Fine grain Emotion Analysis in Spanish using linguistic features and transformers

Alejandro Salmerón-Ríos[1], **José Antonio García-Díaz**[1], **Ronghao Pan**[1], and **Rafael Valencia-García**[1]

[1]**Informatics and Systems, Universidad de Murcia, Murcia, Spain**

Corresponding author:
José Antonio García-Díaz[3]

Email address: joseantonio.garcia8@um.es

ABSTRACT

Mental health issues are a global concern, with a particular focus on the rise of depression, which affects millions of people worldwide and is a leading cause of suicide, especially among young people. Recent surveys indicate an increase in depression cases during the COVID-19 pandemic, which will affect approximately 5.4% of the population in Spain in 2020. Social media platforms such as Twitter have become important hubs for health information as more people turn to these platforms to share their struggles and seek emotional support. Researchers have discovered a correlation between emotions and mental illnesses such as depression. This correlation provides a valuable opportunity for automated analysis of social media data to detect changes in mental health status that might otherwise go unnoticed, potentially preventing the development of more serious health consequences. This research explores the field of emotion analysis in Spanish towards mental illness. There are two contributions in this area. On the one hand, the compilation, translation, evaluation and correction of a novel dataset composed of a mixture of other existing datasets in the bibliography. This dataset compares a total of 16 emotions, with an emphasis on negative emotions. On the other hand, the in-depth evaluation of this novel dataset with several encoder-only and encoder-decoder architectures based on state-of-the-art transformers, including monolingual, multilingual and distilled models as well as feature integration techniques. The best results are obtained with the encoder-only MarIA model, with a macro-average F1 score of 60.4771% over the test set.

1 INTRODUCTION

Mental health is a significant global public health problem. Specifically, depression affects approximately 300 million people worldwide and is a major contributor to suicide; depression is the third leading cause of death among people aged 10 to 24. According to the National Statistics Institute[1](INE) and the latest European Mental Health Survey conducted in Spain, an increase in cases of depression in the population due to the pandemic was observed between July 2019 and July 2020 (World Health Organization and others, 2022). In 2020, a total of 5.4% of the population (about 2.1 million people) will experience some form of depression. Early diagnosis of mental health problems is critical to effective treatment, as individuals are typically reluctant to seek help from specialized clinicians to treat their conditions. However, social media platforms are often used by these individuals to discuss their difficulties and find emotional support. This presents a significant opportunity for automated analysis of social media data to detect potential changes in their mental state that would otherwise go unnoticed.

Emotion Analysis (EA) is an aspect of Automatic Document Classification (ADC) that attempts to identify emotions conveyed in text documents. Ongoing research and development in this area is motivated by the growing awareness of mental health, the increased use of social networks, and the need to identify users' states of mind. Advances in Natural Language Processing (NLP) research have led to the development of innovative techniques that show remarkable performance in tasks such as EA, as demonstrated by Transformer-based models (Acheampong et al., 2021), (García-Díaz et al., 2021). This interest in NLP has led to exciting advances in the field.

---

[1]https://www.ine.es/

One way to identify mental disorders is through everyday communication. Research conducted in (De Choudhury et al., 2014) and (Guntuku et al., 2017) has shown that individuals with mental disorders exhibit variations in language and behavior, including increased expression of negative emotions and self-absorption. As a result, clues to an individual's altered mental state can be found in their online posts, which often convey negativity. Thus, textual emotion detection is an important aspect of natural language, where emotions in written material are identified using existing emotion-tagged datasets and algorithms. Unlike Sentiment Analysis (SA), which generally categorizes text into broad classes(Medhat et al., 2014), EA requires fine-grained analysis using emotional scales. The evaluation of emotions is a well-researched topic with numerous papers, and the scales typically include Paul Ekmann's six basic emotions (Ekman and Davidson, 1993). However, there is limited attention to the Spanish language in the literature, and the majority of studies only address the basic emotions outlined by Ekmann. Nevertheless, the importance of mental health during this period, along with related conditions such as depression, loneliness, and hopelessness, have been key drivers of this shift and the emotions associated with it.

For this paper, we have defined the following research questions related to EA for the detection of mental disorders:

- **RQ1.** What is the reliability of identifying negative fine-grained emotions?

- **RQ2.** What is the best approach to face the emotion analysis using the text as input?

- **RQ3.** Are generative models effective in identifying different emotions?

The paper makes significant contributions to the field of EA for the detection of mental disorders: (1) The dataset we compile and evaluate includes 16 different emotions using a multi-classification scheme. This method provides a unique approach by including emotions and states beyond those defined by Ekman's basic emotions. This dataset includes emotions such as loneliness, depression, suicide, and hopelessness. (2) We evaluate this aforementioned dataset with several encoder-only, encoder-decoder, and feature integration models for text generation in EA due to the promising results reported in several studies (Plaza-del Arco et al., 2023), (Brown et al., 2020a).

The paper is organized as follows: Section 2 presents a summary of the current literature on NLP and EA, as well as the state of the art in emotion detection in mental disorders. Section 4 describes the materials and methods used in this study, including a detailed description of our proposed pipeline. Section 3 describes the compiled dataset and the set of experiments performed is presented. Next, Section 5 illustrates the experiments conducted to evaluate the different strategies analyzed and discusses the results. Finally, Section 6, presents the implications of the results and possible future work.

2 STATE OF THE ART

This section contains a review of the recent research and literature investigating the use of NLP techniques for analyzing emotions (see Section 2.1) and their application in the medical field (see Section 2.2).

2.1 Natural Language Processing for Emotion Analysis

EA serves as a tool for identifying specific human emotions, including anger, sadness, and fear, among others. With the advancement of Internet services, people are increasingly using social media platforms to express their emotions, participate in discussions, and exchange views on a variety of topics. In addition, some users provide feedback and review various products and services on e-commerce websites. In today's digital age, organizations across all industries are experiencing a digital revolution, resulting in a significant increase in both structured and unstructured data. A critical responsibility for these organizations is to transform unstructured data into valuable insights that can inform the decision-making process (Munezero et al., 2014). Identifying emotions from user-generated text enables the recognition of their emotional state and perspective on products or services. As a result, vendors and service providers are inspired to improve their current systems, products or services.

Sentiment and Emotion Analysis are vital in the education sector for both students and teachers. According to (Sangeetha and Dhandayudam, 2021), a teacher's effectiveness depends not only on their academic qualifications but also on their enthusiasm, talent, and dedication. Therefore, getting regular feedback from students is an efficient way for teachers to improve their teaching methods. Automated processing methods such as sentiment and emotion analysis can help interpret a student's textual feedback. This can provide valuable insights to help teachers and institutions make improvements.

