid for understanding the overall impact of celebrity social media posts. Consequently, this research is neither conclusive nor definitive, but has been designed to be flexible and to allow for inclusion and feedback with new comments, as well as different social networks or even the integration of a wider range of emotions in future studies. On the other hand, cultural, social, or gender factors may influence the emotional response of individuals. The gender of the person making the disclosure could be relevant to the emotional response it triggers, as well as the gender of the person responding. Is it the same for a man to reveal a mental health issue as it is for a woman? There are likely to be differences in the responses they receive depending on their gender or cultural factors. Furthermore, another limitation is the lack of data on the geographical origin of the commentators, which prevents a more detailed exploration of how cultural norms might influence emotional responses. In fact, previous existing research has shown significant cross-cultural differences in Internet use and psychological factors between Spain and Latin America, highlighting the need to account for these variations in studies with Spanish-speaking populations [69]. Incorporating metadata related to geographic origin in future research could provide valuable insights into how regional and cultural contexts shape emotional interactions on platforms like Instagram and TikTok. In this regard, this set of case studies with a high social impact represent a particularly promising line of research, which we will focus on in the future using the AI tools developed in this study. Another future direction will involve integrating additional techniques, such as data processing and learning strategies, to further enhance the study's scope and effectiveness.

IX. CONCLUSIONS

This research makes three crucial contributions that address some challenges related to mental health issues on social networks by integrating machine learning and natural language processing techniques. First, the study extends a previous corpus labeled with emotions and polarity by including new comments from Instagram and TikTok posts related to celebrity and influencer disclosures on mental health, and by also expanding the data to include posts made by men. This corpus is the first Spanish corpus designed to analyze the impact of social responses to mental health narratives on two of the most used social networks, particularly among young people.

Second, the research applies LLM-based classification models to enhance the efficient detection of emotions and polarity in these virtual environments. The three algorithms demonstrated high accuracy in detecting emotions and polarities, with MenTaiBERT, a BERT model with a specific classification layer, outperforming both RoBERTa and GPT-3.5 Turbo. In fact, MenTaiBERT outperforms in emotion detection with 99% accuracy, compared to GPT-3.5 Turbo's

88% and RoBERTa's 85.5%. For polarity, it also leads with 98% accuracy, surpassing RoBERTa's 89% and GPT-3.5 Turbo's 90%.

Finally, a user-friendly software tool named MenT-AI has been designed in Python to offer an intuitive and efficient user experience. This tool allows users to input comments for individual predictions and to load data in tabular format to easily and visually obtain emotional predictions in the same format. This software tool and the modelled AI algorithms are particularly useful for researchers and practitioners who want to understand patterns and behaviors in posts related to mental health disclosures and narratives on social networks used by young people. Specifically, these software tools can be used to conduct analyses of various use cases related to emotional responses, and how these might vary depending on factors such as the gender of the public figures making social media posts, the types of posts (text, video, songs, text with video), as well as other relevant social components. For example, it will be possible to explore how emotional responses may be influenced by aspects such as cultural identity, age, or the social context of the followers, which will allow for a more comprehensive analysis of the emotional effects of social media posts. It is important to note that the corpus, the categorization decalogue, and the LLM-based classification models are openly accessible in a GitHub repository for researchers.

In summary, analyzing the emotional impact generated by celebrity posts on social media platforms such as Instagram and TikTok is crucial, especially among young people, as these platforms can significantly influence their self-esteem, perception of reality, and emotional well-being.

ACKNOWLEDGMENT

This research was supported by these projects: POPTEC Equisam project (ref. 0298_EQUISAM_2_E). All authors have contributed equally to the research..

