Detection of Depression Symptoms on Social Networks through Machine Learning and Facial Analysis: An Integrated Approach

Detección de síntomas de depresión en Redes Sociales mediante Machine Learning y Análisis Facial: un enfoque integrado

DOI: http://dx.doi.org/10.17981/cesta.07.01.2026.03
Scientific research article
Date received: May 04, 2026, Date accepted: June 04, 2026.

Javier Antonio Ballesteros Ricaurte
Universidad Pedagógica y Tecnológica de Colombia (Colombia)
javier.ballesteros@uptc.edu.co

How to cite:
J. Ballesteros Ricaurte, “Detection of Depression Symptoms on Social Networks through Machine Learning and Facial Analysis: An Integrated Approach”, J. Comput. Electron. Sci.: Theory Appl., vol. 7, no. 1, pp. 20-34, 2026. DOI: 10.17981/cesta.07.01.2026.03

Abstract

Introduction: The early detection of depressive symptoms is a critical challenge in mental health, particularly due to the limitations of traditional diagnostic tools and the growing relevance of digital platforms as sources of behavioral and emotional information.
Objective: To develop an automated system for detecting depressive indicators through the analysis of Instagram content, integrating web scraping techniques, natural language processing, facial emotion recognition, and machine learning algorithms.
Method: A computational approach was employed, combining textual classification and facial emotion analysis. For the textual component, Naive Bayes, Logistic Regression, and Random Forest algorithms were evaluated using a dataset of approximately 200,000 text records from Reddit, which was used exclusively for training and evaluating the text classifier. Facial emotion analysis was performed independently on images extracted from the analyzed Instagram profiles. The system was implemented through a Streamlit interface.
Results: The results showed that the Naive Bayes model achieved the best performance for textual classification, with an accuracy of 90% and high recall in detecting depressive indicators. The integration of text analysis and facial emotion recognition allowed the system to compensate for the limitations of each method when used separately and to generate both individual and global reports.
Conclusions: The proposed system contributes to the development of non-invasive tools for the early detection of depressive symptoms, supporting timely interventions and demonstrating the potential of combining machine learning, natural language processing, and facial emotion recognition in digital mental health contexts.

Keywords

Depression; Artificial Intelligence; Instagram; Machine Learning; Facial Recognition; Web Scraping.

Resumen

Introducción: La detección temprana de síntomas de depresión representa un desafío relevante en salud mental, debido a las limitaciones de las herramientas diagnósticas tradicionales y al uso creciente de plataformas digitales como fuentes de información conductual y emocional.
Objetivo: Desarrollar un sistema automatizado para detectar indicadores depresivos mediante el análisis de contenido de Instagram, integrando web scraping, procesamiento de lenguaje natural, reconocimiento facial de emociones y algoritmos de aprendizaje automático.
Método: Se empleó un enfoque computacional que combina clasificación textual y análisis de emociones faciales. Para el componente textual, se evaluaron Naive Bayes, Regresión Logística y Random Forest con un conjunto de aproximadamente 200.000 registros de Reddit, usado para entrenar y evaluar el clasificador. El análisis facial se realizó de forma independiente sobre imágenes extraídas de perfiles de Instagram. El sistema se implementó mediante una interfaz en Streamlit.
Resultados: El modelo Naive Bayes obtuvo el mejor desempeño en la clasificación textual, con una precisión del 90% y alta capacidad para identificar indicadores depresivos. La integración del análisis textual y facial permitió compensar las limitaciones de cada método por separado y generar reportes individuales y globales.
Conclusiones: El sistema propuesto aporta al desarrollo de herramientas no invasivas para la detección temprana de síntomas depresivos, favorece intervenciones oportunas y evidencia el potencial de combinar aprendizaje automático, procesamiento de lenguaje natural y reconocimiento facial en salud mental digital.

Palabras clave

Depresión; Inteligencia Artificial; Instagram; Machine Learning; Reconocimiento Facial; Web Scraping.

INTRODUCTION

The early detection of affective disorders such as depression represents a major challenge, as commonly used clinical tools are often costly or invasive [1]. Today, social networks have become spaces where users routinely share their experiences and emotional states, making them valuable data sources for public health monitoring [2].

This study proposes the development of an automated system that uses social media data scraping techniques and machine learning algorithms to detect symptoms of depression. This system integrates textual content analysis and facial recognition to identify potential indicators of depressive states in a non-invasive and accessible manner.

The project is based on the application of technologies such as natural language processing, image analysis, and machine learning. Through the collection and analysis of social media posts, specifically on Instagram, the system seeks to identify linguistic patterns and facial expressions that may be associated with depressive symptoms.

The implementation of this system not only represents an achievement in the integration of multiple data sources for mental health analysis, but also offers a potential tool for the early detection of depressive symptoms. The system could serve as a complement for mental health professionals by providing additional data for their assessments and facilitating more timely interventions.

The objective of the project is to create an application capable of detecting symptoms of depression based on facial and textual analysis of posts from Instagram profiles. The analysis process was carried out using machine learning algorithms, scraping techniques, and the integration of facial analysis APIs.

THEORETICAL AND CONCEPTUAL FRAMEWORK

Depression is defined as an affective disorder characterized by symptoms such as depressed mood, anhedonia, sleep disturbances, and chronic fatigue, present for at least two weeks [5]. Anxiety, in turn, involves excessive and uncontrollable worry that often coexists with depression. This project uses these clinical foundations to characterize symptoms that may be detected through technology.

Machine Learning (ML) enables systems to learn patterns from historical data in order to make predictions [2]. In this context, web scraping is used as a technique for the automated collection of heterogeneous data from the web [6]. Instagram was selected as the object of study due to its predominantly visual nature, which facilitates psychological and facial analysis [3].