In recent years, researchers have made efforts to automate emotion analysis, but emotion detection from text remains a challenging task due to ambiguities and the introduction of new slang or terminologies. Broadly speaking, emotion models can be divided into two categories:

- **Dimensional model of emotions**: This model represents emotions using three parameters: valence, arousal, and power. Valence refers to the polarity of an emotion, while arousal measures the level of intensity associated with a feeling. Power or dominance refers to the degree of control over an emotion.

- **Categorical model of emotions**: This model represents emotions as discrete parameters, such as sadness, happiness, anger, and others. Emotions are categorized into varying number of groups, ranging from four to eight depending on the specific model.

In the field of EA using categorical models of emotion, most researchers prefer Ekman's or Plutchik's (Plutchik, 1980) emotion models as a fundamental basis. These models define emotional states to be used when labeling sentences or documents. For example, Ekman's six basic emotions have been used by (Batbaatar et al., 2019) and (Becker et al., 2017), while some researchers have introduced one or two additional emotional states, such as Plutchik's (Mohsin and Beltiukov, 2019), (Park et al., 2020) to develop customized emotion models. In addition, in existing studies or cases, researchers have used a classification system that includes more than 10 emotions. For example, (Cowen and Keltner, 2017), employed a self-report measure that captured 27 different categories of emotion, bridged by a continuous gradient. Similarly, (Lazarus and Lazarus, 1994), considered emotional analysis as a set of 15 emotions.

Text-based EA is usually considered as a supervised machine learning classification problem, which involves assigning one or more emotion categories to sentences or text selections. According to Saffar et al. (2023), the common process that takes place in text-based EA in most supervised machine learning classification systems is: data preprocessing, tokenization and lemmatization, feature extraction, and machine learning algorithms. The first step, common to most NLP tasks, is the preprocessing of textual data. On social media platforms, individuals often express their emotions informally, resulting in highly unstructured data that makes sentiment and emotion analysis difficult for machines. Data preprocessing therefore plays a key role in ensuring data quality, as it has profound impact on subsequent analysis. Tokenization, lemmatization, stop word removal, and spell checking are the most common operations at this stage. Once the text is cleaned and normalized, it moves on to feature extraction, model training, development, and system evaluation. Next, in the feature extraction stage, the text is segmented into sentences and words through tokenization. These tokens are then transformed into understandable numerical vectors suitable for machine learning algorithms. Prominent methods for feature extraction include Bag of Words (BOW), Term Frequency–Inverse Document Frequency (TF–IDF), and word embedding.

Text-based EA models use these input vectors to establish mapping relationships between input text and output emotion labels or scores. Various techniques for EA fall into five broad categories: (1) lexicon-based approaches, (2) machine learning-based approaches, (3) deep learning-based approaches, (4) hybrid approaches, and (5) transfer learning methods (Nandwani and Verma, 2021a).

Lexicon-based methods rely on affective keywords associated with psychological states (Murthy and Kumar, 2021), often using lexicons such as WordNet-Affect. WordNet-Affect (Strapparava and Valitutti, 2004), an extended form of WordNet (Miller, 1995), contains affective words annotated with emotion labels. However, these methods face challenges such as keyword ambiguity and limited linguistic information.

To address domain-specific problems, corpus-based methods have emerged, albeit with limitations in generalizability. Naïve Bayes and Support Vector Machines (SVMs) are machine learning models that can detect emotions and can be adapted to new tasks by using training and test datasets (Nabeel et al., 2021). Deep learning networks, especially those using word or sentence embeddings, have shown excellent performance in identifying emotions in text. Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), FastText (Joulin et al., 2017), and ELMO (Peters et al., 2018) are commonly used embedding techniques.

Hybrid models combine different methods to overcome limitations, which may include machine learning, lexicon-based approaches, or deep learning networks. The need for hybrid models arises from the complexity and ambiguity involved in expressing emotions in natural language, with some labels

having insufficient training instances (Nandwani and Verma, 2021b). Traditionally, datasets were assumed to belong to the same domain. However, when the domain changed, a new model was required. Transfer learning provides a solution by allowing the reuse of pre-trained models in different domains, thereby eliminating the need to develop new models for each domain change.

Although EA is a frequently studied topic with a considerable amount of literature available, there has been a lack of attention given to the Spanish language. One study that has addressed this issue is the EmoEvalEs shared task (Plaza-del Arco et al., 2021) organized by IberLEF 2021, which aims to promote the recognition and evaluation of emotions in the Spanish language. Our research group participated in this task (García-Díaz et al., 2021), where we presented a method that merges explainable linguistic features with BETO (Cañete et al., 2020), resulting in an accuracy rate of 68.5990%.

2.2 Emotion Analysis in the medical domain

Mental illness alters the way a person feels, thinks, or acts and is a major public health problem, causing disability and poor well-being worldwide Rehm and Shield (2019). The World Health Organization (WHO) highlights that one in eight people struggle with mental illness, posing a significant economic challenge to governments[2]. The Covid-19 pandemic has exacerbated this problem and increased its impact (Skaik and Inkpen, 2020). Early detection of mental illness can prevent its progression to a severe state and allow for intervention (Leiva and Freire, 2017). However, most patients with mental illness do not receive effective diagnosis and treatment due to ignorance about mental health assessment and the stigma associated with these illnesses.

Social media platforms such as X (formerly known as Twitter) have become indispensable sources of information for the healthcare industry, as more and more people turn to these platforms to share their ailments and seek emotional support. This presents an important opportunity for automated processing of social media data to identify changes in mental health status that might otherwise go unnoticed, before they develop into more serious health consequences (Zhang et al., 2023a). It is therefore imperative to detect mental illness in individuals at an early stage, as this could have life-saving implications. NLP is playing an increasingly important role in the processing of social media data and has been used to facilitate tasks such as sentiment and emotion analysis and mental health assessment.

The CLEF eRisk Lab[3] has been held annually since 2017 as a shared task aimed at identifying early signs of mental disorders from social media posts. The 2023 edition of eRisk (Parapar et al., 2023) focused on detecting early signs of depression, pathological gambling, and measuring the severity of eating disorder symptoms. In the working note published by the UMUTeam on MentalRisk (Pan et al., 2023b), they have demonstrated good performance of pre-trained models based on Transformers for the detection of mental disorders. Furthermore, in another study by the same team (Pan et al., 2023a), it has been shown that emotions are a feature that complements depression detection models and could help improve their performance.

In addition, previous computational studies have shown that individuals with mental disorders consistently exhibit changes in their speech and behavior, including an increased prevalence of negative emotions and a heightened focus on self-attention (Guntuku et al., 2017). As a result, daily communication is essential in detecting mental disorders. To protect patients from mental health problems, such as depression, physicians should use automated sentiment and emotion analysis, as recommended in Singh et al. (2021).