REFERENCES

- [1] DataReportal (2024, February 21). Digital 2024: Spain [Online] Available: <https://datareportal.com/reports/digital-2024-spain>
- [2] M. Asper, W. Osika, C. Dalman, E. Pöllänen, O. Simonsson, P. Flodin, A. Sidorchuk, L. Marchetti, F. Awil, R. Castro, and M. E. Niemi, "Effects of the COVID-19 pandemic and previous pandemics, epidemics and economic crises on mental health: systematic review," *BJPsych Open*, vol. 8, no. 6, Dec. 2022, Art. no. e181, doi: 10.1192/bjo.2022.587.
- [3] World Health Organization. 2021. Mental health of adolescents. Available at: <https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health> (accessed 18 October 2023).
- [4] Confederación Salud Mental España y Fundación Mutua Madrileña. 2023. La situación de la salud mental en España
- [5] Serrano-Gimeno, V., Diestre, A., Agustín-Alcain, M., Portella, M. J., de Diego-Adeliño, J., Tiana, T., Cheddi, N., Distefano, A., Dominguez, G., Arias, M., Cardoner, V., Puigdemont, D., Perez, V., & Cardoner, N. (2024). Non-fatal suicide behaviours across phases in the COVID-19 pandemic: a population-based study in a Catalan cohort. *The lancet. Psychiatry*, 11(5), 348–358. [https://doi.org/10.1016/S2215-0366\(24\)00065-8](https://doi.org/10.1016/S2215-0366(24)00065-8)
- [6] E. S.-T. Wang and Y.-T. Liao, "Contribution of internet celebrities' self-disclosure to fan-perceived interpersonal attraction and enduring involvement," *Comput. Human Behav.*, vol. 140, Jan. 2023, Art. no. 107601, doi: 10.1016/j.chb.2022.107601.
- [7] J. Klosternann, M. Meißner, A. Max, and R. Decker, "Presentation of celebrities' private life through visual social media," *J. Bus. Res.*, vol. 156, Jan. 2023, Art. no. 113524, doi: 10.1016/j.jbusres.2022.113524.
- [8] D. B. Francis, "Twitter is really therapeutic at times": Examination of Black men's Twitter conversations following hip-hop artist Kid Cudi's