For the selection of the facial analysis API to be used in the scraping process, the following characteristics were considered: accuracy, image processing capacity, cost, compatibility with other tools, and ease of use. These criteria were taken into account in order to process large volumes of images and obtain the necessary data from them.

Table 1 presents a comparison of some of the main APIs for facial analysis, considering key aspects such as implementation capacity, accuracy, use, processing capacity, and cost. It should be noted that some APIs may have different consumption models.

TABLE 1. COMPARISON OF FACIAL ANALYSIS APIS.

API	Implementation Capacity	Accuracy	Use	Processing Capacity	Cost
Microsoft Computer Vision API	Compatible with Microsoft and Windows applications.	96%	Detection and analysis of biometric information.	1,000 to more than 1,000,000 images per month.	USD 0–1
INFERDO	Easy to integrate into medical data management applications and healthcare sector systems.	100%	Pre-trained deep learning models.	5,000 to 50,000 images per month.	USD 0–10
FACE++	Can be integrated into other APIs or programs to analyze images provided	99%	Useful for face comparison, detection, and search, as well as other	Approximately 10,000 images or a consumption rate of 2 queries per	USD 0–100

	through links.		physical traits.	second.	

For the scraping phase, Face++ will be used because it is easy to integrate into Python scripts and does not appear to limit the number of images during the trial period. However, it does limit concurrency and restricts processing to one image per second. The other alternatives will also be explored in case any issues arise or the trial period ends.

METHODOLOGY

This research adopts an applied methodology with a quantitative and computational approach, aimed at developing an automated system for detecting potential symptoms of depression on social networks. To this end, web scraping techniques, natural language processing, machine learning, and facial emotion analysis are integrated to examine both the textual and visual content published on Instagram profiles in a complementary manner. This methodological strategy makes it possible to evaluate different classification algorithms, process information from digital sources, and generate individual and global reports that support the early identification of indicators associated with depressive states.

A. Selection of Algorithms and Tools

Three classification approaches were compared: Naive Bayes, a probabilistic model; Logistic Regression, a linear model; and Random Forest, a tree-based model. For facial analysis, the Face++ API was selected due to its high accuracy of 99% and its ability to process multiple biometric traits, such as the degree of happiness.

It is important to note that the Face++ API quantifies the degree of happiness expressed in the analyzed face. In the context of this project, it is assumed that a low facial happiness score may serve as an indirect indicator of negative emotional states associated with depression. This assumption follows the premise, supported by the API’s own documentation and facial expression studies such as those by Ekman, that happiness is the emotion with the most recognizable facial expression and that its sustained absence may correlate with negative affective states.

However, it is acknowledged that this relationship is not necessarily direct or universal: a low expression of happiness does not automatically indicate a depressive state. Therefore, this indicator is used as a complementary signal within a multimodal analysis, rather than as an isolated diagnostic criterion.

B. Collection of Relevant Social Media Data, Including User Posts through Scraping

Absolute frequencies and relative proportions were calculated for each macro-area in both corpora, enabling the construction of a comparative thematic density matrix.

The Instagram data scraping process was carried out using the Instaloader library in Python. Each step involved in the process of downloading data from an Instagram profile and extracting traits or values related to happiness from the published images through an emotion detection API is described below.

To download an Instagram user profile, it is necessary to enter the profile name. Once it is entered, the following subprocess, shown in Figure 1, is performed to download a public Instagram profile using web scraping and the facial analysis API.

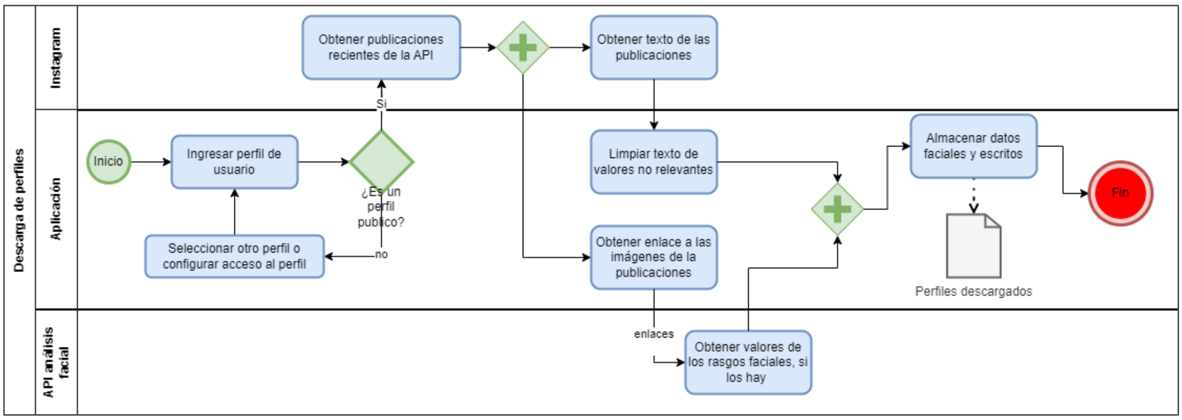

Figure 1. BPMN Diagram

Step 1: Preparation of the Work Environment

The first step is to prepare the environment where the data will be stored. The method verifies whether a folder named dataProfiles exists, where the CSV files will be stored. If it does not exist, it creates the folder using os.makedirs. This ensures that the file structure is ready to store the results of the analysis.

From a text file located in the dataProfiles folder, the access credentials for the Face++ API are loaded in order to use its services. It is recommended to create new credentials to avoid concurrency issues with simultaneous users.

Step 2: Loading the Instagram Profile

The next step is to load the Instagram profile of the provided user using the Instaloader tool, which allows public profile information to be obtained. The method verifies that the profile is not private and that it contains posts. If the profile does not exist or is configured as private, exceptions are raised to inform the user of the issue. This ensures that only accessible and valid profiles are processed for analysis.