Given that emotions are an important part of human nature and can affect people's behavior and mental states (Canales and Martínez-Barco, 2014), a correlation has been found between emotions and mental illnesses such as depression (Compare et al., 2014). For example, in Joormann and Gotlib (2010), it was shown that the severity of depressive symptoms is associated with an increasingly inverse relationship between positive and negative emotions. Furthermore, in Dejonckheere et al. (2019) and Dejonckheere et al. (2018), it was also shown that individuals with depressive symptoms have difficulty regulating emotions, resulting in lower emotion complexity. Therefore, from a psychological perspective, information about emotions is useful in diagnosing mental illness.

Currently, EA is a frequently addressed and studied topic, and there are numerous works and shared tasks dedicated to it. However, few of them focus on the Spanish language, and most of them concentrate exclusively on the use of Paul Ekman's basic emotions. Nevertheless, the importance of mental health at

---

[2]https://www.who.int/news-room/fact-sheets/detail/mental-disorders
[3]https://erisk.irlab.org/

this time and the states related to it, such as depression, loneliness, or hopelessness, are indicators that would help detect mental health problems. For this reason, in this paper, we compiled and evaluated a dataset labeled with 14 different emotions, offering a unique approach by including emotions and states that go beyond the basic emotions defined by Ekman (anger, disgust, fear, joy, sadness, and surprise). This dataset includes emotions such as loneliness, depression, and hopelessness.

Unlike other approaches to mental health detection, such as DAS (Depression, Anxiety, Stress) detection models, EA models can provide greater consistency and enable the visualization of mood changes through published text, thereby avoiding false positives. For example, a semi-supervised machine learning model, DASentimental, has been proposed in Fatima et al. (2021) to extract depression, anxiety, and stress from written text. However, it is only capable of identifying negative emotions.

Our approach to EA is based on creating a multi-classification dataset of 16 different emotions (covering emotions such as loneliness, depression, and hopelessness) and using transfer learning, which involves fine-tuning and evaluating different encoder-only LLMs (including a combination of Spanish and multilingual models). These models include (1) BETO (Cañete et al., 2020), (2) ALBETO (Cañete et al., 2022), (3) BERTIN (de la Rosa et al., 2022), (4) XLM (Conneau et al., 2020), (5) DistilBETO (Cañete et al., 2022), (6) MarIA (Gutiérrez-Fandiño et al., 2022), (7) multilingual BERT (Devlin et al., 2019), (8) multilingual DeBERTA (He et al., 2021), and (9) TwHIN (Zhang et al., 2023b) as well as encoder-decoder text generation models for text classification tasks, such as BLOOM (Scao et al., 2022), BART (Lewis et al., 2020), GPT-2 (Radford et al., 2019), and Llama-2 (Touvron et al., 2023).

3 DATASET

In this research, a novel dataset is formulated by combining, translating, and relabeling pre-existing datasets using a detailed emotional classification system.

The first step in developing this corpus is to establish a taxonomy of emotions. We start with Ekman's taxonomy, which includes negative emotions such as sadness and anger. Next, we include Plutchik's Wheel of Emotions, which is a broader representation than Ekman's and considers fine-grained emotions such as grief, disgust, or remorse. We complement this taxonomy with research by Leis et al. (2019), which identified a list of words that might indicate signs of depression. The choice of words was made by psychiatrists, members of the Institute of Neuropsychiatry and Addictions (INAD), Parc Salut Mar, Barcelona, Spain. It includes both the words and and a score assigned to them, which is the sum of the scores given by each given by each evaluator on a Likert scale (from 1 to 5) according to the relevance of that word to the of that word to a patient with depression. The alternative state or emotion that some of them could reflect was extracted. Some of the emotions are depressed, disappointed, ashamed, hopeless, lonely, regretful, nervous, and suicidal. Keeping in mind that the terms provided could represent other conditions in addition to identifying signs of depression, the terms with the highest scores were extracted and included in the taxonomy. The final list of emotions are: (1) anger, (2) depressed, (3) disappointment, (4) disgust, (5) embarrassment, (6) fear, (7) grief, (8) hopeless, (9) joy, (10) lonely, (11) nervousness, (12) neutral, (13) remorse, (14) sadness, (15) suicidal, and (16) surprise.

Once the emotions were identified, the next step was to find corpora that contained at least one or more of these emotions.

- Merging Datasets for EA (De Arriba et al., 2021). The International Workshop on Software Engineering Automation: A Natural Language Perspective (NLP-SEA) presented this dataset in 2021. It contains 5260 Spanish documents classified as: (1) sadness, (2) not relevant, (3) fear, (4) happiness, (5) anger or (6) surprise. The dataset was collected from Twitter and it is related to the COVID-19 outbreak.

- Detecting signs of depression in tweets in Spanish: behavioral and linguistic analysis (Leis et al., 2019). It contains documents about detecting signs of depression in Spanish.

- Detecting Depression in Social Media Via Twitter Usage[4]. This dataset is compiled from Twitter and contains texts related to the labels depressed, hopeless, lonely and suicidal. The keywords used to retrieve the dataset are depressed, depression, hopeless, lonely, suicide, and antidepressant.

---

[4]https://github.com/ram574/Detecting-Depression-using-Tweets

- GoEmotions (Demszky et al., 2020). This dataset was collected by Google from popular English subreddits and manually tagged with 27 emotion categories, including 12 positive emotion categories, 11 negative, 4 ambiguous, and 1 neutral. To validate that the taxonomic choices were consistent with the given emotions, a principal component analysis was performed, which allowed two datasets to be compared by extracting linear combinations with greater variability.

Next, we show some translated examples of the dataset. There are several messages concerning coronavirus: *I think no one is going to forget the "COVID-19" pandemic... not even the children.* (sadness), *I'm so scared to go to my doctor's appointment tomorrow because I don't want to catch the corona virus, I hope no one there has it when I go tomorrow.* (fear), or *The European Medicines Agency will make a statement shortly. Yes, there is a link between the ASTRA-ZENECA vaccine and blood clots. Better late than never.* (anger). Another examples are *I have been going to a Mexican restaurant in my neighborhood every month for the past 2.5 years, but no one knows my name.* (disappointment) and *Every night I cry in my bed and think of ways to kill myself. I have wanted to end my life for years, but I couldn't. If I didn't know that suicide is a sin, I would have done it long ago.* (suicidal).

Once the sentences were compiled, we translated the English sentences into Spanish using SYSTRAN[5] and we checked them manually.

Another problem we faced was that the sentences from Reddit were too long for some models, such as Transformers. Since these documents were too complex, we decided to split them into sentences and annotate each sentence individually. The annotation phase was carried out by members of our research group, where each document was revised three times and the emotion of the sentence was decided in a final meeting.

A cleaning process is performed as a final step, removing retweets and stripping the documents of line breaks, mentions, hyperlinks, images, and ensuring that there are no duplicate documents.