> REPLACE THIS LINE WITH YOUR MANUSCRIPT ID NUMBER (DOUBLE-CLICK HERE TO EDIT) <

- depression disclosure, *Health Commun.*, vol. 36, no. 4, pp. 448-456, 2021, doi: 10.1080/10410236.2019.1700436.
- [9] P. C. Gronholm and G. Thornicroft, "Impact of celebrity disclosure on mental health-related stigma," *Epidemiol. Psychiatr. Sci.*, vol. 31, Art. no. e62, 2022, doi: 10.1017/S2045796022000488.
- [10] M. Withers, T. Jahangir, K. Kubasova, and M-S. Ran, "Reducing stigma associated with mental health problems among university students in the Asia-Pacific: A video content analysis of student-driven proposals," *International Journal of Social Psychiatry*, vol. 68, Art. No. 4, pp. 827-835, 2021 <https://doi.org/10.1177/00207640211007511>
- [11] A. Zsila, G. Orosz, L.E. McCutcheon, and Z. Demetrovics, "Individual Differences in the Association Between Celebrity Worship and Subjective Well-Being: The Moderating Role of Gender and Age," *Frontiers in Psychology*, vol. 12, 2021, <https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2021.651067>
- [12] D.B. Fransen, "The celebrityization of self-care: The celebrity health narrative of Demi Lovato and the sickscape of mental illness," *European Journal of Cultural Studies*, Vol. 23, Art. No. 1, pp. 89-111, 2019, <https://doi.org/10.1177/1367549419861636>
- [13] E. Cambria and B. White, "Jumping NLP curves: A review of natural language processing research," *IEEE Comput. Intell. Mag.*, vol. 9, no. 2, pp. 48-57, May 2014.
- [14] T. Wolf et al., "Transformers: State-of-the-art natural language processing," arXiv preprint arXiv:1910.03771, 2020.
- [15] M. V. Mäntylä, D. Graziotin, and M. Kuuttila, "The Evolution of Sentiment Analysis - A Review of Research Topics, Venues, and Top Cited Papers," *Comput. Sci. Rev.*, vol. 27, pp. 16-32, 2016, doi: 10.1016/j.cosrev.2017.10.002.
- [16] M. Malgaroli et al., "Natural language processing for mental health interventions: a systematic review and research framework," *Transl. Psychiatry*, vol. 13, no. 1, p. 309, 2023, doi: 10.1038/s41398-023-02592-2.
- [17] N. K. Iyortsuun et al., "A review of machine learning and deep learning approaches on mental health diagnosis," *Healthcare*, vol. 11, no. 3, p. 285, Jan. 2023, doi: 10.3390/healthcare11030285.
- [18] A. Khan and R. Ali, "Unraveling minds in the digital era: a review on mapping mental health disorders through machine learning techniques using online social media," *Soc. Netw. Anal. Min.*, vol. 14, no. 1, pp. 1-33, 2024, doi: 10.1007/s13278-024-01205-0.
- [19] A. Le Glaz et al., "Machine learning and natural language processing in mental health: systematic review," *J. Med. Internet Res.*, vol. 23, no. 5, p. e15708, 2021, doi: 10.2196/15708.
- [20] N. H. Di Cara et al., "Methodologies for monitoring mental health on Twitter: systematic review," *J. Med. Internet Res.*, vol. 25, p. e42734, 2023, doi: 10.2196/42734.
- [21] N. Merayo, A. Ayuso-Lanchares, and C. González-Sanguino, "Machine learning and natural language processing to assess the emotional impact of influencers' mental health content on Instagram," *PeerJ Comput. Sci.*, vol. 10, p. e12251, 2024, doi: 10.7717/peerj-cs.2251.
- [22] GitHub - GCODeveloper/Mental-Health-Dataset. n.d. Available: <https://github.com/GCODeveloper/Mental-Health-Dataset> (accessed Jul. 27, 2024).
- [23] T. Nguyen, D. Phung, B. Dao, S. Venkatesh, M. Berk, "Affective and Content Analysis of Online Depression Communities," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 217-226, 2014, doi: 10.1109/TAFFC.2014.2315623.
- [24] A. Malhotra and R. Jindal, "Xai transformer-based approach for interpreting depressed and suicidal user behavior on online social networks," *Cogn. Syst. Res.*, vol. 84, p. 101186, 2024, doi: 10.1016/j.cogsys.2023.101186.
- [25] I. Levkovich, M. Omar, "Evaluating of BERT-based and Large Language Mod for Suicide Detection, Prevention, and Risk Assessment: A Systematic Review," *J Med Syst*, vol. 48, Art.113, 2024, doi: <https://doi.org/10.1007/s10916-024-02134-3>
- [26] J. Gorai, D.K. Shaw, "A BERT-encoded ensembled CNN model for suicide risk identification in social media posts," *Neural Comput & Applic.*, vol. 36, pp. 10955-10970, 2024, doi: <https://doi.org/10.1007/s00521-024-09642-w>
- [27] S. T. Rabani et al., "Detecting suicidality on social media: Machine learning at rescue," *Egypt. Inform. J.*, vol. 24, no. 2, pp. 291-302, 2023, doi: 10.1016/j.eij.2023.04.003.
- [28] A. M. Schoene, L. Bojanić, M. -Q. Nghiem, I. M. Hunt and S. Ananiadou, "Classifying Suicide-Related Content and Emotions on Twitter Using Graph Convolutional Neural Networks," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 1791-1802, 2023, doi: 10.1109/TAFFC.2022.3221683.