Step 3: Creation of the CSV File

Once the profile has been validated, the method creates a CSV file to store the information downloaded from the posts. This file is essential for structuring and organizing the data that will later be processed.

The columns "Text," "Date," and "Happiness Percentage" are defined, structuring the analysis around these three key aspects of the study: the textual content, the time at which it was shared, and the emotional analysis of the images.

Step 4: Iteration over the Posts

The method extracts up to 100 posts from the profile in order to analyze them one by one. The number of posts was limited to 100 because some profiles contain a large volume of content, and only the most recent data were considered. During this iteration, several subprocesses are performed to prepare and store the information from each post.

The caption text of each post is obtained. This text is processed using the clean_text method, which removes emojis and special characters. If the text is empty after cleaning, it is replaced with the value "NotApplicable." This processing is relevant because the use of emojis on social media is common, but they are not useful for subsequent analysis.

The publication date is obtained and formatted in a standard way (year-month-day hour) so that it can be easily analyzed later.

Step 5: Analysis of Emotions in Images

For each post, the method obtains the image URL and then makes a call to the Face++ API to analyze the emotions present in the faces within the image.

The API returns a happiness percentage for each face detected in the image. In individual Instagram profiles, it is assumed that the main face of the profile owner is the one that appears most frequently and prominently in the posts. However, when multiple faces are detected in an image, the automatic selection of the face with the highest happiness level does not guarantee that it corresponds to the profile being studied, which represents a limitation of the method.

As an approximation to mitigate this issue, priority was given to the analysis of profiles in which the user is the predominant subject of the photographs, and the value "N/A" is recorded when it is not possible to identify the profile owner's face with certainty. This limitation is acknowledged as part of the scope of the study, and future work is encouraged to incorporate facial identification mechanisms based on biometric recognition linked to the profile.

This emotional analysis is central to the project, as it makes it possible to associate visual content with the emotional state expressed.

If errors occur due to the API's concurrency limit, the method waits one second and retries the request. This ensures that the emotional analysis remains robust, even under API restrictions. Although the API does not limit the number of images that can be processed, it does restrict highly concurrent use.

Step 6: Storage in the CSV File

Once the textual information, date, and happiness percentage have been processed, the method stores these data in the CSV file.

Each post is saved as a row in the CSV file, including the processed text, the publication date, and the happiness percentage detected in the image. If no face could be detected or the image could not be analyzed, “N/A” is recorded instead of the happiness percentage. This ensures that the CSV file contains a clear and complete representation of the analyzed posts.

This storage method was chosen over a database because it is easier to implement in a local or portable application, allowing for simpler installation and facilitating data access, as well as the application of an indefinite number of tests or symptom detection processes according to the user’s needs.

C. Integration and Use of the Facial Detection API

One of the data points to be extracted from Instagram posts is the happiness value in the published photographs. For this purpose, the Face++ facial analysis API was selected. This API allows a photograph to be analyzed free of charge and enables the extraction of values such as age, gender, beauty, ethnicity, and, most importantly, emotions. It also allows multiple faces to be analyzed simultaneously.

A Face++ account was created. Once the account was created, the facial recognition service was configured. This service provides an API key and secret key to access the service, as shown in Figure 2.

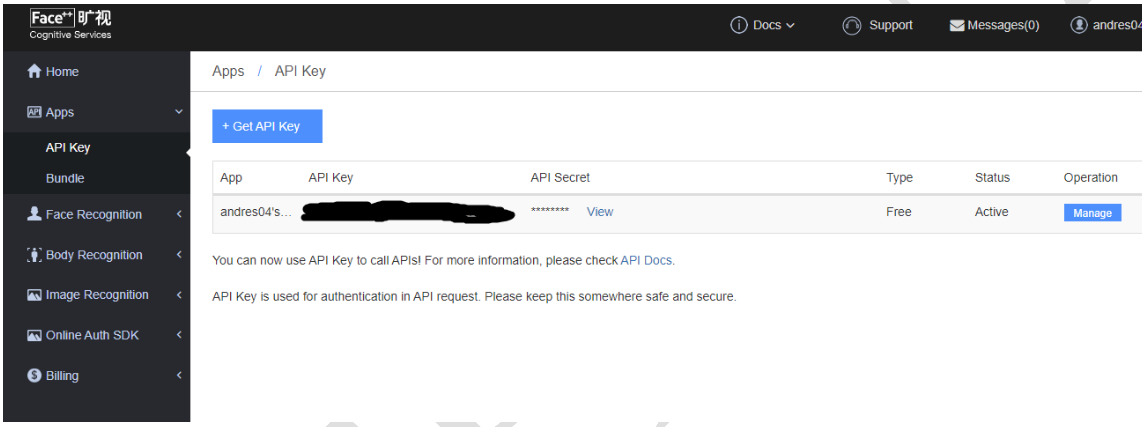

Figure 2. Obtaining New Face++ API Credentials

D. Training of Machine Learning Models

For the development of the machine learning model, a dataset downloaded from Kaggle was selected: https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch. This dataset consists of a file containing approximately 200,000 rows, with each row including the following information:

- Index: a numerical identifier.
- Text: a text string in English, extracted from Reddit posts, which may or may not be related to a person showing symptoms of depression.
- Class: the diagnosis or classification assigned to the text, with values of “suicidal” or “non-suicidal.”

The process consisted of training a machine learning model capable of classifying the text of a post as suicidal or non-suicidal, using this dataset as training data.