Next, we split the dataset into training, validation, and test in a ratio of 60-20-20. Table 1 shows the statistics of the final dataset and the number of instances in each split. This dataset contains 38,559 sentences annotated with 16 emotions. It can be observed that the emotions are not balanced, and there is a significant lack of emotions labeled as surprise. Joy is another underrepresented emotion, with only 186 instances. However, sentences related to negative emotions, such as depression, disappointment, lonely, or sadness contain several examples.

Table 1. Dataset statistics, including the number of emotions per label and split

Emotion	Train	Val	Test	Total
anger	775	259	259	1,293
depressed	3,520	1,174	1,174	5,868
disappointment	2,823	941	941	4,705
disgust	1,295	432	432	2,159
embarrassment	730	243	244	1,217
fear	1,280	427	427	2,134
grief	184	61	62	307
hopeless	739	247	247	1,233
joy	111	37	38	186
lonely	2,931	977	977	4,885
nervousness	358	120	120	598
neutral	3,358	1120	1,120	5,598
remorse	596	199	199	994
sadness	639	213	214	1,066
suicidal	3,786	1262	1,263	6,311
surprise	3	1	1	5
Total	23,128	7,713	7,718	38,559

Figure 1 illustrates the information gain of the linguistic features extracted with UMUTextStats

[5]https://www.systran.net/en/translate/

281 (García-Díaz et al., 2022) for each emotion. [6] It can be observed that the most informative features
282 are related to stylometry, specifically the length of the corpus and the number of syllables and words.
283 Other significant features are health-related lexical items, which are correlated with sadness and suicidal
284 feelings. Other relevant features include qualifying adjectives and negative processes.
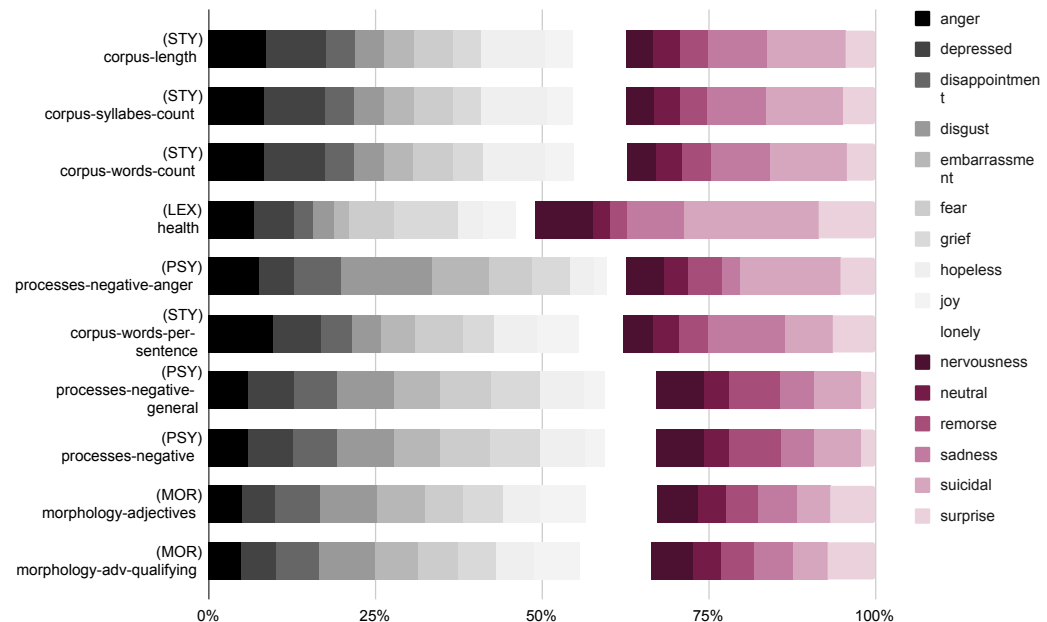
Figure 1. Information gain of the top 10 linguistic features of the dataset. The values are normalized in order to fit the 100%.

285   The compiled dataset is available to the scientific community[7]

4 MATERIALS AND METHODS

287 To evaluate the dataset, our evaluation primarily used two techniques that involve fine-tuning different
288 LLM models to determine the emotion. Specifically, we analyzed two different language models for the
289 classification task: encoder-only models based on MLM (Masked Language Models) and encoder-decoder
290 autoregressive transformers for text generation, including BART, T5, GPT-2, BLOOM, and Llama-2.
291 We also tested several model combination techniques, including knowledge integration and ensemble
292 learning.
293   Our pipeline is shown in Figure 2. First, the data preprocessing module is used to cleanse the texts.
294 Second, the splitter module separates the corpus into training, validation, and test sets. Third, feature
295 sets are extracted to train the classification model. Fourth, we fine-tune different LLMs for emotion
296 classification. Finally, we evaluate different ensemble learning methods for text classification.

4.1 Data pre-processing and splitter

298 We used a variety of feature sets, including linguistic features for the baseline and different types of
299 embedding. A standard preprocessing procedure was performed, which consisted of removing all social
300 media jargon, such as hyperlinks, hashtags, or mentions. In addition, all percentages and numbers
301 were replaced with fixed tokens to prevent the classifiers from learning specific quantities. Finally, the
302 normalized version is used to extract tokens of the embedding-based features and then to fine-tune various
303 LLMs.
304   The splitter module plays a crucial role in extracting the training, development, and test datasets, and
305 these splits vary depending on the task at hand. In this particular scenario, we constructed a dataset with a

---

[6]Please, refer to Section 4.2 to a description of the linguistic categories of this tool.

[7]The link will be provided upon acceptance.
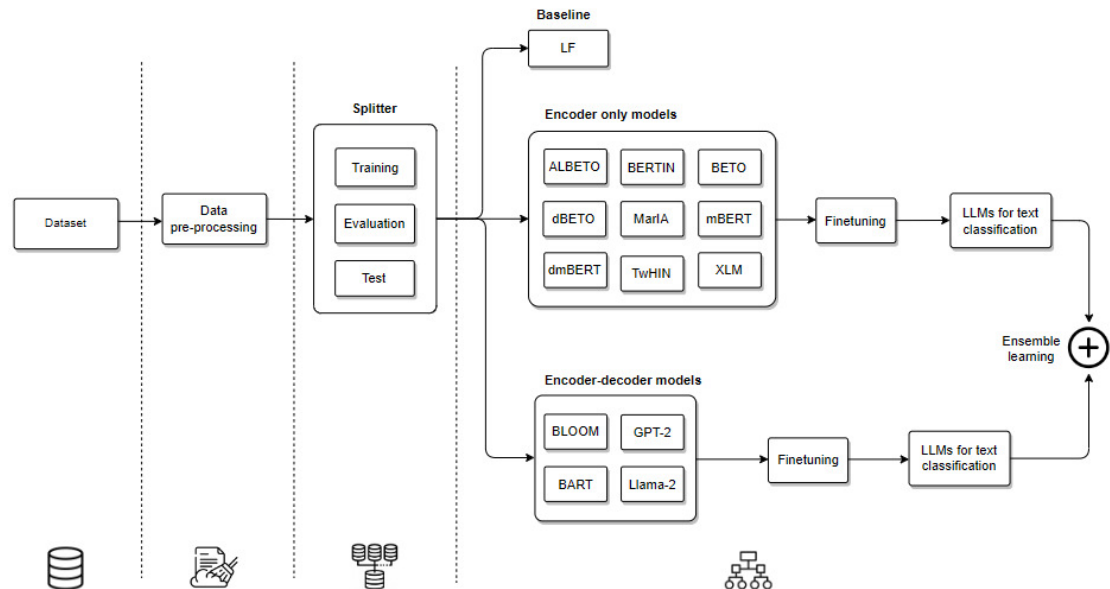
Figure 2. General architecture of our pipeline for emotion classification.