- [29] L. Cao, H. Zhang, X. Wang and L. Feng, "Learning Users Inner Thoughts and Emotion Changes for Social Media Based Suicide Risk Detection," in *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1280-1296, 2023, doi: 10.1109/TAFFC.2021.3116026.
- [30] A. M. Schoene, A. P. Turner, G. De Mel and N. Dethlefs, "Hierarchical Multiscale Recurrent Neural Networks for Detecting Suicide Notes," in *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 153-164, 2023, doi: 10.1109/TAFFC.2021.3057105.
- [31] K. M. Hasib et al., "Depression detection from social networks data based on machine learning and deep learning techniques: An interrogative survey," *IEEE Trans. Comput. Soc. Syst.*, vol. 10, no. 4, pp. 1568-1586, 2023, doi: 10.1109/TCSS.2023.3263128K.
- [32] S. Khan and S. Alqahtani, "Hybrid machine learning models to detect signs of depression," *Multimed. Tools Appl.*, pp. 1-19, 2024, doi: 10.1007/s11042-023-16221-z.
- [33] B. Rohit, and S. Pavi, "A Hybrid BERT-CNN Approach for Depression Detection on Social Media Using Multimodal Data," *The Computer Journal*, Vol. 67, no. 7, pp. 2453-2472, 2024, doi: <https://doi.org/10.1093/comjnl/bxae018>.
- [34] S. D. Pande et al., "Depression detection based on social networking sites using data mining," *Multimed. Tools Appl.*, vol. 83, no. 9, pp. 25951-25967, 2024, doi: 10.1007/s11042-023-16564-7.
- [35] Z. N. Vasha et al., "Depression detection in social media comments data using machine learning algorithms," *Bull. Electr. Eng. Inform.*, vol. 12, no. 2, pp. 987-996, 2023.
- [36] L. Ilias, S. Mouzakitis and D. Askounis, "Calibration of Transformer-Based Models for Identifying Stress and Depression in Social Media," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 2, pp. 1979-1990, 2024, doi: 10.1109/TCSS.2023.3283009.
- [37] A. Pourkeyvan, R. Safa and A. Sorourkhah, "Harnessing the Power of Hugging Face Transformers for Predicting Mental Health Disorders in Social Networks," *IEEE Access*, vol. 12, pp. 28025-28035, 2024, doi: 10.1109/ACCESS.2024.3366653.
- [38] A. Wongkoblap, M. A. Vellido, and V. Curcin, "Deep learning with anaphora resolution for the detection of tweeters with depression: Algorithm development and validation study," *JMIR Ment. Health*, vol. 8, no. 8, p. e19824, 2021, doi: 10.2196/19824.
- [39] L. Tong et al., "Cost-Sensitive Boosting Pruning Trees for Depression Detection on Twitter," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 1898-1911, 2023, doi: 10.1109/TAFFC.2022.3145634.
- [40] A. Tyagi, V. P. Singh, and M. M. Gore, "Towards artificial intelligence in mental health: a comprehensive survey on the detection of schizophrenia," *Multimed. Tools Appl.*, vol. 82, no. 13, pp. 20343-20405, 2023, doi: 10.1007/s11042-022-13809-9.
- [41] M. E. Aragón, A. P. López-Monroy, L. C. González-Gurrola and M. Montes-y-Gómez, "Detecting Mental Disorders in Social Media Through Emotional Patterns - The Case of Anorexia and Depression," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 211-222, 2023, doi: 10.1109/TAFFC.2021.
- [42] W. C. de Melo, E. Granger and A. Hadid, "A Deep Multiscale Spatiotemporal Network for Assessing Depression From Facial Dynamics," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1581-1592, 2022, doi: 10.1109/TAFFC.2020.3021755.
- [43] M. Bishay, P. Palasek, S. Priebe and I. Patras, "SchNet: Automatic Estimation of Symptoms of Schizophrenia from Facial Behaviour Analysis," *IEEE Transactions on Affective Computing*, vol. 12, no. 4, pp. 949-961, 2021, doi: 10.1109/TAFFC.2019.2907628.
- [44] M. Kavaklı, "Why do we have emotions? The social functions of emotions," *Research on Education and Psychology (REP)*, Art. 3, no. 1, pp. 11-20, 2019, doi: <https://doi.org/10.1080/026999399379168>.
- [45] M. Schreiner, T. Fischer, and R. Riedl, "Impact of content characteristics and emotion on behavioral engagement in social media: Literature review and research agenda," *Electronic Commerce Research*, Vol. 21, no. 3, pp. 329-345, 2019, doi: <https://doi.org/10.1007/s10660-019-09353-8>.
- [46] S. Sciarra, D. Villani, A.F. Di Natale, and C. Regalia, "Gratitude and social media: A pilot experiment on the benefits of exposure to others' grateful interactions on Facebook," *Frontiers in Psychology*, Vol. 12, pp. 667052, 2021, doi: <https://doi.org/10.3389/fpsyg.2021.667052>.
- [47] J.E. Stellar, A.M. Gordon, P.K. Piff, D. Cordaro, Y. Bai, C.L. Anderson, and D. Keltner, D. "Self-transcendent emotions and their social functions: Compassion, gratitude, and awe bind us to others through prosociality,"