For each algorithm, similar scenarios were prepared. The same dataset was loaded, with 80% of the data used to train the model and the remaining 20% used for evaluation. The same training metrics were applied (precision, recall, and F1-score) along with the same preprocessing steps. The variables used were text, which represents the texts to be analyzed, and class, which indicates whether the text is suicidal or non-suicidal. Vectorization was performed using the Bag-of-Words method through scikit-learn’s CountVectorizer, followed by scaling. From that point, the corresponding model training method was applied for each algorithm.

Once the data had been loaded and the models trained, a comparison was performed using the defined metrics, with the aim of selecting the model that achieved the highest metric values compared with the others.

E. Evaluation of the Machine Learning Models

Once the models had been trained, they were evaluated using the test set in order to measure their performance. To do this, predictions were made on the test data (X_test) and compared with the actual labels (y_test). The results are shown in Table 2.

TABLE 2. RESULTS OF THE ALGORITHM COMPARISON

Algorithm	Accuracy	Precision	Recall	F1-Score	Support
Naive Bayes	0.8986	0.95	0.84	0.89	23287
Logistic Regression	0.8683	0.98	0.75	0.85	23287
Random Forest	0.8239	0.79	0.88	0.83	11675

For the interpretation of the metrics, it is important to clarify that the training dataset uses two classification classes: Class 0 corresponds to non-suicidal posts, that is, posts without indicators of depression or suicidal ideation, while Class 1 corresponds to suicidal posts, that is, posts with indicators of suicidal ideation or marked depressive symptoms. This distinction is essential for evaluating the performance of each model according to the objective of the project: to correctly detect positive cases, corresponding to Class 1.

Observations:

- Naive Bayes shows good performance, standing out for its high recall in Class 1 (0.84), which means that it correctly identifies most of the true positives in that class.
- Naive Bayes has high precision for Class 0 (0.95), indicating that it produces few false positives in that class.
- Logistic Regression has high precision for Class 0 (0.98), but its recall in that class is lower (0.75), indicating that although it predicts accurately when it classifies a case as Class 0, it does not capture all true cases of that class.
- Overall, Logistic Regression achieves good accuracy of 0.87, but it is less balanced compared with Naive Bayes.
- Random Forest has the lowest performance among the three algorithms, with an accuracy of 0.82.
- Random Forest is the algorithm with the lowest balance among the metrics, which may be due to the nature of the model or to a reduction in the sample size.

For this project, the Naive Bayes algorithm was implemented, as it obtained the highest metrics, especially in terms of precision and recall, which allow true positive cases to be detected more effectively.

F. Text Analysis Based on Keyword Detection

In addition to the analysis of posts through machine learning models, a method was implemented to identify keywords in order to specify some of the possible symptoms present in the posts. This function was added because, although the machine learning model can classify posts into categories such as “depressive” or “non-depressive,” it does not have the ability to identify specific causes or symptoms related to depression.

To identify the keywords, the following method was developed:

- First, the symptoms to be evaluated in this project were established. Four symptoms were initially defined as a basis, although more can be added if necessary.
- Frequent use of past-tense verbs: To identify these words, an English grammar library was added to the keyWordsDetector class.
- Mentions of pain or illness: A list of words related to the sensation of pain was created, including both physical pain, as a probable cause, and emotional pain. This type of symptom is mentioned in the DSM-5 Guide to the Diagnostic and Statistical Manual of Mental Disorders.
- Mentions of tiredness or lack of energy: A list of words related to tiredness, lack of sleep, or lack of energy was created. This type of symptom is mentioned in the DSM-5 Guide to the Diagnostic and Statistical Manual of Mental Disorders [2].
- Use of absolutist words: A list of absolutist terms was created, including words such as always, never, nobody, every time, among others. This symptom was examined in studies such as “Corrigendum: In an Absolute State: Elevated Use of Absolutist Words Is a Marker Specific to Anxiety, Depression, and Suicidal Ideation” [19].

After the lists of words and symptoms were created, they were stored in single-column CSV files in order to facilitate the insertion of symptoms by the end user. The only exception was past-tense verbs, for which an English grammar library was used.

G. Keyword Identification Tool

A class or file named keyWordsDetector was developed. This tool allows a list of texts from posts and a CSV file containing words associated with a specific symptom to be entered. Using the word_tokenize library, the tool verifies whether a word or phrase from the CSV file appears in the text of the post. If a match is found, an attribute with the name of the symptom is assigned a value of 1; otherwise, it is assigned a value of 0.

After this analysis is completed, the posts from the downloaded profile contain new attributes, each named after one of the symptoms. This makes it possible to generate individual or global reports for each profile and helps to better identify symptoms or their possible causes.

H. Interface Development

Once the profile download modules and the machine learning algorithms had been developed, the main module and interface were implemented. For this purpose, three user interface modules were designed: the profile download and preview module, the profile analysis module, and the global or general analysis module for all downloaded and analyzed profiles.

The profile analysis module allows the user to select a downloaded profile for preview. Once the profile is selected, the analysis option appears. When initiated, the module enters the text of the posts from that profile, one by one, into the trained machine learning algorithm for analysis. Then, the symptom or keyword analysis is performed, and finally, the profile results are displayed. These results include facial happiness over time, the outputs of the machine learning model, and the occurrence of symptoms based on keywords.

The global analysis module presents the results of all analyzed profiles in a manner similar to the individual profile analysis module, while also including additional statistical reports.

For the implementation of the main module and interface, the Streamlit framework was selected. Streamlit allows the development of simple and easily deployable Python-based web applications, especially for projects focused on data analysis and machine learning. Development in this framework is centered on a main file in which library functions, data analysis processes, and other elements are executed directly, similar to a simple Python script. This is combined with prebuilt user interface components, which facilitates the development of the web user interface and event handling.

The overall operation of the application is defined in the BPMN diagram shown in Figure 3, which describes, step by step, the process of downloading and analyzing profiles to generate a report on the emotions or possible symptoms found in their posts.