total of 38.559 documents with different emotions and some of them are unbalanced, as shown in Table 1.
Therefore, we have divided the dataset in a 60-20-20 ratio with stratified sampling to ensure that each
emotion is correctly represented in each sample.

4.2 Baseline

To compare the results of the encoder-only and encoder-decoder strategies, we build a baseline based
on linguistic features (LFs). LFs are a means of representing documents as a vector formed by the
percentages and raw counts of linguistically relevant features that indicate what a text says and how it
says it (Tausczik and Pennebaker, 2010). To extract the LFs, we rely on UMUTextStats (García-Díaz
et al., 2022). This tool captures 365 linguistic features, organized as follows:

- **Correction and style of written communication (COR).** The analysis detects a variety of errors,
  including orthographic errors such as misspelled words, stylistic issues, and performance errors.
  These performance errors can consist of sentences beginning with numbers or the same word, as
  well as identifying common errors and unnecessary phrases.

- **Phonetics (PHO).** It documents expressive lengthening, in which certain letters are deliberately
  lengthened to emphasize their meaning.

- **Morphosyntax (MOR).** The structure of words is recorded along with grammatical attributes,
  including gender and number, and various affixes, such as nominal, adjectival, verbalizing, adverbial,
  augmentative, diminutive, and derogatory suffixes. In addition, these features are organized
  according to their respective part-of-speech categories, such as verbs, nouns, and adjectives.

- **Semantics (SEM).** Includes sound words, polite expressions, derogatory terms, and figures of
  speech in which the part represents the whole.

- **Pragmatics (PRA).** The use of figurative language devices, including understatement, rhetorical
  questions, hyperbole, idiomatic expressions, verbal irony, metaphors, and similes.

- **Stylometry (STY).** It records punctuation symbols, corpus statistics, and metrics related to the
  number of words, syllables, or sentences.

- **Lexical (LEC).** The text provides a thorough identification and analysis of the topics, covering
  both abstract and general topics.

- **Psycho-linguistic processes (PLI)**. This category contains emoticons and lexicons related to emotions and feelings.

- **Register (REG)**. It emphasizes the distinction between informal and formal language, and addresses topics that may be considered offensive.

- **Social media jargon (SOC)**. This category captures features related to the speaker's mastery of social media jargon.

For the training the baseline based on the linguistic features, we relied on a Multilayer Perceptron (MLP) because these features do not contain sequential information such as text.

4.3 Encoder-only classification model

For the encoder-only classification model, we evaluated several encoder-only language models. Figure 3 shows the pipeline of our encoder-only approach and different feature-integration strategies, such as knowledge integration.

Regarding sentence embeddings, several pre-trained transformer-based models are evaluated, which can be classified into BERT-based and RoBERTa-based models. The main difference between these two architectures is that RoBERTa performs masking during training, whereas BERT performs masking at the beginning of training. Transformer-based models can be pre-trained with a multilingual corpus or with a monolingual corpus of a specific language. Thus, we have performed an evaluation of several multilingual and monolingual Spanish models for this task. It is possible to categorize models using BERT and RoBERTa-based architectures as follows:

- **BERT-based monolingual model**. The study used the Spanish variant of BERT called BETO (Cañete et al., 2020). The evaluation of BETO model also included two lighter versions derived from it: ALBETO (Cañete et al., 2022) and DistilBETO (Cañete et al., 2022).

- **BERT-based multilingual model**. In terms of BERT-based models pre-trained with a multilingual corpus, the models used are: (1) TwHIN-BERT (Zhang et al., 2023b), a multilingual tweet language model trained with 7 billion tweets from more than 100 different languages; (2) M-BERT (Devlin et al., 2019) is a transformer model pre-trained with a large corpus of multilingual data in a self-supervised manner; (3) M-DistilBERT (Sanh et al., 2019), a distilled version of the multilingual BERT base model.

- **RoBERTa-based monolingual model**. In this paper, we have evaluated different models pre-trained with the Spanish corpora, such as MarIA (Gutiérrez-Fandiño et al., 2022) and BERTIN (de la Rosa et al., 2022).

- **RoBERTa-based multilingual model**. Regarding RoBERTa-based models pre-trained with a multilingual corpus, we evaluated XLM-RoBERTa (100-1280) (Lample and Conneau, 2019), which is a multilingual version of RoBERTa trained with data filtered from CommonCrawl from 100 different languages.

In order to produce more robust solutions, different strategies based on the integration of feature sets have been evaluated. In particular, two strategies are evaluated: single feature evaluation and knowledge integration. The first strategy, known as single feature evaluation, does not combine feature sets. Multiple models are trained individually for each feature set, including hyperparameter tuning. Then, the best model is selected based on the validation split. This process is also called fine-tuning, which consists of taking pre-trained models and adapting them to a specific task or domain. The second approach is called knowledge integration. This method involves training a multi-input neural network from scratch, incorporating all the sentence embeddings for each encoder-only mode. The idea behind this approach is that the network learns during training how to exploit the strengths of each feature set. In this work, we adopt a network architecture design where each feature set is connected to a separate stack of hidden layers. The output of each layer is then concatenated and connected to a new set of hidden layers, which in turn are connected to the final output layer.
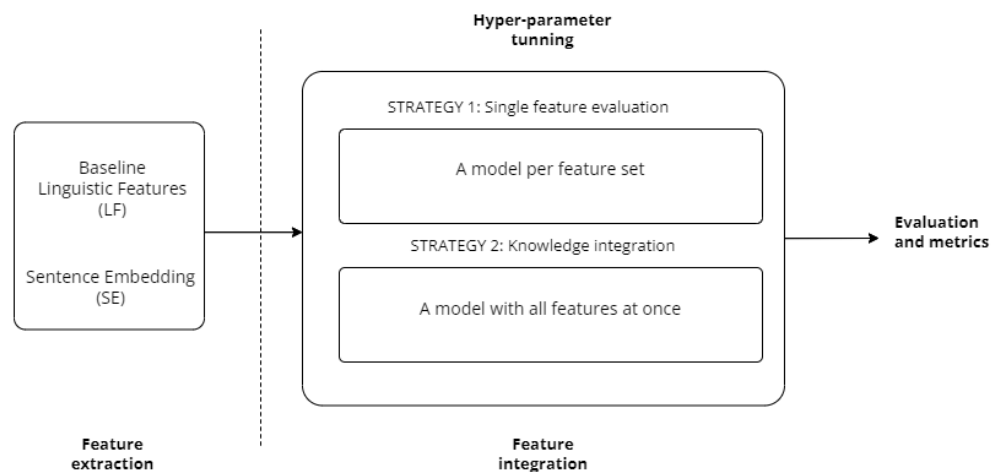
Figure 3. General architecture of the feature-based classification model for EA.