> REPLACE THIS LINE WITH YOUR MANUSCRIPT ID NUMBER (DOUBLE-CLICK HERE TO EDIT) <

- Emotion Review, Vol. 9, no. 3, pp. 200–207, 2017, doi: <https://doi.org/10.1177/1754073916684557>.
- [48] Y. H. Lee, C. W. Yuan, and D. Y. Wahn, "How video streamers' mental health disclosures affect viewers' risk perceptions," *Health Commun.*, vol. 36, pp. 1931–1941, 2021, doi: 10.1080/10410236.2020.1808405.
- [49] M. A. Alvarez-Mon et al., "Assessment of antipsychotic medications on social media: Machine learning study," *Front. Psychiatry*, vol. 12, p. 737684, 2021, doi: 10.3389/FPSYT.2021.737684.
- [50] S. Jilka et al., "Identifying schizophrenia stigma on Twitter: a proof of principle model using service user supervised machine learning," *Schizophrenia*, vol. 8, pp. 1–8, 2022, doi: 10.1038/s41537-021-00197-6.
- [51] N. Oscar et al., "Machine learning, sentiment analysis, and tweets: An examination of Alzheimer's disease stigma on Twitter," *The Journals of Gerontology, Series B: Psychological Sciences*.
- [52] S. Bograd, B. Chen, and R. Kavuluru, "Tracking sentiments toward fat acceptance over a decade on Twitter," *Health Informatics J.*, vol. 28, no. 1, pp. 14604582211065702, 2022, doi: 10.1177/14604582211065702.
- [53] Xue, J., Chen, J., Hu, R., Chen, C., Zheng, C., Su, Y., & Zhu, T., "Twitter Discussions and Emotions About the COVID-19 Pandemic: Machine Learning Approach," *Journal of Medical Internet Research*, vol. 22, no. 11, pp. e20550, 2020, doi: 10.2196/20550.
- [54] One Click Comment Extractor for IG - Chrome Web Store. n.d. Available at: <https://chrome.google.com/webstore/detail/comment-exporter/ckkachhlpdnnmchlhaepfcmhmadmpbgp> (accessed 18 October 2023).
- [55] A. A. Mangino, K. A. Smith, and W. H. Finch, "Modeling Responsibly Toward a Fair, Interpretable, and Ethical Machine Learning for the Social Sciences," in *TMS Proceedings 2021*, PubPub, Nov. 2021.
- [56] S. Delany, F. Benamara, V. Moriceau, F. Olivier, y J. Mothe, "Psychiatry on Twitter: Content Analysis of the Use of Psychiatric Terms in French," *JMIR Formative Research*, vol. 6, pp. e18539, 2022, doi: 10.2196/18539.
- [57] P. Ekman, "Ekman, P. (2004, April). What we become emotional about. In Feelings and emotions: The Amsterdam symposium pp. 119–135, 2024.
- [58] L. Williams, M. Arribas-Ayllon, A. Artemiou, and I. Spasić, "Comparing the utility of different classification schemes for emotive language analysis," *Journal of Classification*, vol. 36, pp. 619–648, 2019, <https://doi.org/10.1007/s00357-019-9307-0>.
- [59] R. Plutchik, "What is an emotion," *The Journal of psychology*, Vol. 61, no. 2, pp. 295–303, 1965.
- [60] B.L. Fredrickson, "Positive emotions broaden and build," *Advances in experimental social psychology*, Vol. 47, pp. 1–53, 2013, <https://doi.org/10.1016/B978-0-12-407236-7.00001-2>.
- [61] A. Vaswani et al., "Attention is all you need," *Advances in neural information processing systems*, pp. 5998–6008, 2017.
- [62] Y. Liu et al., "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
- [63] ROBERTuito. 2022. Available at: <https://huggingface.co/pysentimiento/robertuito-emotion-analysis>, (accessed 10.18.23).
- [64] OpenAI. Available at: <https://platform.openai.com/docs/models>
- [65] Hugging Face. Available at: <https://huggingface.co/google-bert/bert-large-uncased>
- [66] C. M. Bishop, "Pattern recognition and machine learning," Springer, 2006
- [67] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," *arXiv preprint arXiv:1801.06146*, 2018
- [68] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *Journal of machine learning technologies*, vol. 2, no. 1, pp. 37–63, 2011.
- [69] M. Varcheta, G. Tagliaferri, E. Mari, A. Quaglieri, C. Cricenti, A.M. Giannini, and M. Martí-Vilar, "Exploring Gender Differences in Internet Addiction and Psychological Factors: A Study in a Spanish Sample", *Brain Sciences*, Vol. 14, Art. No. 10, pp.1037, 2024 <https://doi.org/10.3390/brainsci14101037>.

Fellow at the University of Hertfordshire in the Optical Networks Group, Science and Technology Research Institute, the TOyBA research group (University of Zaragoza), and the Technology University of Munich (TUM). Her research focuses on the design and performance evaluation of optical networks and the application of artificial intelligence techniques in multidisciplinary fields such as mental health, video games and optical networks.



Alba Ayuso-Lanchares was awarded her PhD in October 2021 from the University of Valladolid (Spain). She earned a degree in Speech Therapy from the University of Valladolid in 2014 and a master's degree in Neuropsychology and Education from the University of La Rioja in 2016. Since 2014, she is an Assistant Professor with the Department of Pedagogy at the University of Valladolid and a Collaborating Professor at the University of La Rioja since 2017. Her research interests include language development studies, speech therapy interventions, and education. Ms. Ayuso is a member of the Association of Speech Therapists of Spain.



Clara González-Sanguino, PhD in Psychology with international mention (2021), currently Lecturer at the Faculty of Education and Social Work at the University of Valladolid (UVA). Mental health and stigma are her main line of research, along with psychological assessment, with several scientific contributions of impact in the area with national and international authors. She has collaborated in several research projects and is member of the University Chair Against Stigma Universidad Complutense de Madrid (UCM)-Group 5, and the recognized research groups of Evaluation and Psychological Research in Mental Health and Society of the UCM, and Psychology, Health and Neuroeducation of the UVA.



Noemí Merayo received the Telecommunication Engineer degree from the Valladolid University, Spain, in February 2004 and the Ph.D. degree at the same university, in July 2009. She works as Lecturer at the Universidad de Valladolid. She has also been a Visiting Research