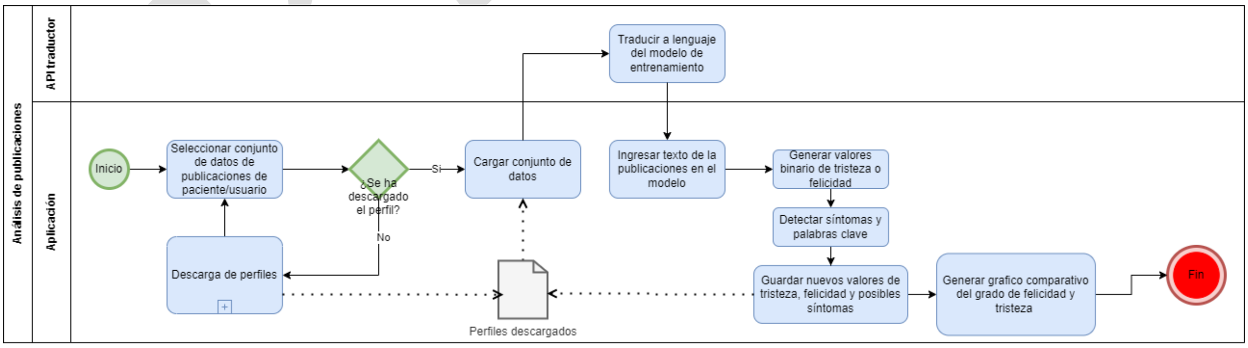

Figure 3. BPMN Diagram of the Application

- Dataset selection: Through the interface, one of the analyzed profiles is selected for preview.
- Dataset loading: The list of posts and their attributes is loaded for analysis, along with the trained machine learning model.
- Translation of posts: Before the text is entered into the algorithm, the posts are translated into English because the model was trained with English-language posts; therefore, it can only generate results based on posts in that language.

- Input of posts into the model: Once translated, the posts are entered into the model. This may take some time depending on the number of posts.
- Generation of values by the model: Once each post has been analyzed, a binary value, either 0 or 1, is generated and assigned to each post. A value of 1 indicates the presence of symptoms or patterns related to depression, while a value of 0 indicates their absence.
- Keyword detection: After the values have been generated using the machine learning algorithm, the different lists of keywords related to possible symptoms of depression are reviewed. The system verifies whether any word or phrase exists in the post. If a match is found, an attribute with the name of the symptom is assigned a value of 1; otherwise, it is assigned a value of 0.
- Saving of results: Once the new attributes have been generated, they are stored for future analysis.
- Generation of charts and results: Two types of reports are generated: a global or general report and an individual report. The individual report generates a chart of facial happiness recorded over time and another chart related to possible symptoms of depression in the text. It also includes a table showing the percentage of occurrence of words related to each symptom. For the general report, the previously analyzed profiles are used to generate a treemap chart showing the percentage of symptom occurrence, as well as another report related to facial happiness and textual sadness.

I. Application Architecture and Modules

In order to establish an efficient organization and ensure optimal operation of the application, the structure shown in Figure 4 was designed. The function of each module in the model is explained below.