4.4 Encoder-decoder models

LLMs have revolutionized many aspects of NLP. According to Brown et al. (2020a), LLMs have the ability to learn from few or even no examples due to their inherent ability to "understand" language. New models such as Llama-2 (Touvron et al., 2023) and GPT-3 (Brown et al., 2020b), which have been trained on large and diverse text corpora, have further enhanced these capabilities. This has opened up a wide range of possibilities in the field of NLP, such as direct label prediction through prompting. In fact, several articles have tested this hypothesis and achieved good performance across on a wide variety of NLP tasks (Plaza-del Arco et al., 2023; Wang et al., 2023).

In this study, we evaluated several popular generative LLMs such as GPT-2 (Radford et al., 2019), BLOOM (Scao et al., 2022), BART (Lewis et al., 2020), and Llama-2 (Touvron et al., 2023), which are models pre-trained on massive amounts of text data, enabling them to generate coherent and contextually relevant text in emotion classification tasks. However, since we have a total of 16 possible emotions and some of them are similar, it poses a challenge for zero-shot and few-shot learning for these models. Therefore, we performed a fine-tuning process on the LLMs. Fine-tuning involves taking a pre-trained LLM and adapting it to a specific task or domain. This is done by training the LLM on a smaller dataset that is specific to the task or domain and adjusting the model weights and parameters to better fit the new data. In this way, the model can achieve good performance without the need for extensive training data or abundant computational resources. The models tested in this study are the following:

- **BLOOM**. It is an autoregressive LLM that has been trained on extensive text data using industrial-scale computing resources. It excels at generating coherent text in 46 languages and 13 programming languages, making it nearly indistinguishable from human-generated text (Scao et al., 2022). In this paper, we have used a smaller version of BLOOM, known as BLOOM-3B[8], which has 3 billion parameters.

- **GPT-2**. It is a pre-trained model using a Causal Language Modelling (CLM) target and a Transformer-based model that has undergone extensive pre-training on large corpora in a self-supervised manner (Radford et al., 2019). In this case, we used a Spanish version of GPT-2 called "Spanish GPT-2[9]". This model was trained from scratch on the large Spanish corpus, also known as the BETO corpus, using Flax.

---

[8] https://huggingface.co/bigscience/bloom-3b
[9] https://huggingface.co/mrm8488/spanish-gpt2

- **BART**. It is a transformer encoder-decoder (seq2seq) model with a bidirectional encoder, similar to BERT, and an autoregressive (GPT-like) decoder. BART is pre-trained by two key steps: (1) introducing noise into the text using a flexible noise function, and (2) training a model to recover the original, uncorrupted text. This model shows remarkable effectiveness when fine-tuned for text generation tasks such as summarization and translation. However, it also performs well in comprehension tasks, including text classification and question answering. Specifically, we have used *mBART-50* which is a pre-trained multilingual Sequence-to-Sequence model using the "Multilingual Denoising Pretraining" target (Lewis et al., 2020).

- **Llama-2**. Llama 2 contains a set of pre-trained and fine-tuned generative text models with parameters ranging from 7 billion to 70 billion. In this article we have used a version of 7B optimized for dialog cases (*meta-llama/Llama-2-7b-chat-hf*) (Touvron et al., 2023).

4.5 Ensemble learning

After acquiring different models for identifying emotions, we evaluated different techniques for ensemble learning, which consists of combining predictions from many individual estimators to produce a more robust estimator. During our research, we investigated two methods of averaging to combine predictions from the best encoder-only model and the best encoder-decoder model by (1) computing the mean, which averages the probabilities produced by each model; or (2) selecting the label with the highest probability, which involves observing the probabilities associated with each model and selecting the one with the highest probability.

5 RESULTS AND ANALYSIS

In this section, we present and discuss the results obtained using the encoder-only and encoder-decoder strategies and the feature integration techniques with the test split and compared to the baseline.

Since our problem involves unbalanced classification, we evaluate the performance of the classification models using both the macro average F1 score and the weighted average F1 score. These metrics serve as indicators of a classification model's performance in terms of accuracy and recall, albeit with different computational methods and weighting schemes. The macro average F1 score evaluates both the precision and recall of each class, combining the results equally without regard to class imbalance. In this way, it allows the selection of the best model that performs equally well for all labels. The weighted average F1 score is a widely used metric in classification problems, especially when dealing with unbalanced datasets where some classes may have many more examples than others. Unlike the macro average F1 score, this metric accounts for class imbalance by assigning a different weight to each class based on its frequency in the dataset. Therefore, in our work, we consider the macro average F1 score as the metric to determine which is the best model, but we keep the weighted average F1 score to reflect the overall performance of the model.

5.1 Results of encoder-only models

First, we report the results of several fine-tuned encoder-type models mentioned in section 4.3. For each model, we added a dense sequential classification layer with the same number of neurons as the output classes for fine tuning and performed hyperparameter optimization to find the best training parameters for these models using the validation split. The hyperparameters under consideration, along with their respective interval ranges, are: (1) weight decay (ranging from 0 to 0.3), (2) training lot size (ranging from 8 to 16), (3) number of training epochs (ranging from 1 to 6), and (4) learning rate (ranging from 1e-5 to 5e-5).

Table 2 shows the best set of hyperparameters obtained for each encoder-only model. It can be observed that most of the models, including ALBERT, BERTIN, Distilled mBERT, TwHIN-BERT, and XLM-RoBERTa, performed better with a training batch size of 16 and a lower learning rate. Most of these models also performed better with a warm-up step of 500, with the exception of BERTIN, BETO, and TwHIN-BERT, which performed better with a value of 250, and Distilled mBERT and XLM-RoBERTa, which performed best with a warm-up step of 0. In terms of weight decay, most of the models performed better with a value less than 0.2.

Table 3 shows the results obtained by the encoder-only models based on Transformers and the baseline. As expected, all models outperformed the baseline in terms of M-F1. In addition, the RoBERTa

Table 2. Best subset of hyperparameters for each encoder-only model based on Transformers.