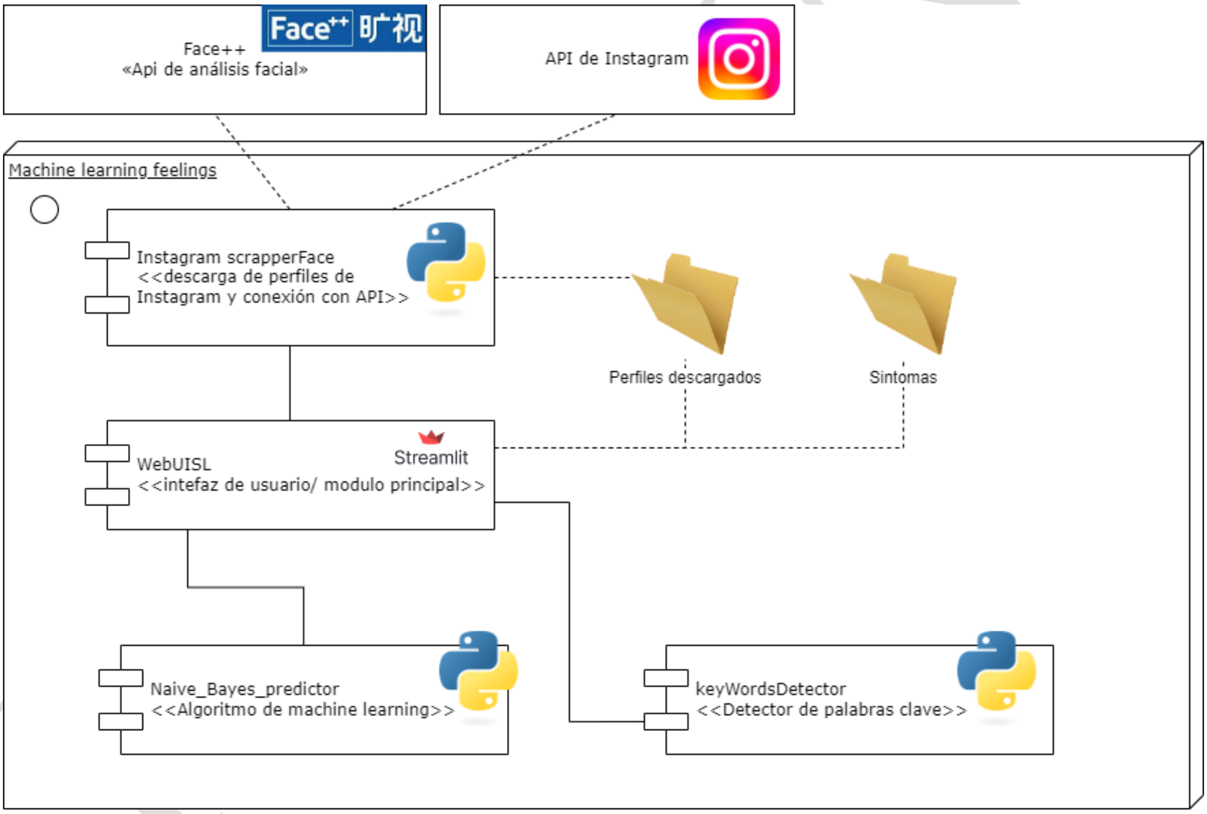

Figure 4. Application Architecture

- Instagram API: This is the official Instagram API, which allows direct access to information from public profiles. To access private profiles, the use of an official Instagram extension and user credentials within this social network is required.
- Face++ API: This is a set of applications and services dedicated to image analysis and biometric recognition. For this project, a free account was created on this platform to use a facial analysis API, which allows facial features, such as emotions expressed in faces, to be extracted when present. Unlike other facial analysis APIs, this one does not impose a limit on the number of images. Instead, it has a limit on queries per second. In other words, although it does not charge based on the total number of images per month, it does limit the number of images processed per second, which increases the time required to analyze a list of posts under the basic plan offered.

- InstagramScraperFace: This module is responsible for downloading public profiles from the Instagram API, provided that they exist and are accessible. It then sends the link of each image to the Face++ API and stores the results. After receiving the profile data and the results from the Face++ API, each profile is stored in CSV format for subsequent analysis.
- Naive Bayes Predictor: This module is responsible for loading the trained machine learning model and retraining it if necessary. It also receives a list of texts belonging to a profile, translates them into English, which is the language of the data used to train the model, and finally enters them into the model to generate results that are later stored.
- keyWordsDetector: Similar to the Naive Bayes Predictor module, this module receives a list of phrases or texts from a profile, translates them, and performs keyword analysis. It adds an attribute to each post with the name of each symptom to be analyzed, assigning a value of 1 if the post contains words related to that symptom, or 0 otherwise. In addition, a special check is performed for a symptom called PastSimple, where an English grammar library is used to detect whether a sentence contains verbs in the simple past tense.
- Downloaded Profiles: This is where downloaded or analyzed profiles are stored.
- Symptoms: This is where the list of keywords related to each symptom is stored in CSV format, so that they can be easily added by the end user.
- WewUISL: This is the main module, where the InstagramScraperFace, Naive Bayes Predictor, and keyWordsDetector modules are used, and where the user interface is generated using the Streamlit framework. It is responsible for handling events, loading certain data, generating charts and tables, and calling methods from the other modules.

RESULTS

Based on the execution of the application following the instructions of each module, the following results were obtained.

Downloaded Profiles: This option allows the user to enter the Instagram profile name of a user for download. Before starting the download, the system verifies whether the profile is public, private, or invalid. Once a profile has been downloaded, it can be viewed in the list, deleted, or previewed, as shown in Figure 5.

Figure 5. Example of the List of Downloaded Profiles

Profile Preview: When the preview option is selected, the list of recent posts from the profile can be viewed, including their text, publication date, and the happiness percentage detected through the Face++ API, as shown in Figure

Archivo seleccionado: 4ndres.g0nzalez04.csv

	Texto	Fecha	Porcentaje de felicidad
0	NoAplica	2024-09-08 22:23:28	0.121
1	Voy a publicar emoticones raros:	2024-09-08 18:21:19	None
2	NoAplica	2024-09-08 18:14:49	None
3	Nueva foto de perfil, no esta mal :)	2024-08-23 01:56:58	1.286
4	Me gusta este cuadro	2024-07-12 15:04:21	None
5	Estoy feliz, mi primera publicación	2024-06-16 00:24:18	99.993
6	Hoy estoy muy feliz acabo de crear mi cuentas	2024-06-16 00:20:54	99.993

Analizar Perfil

Figure 6. Preview of a Profile's Posts

Individual Profile Analysis: If the user clicks the analysis option, after a short period of time, a series of reports related to the posts of the loaded profile is generated. Examples of the analysis results are described below.

Facial Happiness Chart: A time-series chart is generated showing the percentage of happiness found in the photographs of the downloaded profile. This may vary depending on whether the user takes many photographs of other people or of themselves. If no faces are detected, no points are displayed on the chart, as shown in Figure 7.

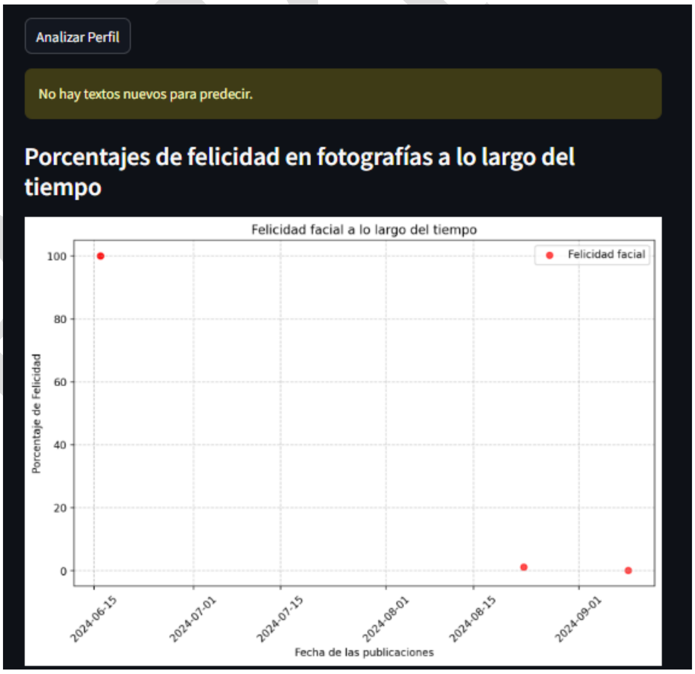

Figure 7. Facial Happiness Chart over Time

Facial Happiness Statistics: Figure 8 presents the statistics of the charted data, including the minimum and maximum happiness values found in the list of photographs, as well as the average value. It also presents the number of photographs corresponding to each percentage range; that is, how many photographs have a happiness percentage below 25%, how many are between 25% and 50%, how many are between 50% and 75%, and how many are between 75% and 100%.