	Learning rate	Epoch	Batch size	Warmup steps	Weight decay
ALBETO	3.5e-05	5	16	500	0.0006
BERTIN	2e-05	5	16	250	0.19
BETO	2.6e-05	4	8	250	0.00089
Distilled BETO	3.8e-05	3	8	500	0.21
MarIA	1.4e-05	4	8	500	0.069
mBERT	2.1e-05	3	8	500	0.25
Distilled mBERT	3.6e-05	5	16	0	0.26
TwHIN-BERT	2.7e-05	4	16	250	0.00076
XLM	3.7e-05	2	16	0	0.098

architecture consistently outperformed BERT. The top two scores were achieved by MarIA (60.47% in M-F1 and 79.42% in W-F1) and the knowledge integration model (58.70% in M-F1 and 78.91% in W-F1). In addition, XLM-RoBERTa outperforms the multilingual BERT (58.27% vs. 56.23% in M-F1). Notably, the lightweight versions of BETO, ALBETO and Distilled mBERT, yielded limited results.

When comparing the results of transformers trained on monolingual datasets with those trained on multilingual datasets, a slight advantage is observed for the models trained only in Spanish. This suggests that obtaining specific pre-trained models for the target language is preferable to using multilingual variants.

Table 3. Benchmark of the different pre-trained models and linguistic feature-based model. The metrics reported for each model and dataset include macro precision (M-P), macro recall (M-R), macro F1-score (M-F1), and weighted F1-score (W-F1).

	Architecture	Language	M-P	M-R	W-F1	M-F1
LF (baseline)	-	Mono	37.6551	41.8719	56.2573	36.9939
ALBETO	BERT	Mono	56.2657	56.3588	76.5421	55.7226
BERTIN	RoBERTa	Mono	56.6446	56.6979	76.9988	55.9180
BETO	RoBERTa	Mono	58.3675	57.8535	78.2100	58.0378
Distilled BETO	BETO	Mono	57.2375	58.0291	77.1875	57.4865
MarIA	RoBERTa	Mono	**61.9512**	**59.4110**	**79.4245**	**60.4771**
mBERT	RoBERTa	Multi	56.5545	56.9628	75.8372	56.2317
Distilled mBERT	BERT	Multi	53.7665	56.3917	74.1021	54.0067
TwHIN-BERT	BERT	Multi	59.0658	57.8394	78.7239	58.2436
XLM	RoBERTa	Multi	56.6756	58.2797	77.5739	58.2797
KI	-	-	59.1886	59.3930	78.9181	58.7096

5.2 Results of encoder-decoder models

In this study, we also evaluated different encoder-decoder models for text generation in EA tasks due to the promising results reported in several studies (Plaza-del Arco et al., 2023), (Brown et al., 2020a). Since these are sequence-to-sequence models, i.e., they take input text and produce output text, we added a linear layer to the pooled output for fine-tuning to ensure that the output corresponds to one of the emotions.

Some of these text generation models, such as BLOOM and Llama-2, are multilingual and pre-trained on a large corpus. The size of these models is quite large. BLOOM-3b has 3 billion parameters and weighs about 6 GB, while Llama-2 has 7 billion parameters and weighs about 13.5 GB. Therefore, we used the LoRA approach to speed up the fine-tuning process and reduce memory consumption. LoRA is based on representing weight updates with two smaller matrices called *update matrices* by low-rank decomposition. These new matrices can be trained to adapt to new data while keeping the total number of changes small. The original weight matrix remains frozen and does not receive any further adjustments.

To produce the final results, both the original and the adjusted weights are combined (Hu et al., 2021). Due to the size of the encoder-decoder models, we only performed hyperparameter optimization with the validation split for the epoch parameter within a range of 5 epochs. We kept other parameters constant: 0.01 for weight decay, 2e-5 for learning rate, and 500 warm-up steps.

Table 4 shows the results obtained with different encoder-decoder models are shown. It can be seen, that the best results are obtained with the Spanish GPT-2 (57.06% in M-F1 and 77.23% in W-F1). In addition, the Spanish GPT-2 is the only monolingual model and the lightest among the models compared. Therefore, we can draw the same conclusion as in the previous case (see Section 5.1), that obtaining specific pre-trained models for the target language is preferable over using multilingual variants. Regarding the multilingual models, the best result was obtained with multilingual BART, with an M-F1 of 52.16%.

Table 4. Benchmark of the different encoder-decoder models. Metrics reported for each model and dataset include macro precision (M-P), macro recall (M-R), macro F1-score (M-F1), and weighted F1-score (W-F1).

	M-P	M-R	W-F1	M-F1
LF (baseline)	37.6551	41.8719	56.2573	36.9939
Spanish GPT-2	**57.1349**	**57.1106**	**77.2334**	**57.0592**
BLOOM-3b	49.2585	45.9486	65.1833	45.5222
Llama-2	52.0913	52.5977	73.8725	51.2868
mBART	52.4731	52.1906	73.8514	52.1617

5.3 Results of ensemble learning

In this section, different ensemble learning techniques are evaluated using the best encoder-type model (MarIA) and encoder-decoder-type model (Spanish GPT-2) for emotion classification. Table 5 shows the results obtained. As can be seen, both the mean-based and the highest probability-based techniques have improved the weighted F1 score compared to MarIA, with an improvement of 0.0556% and 0.0227%, respectively. For the Spanish GPT-2, the ensemble learning techniques significantly improved the weighted F1 score, with improvements of 2.25% for the mean-based approach and 2.21% for the highest likelihood-based approach. However, ensemble learning did not improve the macro F1 score metrics because it did not improve predictions in emotion classes with fewer instances in the test set, such as *joy* and *grief*. However, it did improve predictions in some cases, such as the *depressed* emotion, which has more instances in the test set.

Table 5. Benchmark of the different ensemble learning techniques between the best pre-trained (MarIA) and encoder-decoder (Spanish GPT-2) model. The reported metrics for each model and dataset include macro precision (M-P), macro recall (M-R), macro F1-score (M-F1), and weighted F1-score (W-F1).

	M-P	M-R	W-F1	M-F1
LF (baseline)	37.6551	41.8719	56.2573	36.9939
MarIA	**61.9512**	**59.4110**	79.4245	**60.4771**
Spanish GPT-2	57.1349	57.1106	77.2334	57.0592
mean	61.5379	58.9295	**79.4801**	60.0464
highest probability	61.5675	58.8974	79.4472	60.0270

Table 6 shows the classification reports for the best single model, MarIA, and the best ensemble-learning model, mean-based ensemble learning. The results show that both strategies are similar for fine-grained emotion analysis. The big difference is observed in the *joy* emotion, where MarIA achieves an F1 score of 42.2535%, while the mean-based ensemble learning approach achieves only 31.250%, outperforming the results in both precision and recall. It is also observed that the ensemble model outperforms MarIA on several negative emotions such as *depressed*, *disappointment*, *embarrassment*,

508 *grief*, *nervousness*, *sadness*, and *suicidal*. In terms of precision and recall, both strategies show similar
509 behavior since both metrics are similar. However, ensemble learning based on the mean shows a larger
510 difference between precision and recall for the identification of texts labeled *joy*. Also, neither strategy
511 was able to identify the only instance of *surprise* in the test split.