Figure 8. Statistics of the Facial Happiness Obtained

Textual Sadness and Symptom Statistics: Figure 9 presents statistics such as the average value and the number of posts with sadness or depression according to the machine learning algorithm, in order to compare which state is more common. In addition, a table is generated indicating the percentage of occurrence of keywords associated with depression symptoms and the number of occurrences.

Figure 9. Statistics Related to Sadness and Percentage of Detected Symptoms.

A. Symptom Analysis Configuration

For both global and individual analysis, it is possible to add or remove symptoms. This can be done through the Symptom Analysis Configuration menu, which allows new symptoms to be added or deleted and provides a preview of the terms, as shown in Figure 10. In this case, a CSV file must be added with the name of the symptom. Inside the file, there must be one column containing the words or phrases to be analyzed.

For example, if the objective is to verify whether a profile frequently attends parties, a file named frequentParties can be created with a single column. Each row in that column should contain a related phrase or word. In this case, the terms must be written in English, since some methods require analysis using an English grammar library.

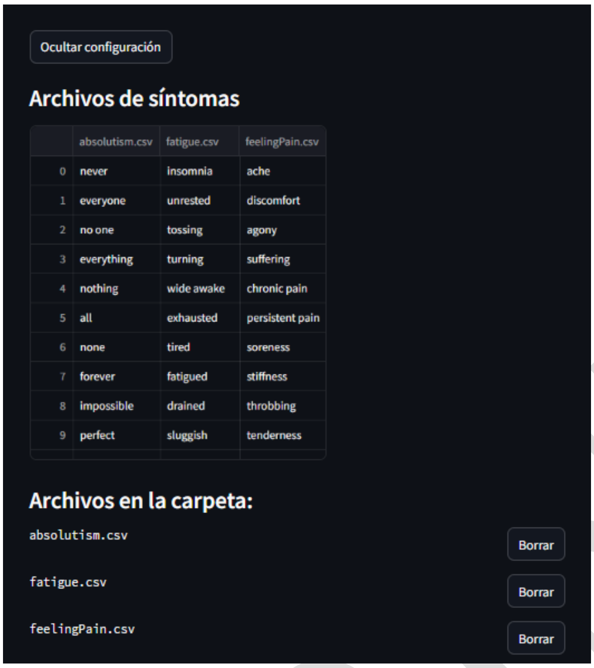

Figure 10. Symptom Analysis Configuration

CONCLUSIONS

Regarding facial analysis, the system uses the happiness percentage reported by the Face++ API as an indirect emotional indicator. This approach is based on the assumption, supported by the literature on facial expression, that the absence of happiness may be associated with negative affective states related to depression. However, this relationship is not deterministic: a low expression of happiness may be due to multiple contextual factors and does not constitute a clinical criterion by itself. For this reason, facial analysis is integrated as a complementary dimension to textual analysis, and the results must be interpreted holistically and under the supervision of mental health professionals.

The integration of textual and facial analysis helps mitigate the lack of information when users publish little text or few images containing faces. Nevertheless, there are technical limitations related to the restrictive policies of social media APIs, which are intended to protect user privacy. This suggests that these tools should be used with academic consent and as a complement to professional assessment. The developed system represents progress toward low-cost, easily deployable preventive mental health tools.

Throughout the research process related to symptoms associated with depression, several specific symptoms were identified that may be useful for formulating a diagnosis. However, some of these symptoms may also be part of other conditions, as in the case of references to pain. Conversely, some indicators may appear to be important symptoms but may not actually be so, such as the excessive use of first-person singular pronouns. Although these pronouns may frequently be used to focus topics, situations, or feelings on oneself, this way of speaking is also quite common on social media. These types of ambiguities make it difficult to select keywords associated with depression or related symptoms.

During the scraping process, it was found that many users tend to publish posts with very little text, only emoticons, or no text at all. This makes textual analysis of profiles difficult. Similarly, some users do not publish many photographs, and when they do, the images may not contain any faces, which prevents facial features from being extracted. In both cases, there is a lack of information; however, by having both textual analysis and facial analysis tools, these shortcomings can be partially addressed.

Tools such as psychological tests or interviews have been developed to generate a diagnosis or obtain key traits from individuals, helping both the patient and the psychologist or researcher identify these symptoms. Developing a tool of similar quality is particularly difficult, not because an equivalent cannot be automated based on these tools and tests, but because of the object of study used in this project. Although people's profiles and posts can provide a considerable amount of information, they are not always reliable. This may be due to several factors, such as the fact that these posts do not necessarily reflect the person's actual emotional state and the control users have over their public image, which may lead to incomplete results.

Based on this project, the following proposals can be considered:

- Regarding the mode of use, some APIs obtain biometric data from individuals or their photographs in order to identify them across images. These functions make it possible to track a person's emotional state over time or isolate their data within a group of people. This may be useful for improving the accuracy and automation of data analysis applications, enabling the implementation of a similar system based on biometric datasets generated through additional services provided by facial analysis APIs.
- For future research, it would be highly useful to work with a dataset that includes facial features, especially emotions, in order to train an algorithm whose parameters include both facial features and text, thereby generating more accurate results.

Ethics Statement

This study uses biometric data obtained through the analysis of images published on public Instagram profiles, processed using the Face++ API. Since these data include facial features associated with emotions, their handling was carried out in accordance with the following ethical principles:

- The analyzed profiles were public, and no private or restricted information was accessed.
- Users' images were not stored; only the numerical emotion values returned by the API were recorded.
- The system does not allow the analyzed individuals to be identified or tracked, since the data are processed in an aggregated and anonymized manner.
- The Face++ API was used within its terms of service for exclusively academic and research purposes.
- Textual posts were processed for linguistic analysis purposes, without disclosing or permanently storing the users' original content.
- The developed system is intended as a support tool for mental health professionals and not as a definitive clinical diagnostic instrument.

This study did not require individual informed consent, since the data were obtained from publicly accessible sources. Nevertheless, the need for more precise regulatory frameworks for the use of biometric data in digital mental health research is acknowledged.

AUTHOR CONTRIBUTION

The author's contributions to this article are as follows:

Javier Antonio Ballesteros Ricaurte: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization, Project administration, Writing–original draft, Writing–review and editing.

The author reviewed the results and approved the final version of the article.

CONFLICT OF INTERESTS

The authors declare that they have no interests or financial relationships that could have influenced this work.

REFERENCES

[1] J. R. Ruiz Avendaño, «EVALUACIÓN DE ALGORITMOS DE APRENDIZAJE DE MÁQUINA PARA LA DETECCIÓN DE SIGNOS DE DEPRESIÓN EN JÓVENES ESTUDIANTES A PARTIR DE INTERACCIONES TEXTUALES EN REDES SOCIALES», Tecnológico de Antioquia, 2021.

[2] A. Abal Aldao, «Uso de algoritmos de apredizaje máquina para la detección de ciberacoso en redes sociales», Universidade da Coruña. Facultade de Informática. Accedido: 8 de junio de 2024. [En línea]. Disponible en: http://hdl.handle.net/2183/33992

[3] J. C. Figuereo-Benítez, F. González-Quiñones, y J. D. Machin-Mastromatteo, «Instagram como objeto de estudio en investigaciones recientes. Una revisión de literatura con enfoque en revistas científicas.», Ámbitos. Revista Internacional de Comunicación, n.o 53, pp. 9-23, 2021, doi: 10.12795/ambitos.2021.i53.01.

[4] M. A. DÜNDAR y F. TUFAN, «Social Comparison Behaviour on Social Network Sites: The Comparison Between Twitter and Instagram», Yeni Medya Dergisi, vol. 2022, n.o 12, pp. 1-16, jun. 2022, doi: 10.55609/yenimedya.1051044.

[5] Asociación Americana de Psiquiatría, «Psychiatry.org - ¿Qué es la depresión?» Accedido: 17 de octubre de 2024. [En línea]. Disponible en: https://www.psychiatry.org/patients-families/la-salud-mental/depresion/que-es-la-depresion

[6] L. Subirats y M. Calvo, «Web scraping», 2019. Accedido: 15 de septiembre de 2024. [En línea]. Disponible en: https://openaccess.uoc.edu/bitstream/10609/147437/1/WebScraping_Modulo1_WebScraping.pdf

[7] World Health Organization, «Depressive disorder (depression)», World Health Organization. Accedido: 16 de septiembre de 2024. [En línea]. Disponible en: https://www.who.int/news-room/fact-sheets/detail/depression

[8] H. Velásquez Cabrejos y J. M. Antón Pérez, «"Validez y confiabilidad del Test de Zung que categoriza la depresión enadultos mayores. – zonas de la Región Cajamarca. 2019"», UNIVERSIDAD NACIONAL PEDRO RUIZ GALLO, 2019. [En línea]. Disponible en: https://hdl.handle.net/20.500.12893/8249

[9] M. José Garciarena Ucelay, U. Nacional de San Luis, L. Cecilia Cagnina, y M. Luis Errecalde, «Análisis de rasgos lingüísticos con técnicas de procesamiento del lenguaje natural en la detección temprana de depresión», Anales de Lingüística, vol. 3597, n.o 7, pp. 89-116, 2021, [En línea]. Disponible en: https://revistas.uncu.edu.ar/ojs3/index.php/analeslinguistica/article/view/5522

[10] M. S. Batista, J. Arana, y J. García, «CARACTERIZACIÓN DEL PERFIL DE LENGUAJE EN EL TRASTORNO DEPRESIVO Y EN EL DOLOR CRÓNICO», Salamanca, jun. 2019. doi: 10.14201/gredos.140363.

[11] M. Al-Mosaiwi y T. Johnstone, «Corrigendum: In an Absolute State: Elevated Use of Absolutist Words Is a Marker Specific to Anxiety, Depression, and Suicidal Ideation (Clinical Psychological Science, (2018), 6, 4, (529-542), 10.1177/2167702617747074)», 1 de mayo de 2019, SAGE Publications Inc. doi: 10.1177/2167702619843297.

[12] J. SUAREZ, «Comparación entre algoritmos de Machine Learning en la implementación del análisis de sentimientos para textos en español», BOGOTÁ D.C., 2021.

[13] J. R. Ruiz Avendaño, «EVALUACIÓN DE ALGORITMOS DE APRENDIZAJE DE MÁQUINA PARA LA DETECCIÓN DE SIGNOS DE DEPRESIÓN EN JÓVENES ESTUDIANTES A PARTIR DE INTERACCIONES TEXTUALES EN REDES SOCIALES», Tecnológico de Antioquia, 2021.

Javier Antonio Ballesteros Ricaurte. Universidad Pedagógica y Tecnológica de Colombia