Table 6. Classification report of Precision (P), Recall (R), and F1-score (F1) of MarIA (left) and Ensemble based on the mean (right) with the test set for each emotion.

Emotion	P	R	F1	P	R	F1
		MarIA			**Ensemble (mean)**	
anger	77.5934	72.2008	74.8000	77.6860	72.5869	75.0499
depressed	93.1330	92.4191	92.7747	93.4708	92.6746	93.0710
disappointment	54.8295	61.5303	57.9870	56.9573	65.2497	60.8222
disgust	44.2516	47.2222	45.6887	45.4756	45.3704	45.4229
embarrassment	34.2466	30.7377	32.3974	35.9813	31.5574	33.6245
fear	60.8911	57.6112	59.2058	60.6061	56.2061	58.3232
grief	41.0256	25.8065	31.6832	38.7755	30.6452	34.2342
hopeless	86.2222	78.5425	82.2034	87.9630	76.9231	82.0734
joy	45.4545	39.4737	42.2535	38.4615	26.3158	31.2500
lonely	96.0685	97.5435	96.8004	95.6871	97.6459	96.6565
nervousness	21.9780	16.6667	18.9573	22.2222	18.3333	20.0913
neutral	93.6057	94.1071	93.8557	90.8457	93.0357	91.9277
remorse	64.9425	56.7839	60.5898	62.3596	55.7789	58.8859
sadness	80.9091	83.1776	82.0276	81.4480	84.1121	82.7586
suicidal	96.0692	96.7538	96.4103	96.6667	96.4371	96.5517
surprise	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

512 Next, we analyze the limitations of these models in predicting emotions. For this purpose, we used
513 the confusion matrix to evaluate the examples misclassified by the MarIA model (see figure 4). We can
514 see that MarIA performs very well in predicting certain emotions such as *depressed*, *lonely*, *neutral*,
515 and *suicidal*, with an accuracy rate of over 90%. However, for more confusing and difficult-to-identify
516 emotions, such as *embarrassment*, *grief*, *joy*, *nervousness*, and *surprise*, the model tends to confuse them
517 with other similar emotions, achieving less than 40% accuracy for these emotions. In most cases, the
518 model confuses these emotions with the emotion *disappointment*.

519 Regarding the Spanish GPT-2, through the confusion matrix (see Figure 5), it can be observed that
520 its behavior is similar to that of MarIA, and it tends to make the same classification errors, except in
521 the case of *surprise*, where the model tends to confuse it with the emotion *embarrassment*. Figure 6
522 shows the confusion matrix of the mean-based ensemble learning model. Comparing it with MarIA's
523 matrix, we can see that ensemble learning has improved the predictions for the emotions *anger*, *depressed*,
524 *disappointment*, *disgust*, *nervousness*, and *sadness* by 1-3%, while it has worsened the predictions for the
525 other emotions in the same range, except for *grief*, which has worsened by up to 5%. For this reason, the
526 weighted F1 score of the ensemble learning is higher than that of MarIA, but not the macro F1 score. As
527 for the Spanish GPT-2 model, we can see that ensemble learning has improved the same emotions as in
528 the case of MarIA, but to a greater extent, in this case between 1% and 8%.

529 By analyzing the results and errors, we can see that pre-trained encoder models have achieved better
530 performance than encoder-decoder models for text generation. Furthermore, it has been shown that the
531 knowledge integration strategy improves the performance of most models, but it does not manage to
532 improve MarIA separately. Regarding ensemble learning techniques, we can observe that the approach
533 of combining the outputs by taking the mean of a fine-tuned encoder model (MarIA) with the outputs
534 of an encoder-decoder fine-tuned model (Spanish GPT-2) significantly improves the performance of
535 the fine-tuned encoder-decoder model, both in terms of weighted and macro F1 scores. It has also
536 improved the performance of predicting some of the more common categories in the test set for the
537 encoder fine-tuned model, resulting in an improvement in the weighted F1 score metric but not in the
538 macro F1 score. Thus, using the macro average F1 score as the reference metric, the MarIA model has
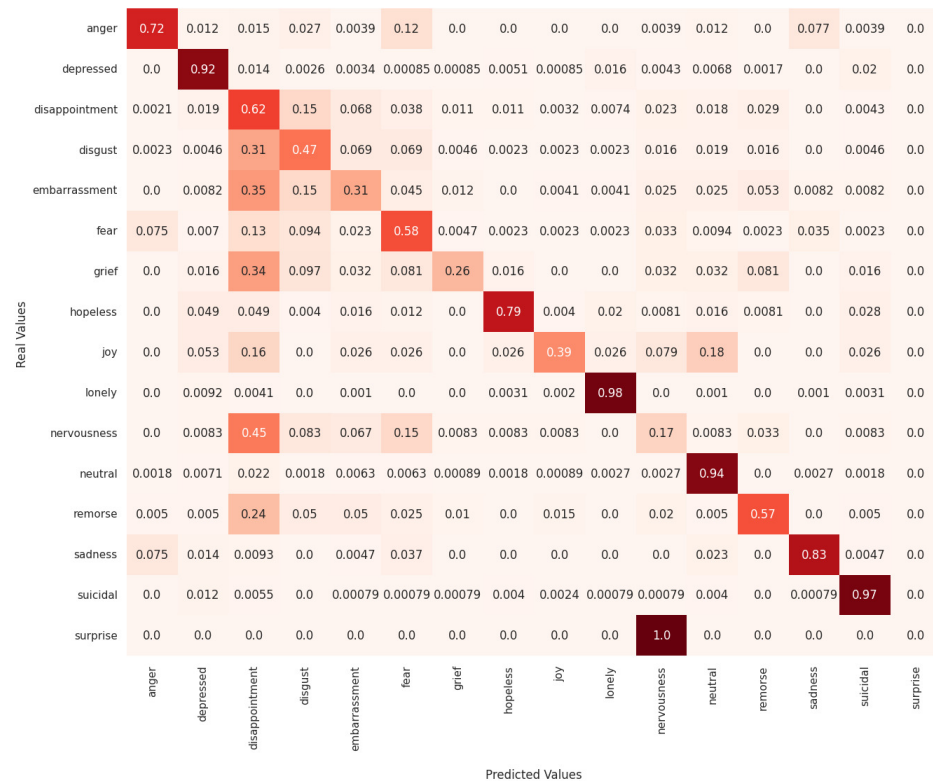
Figure 4. Confusion matrix of the MarIA model
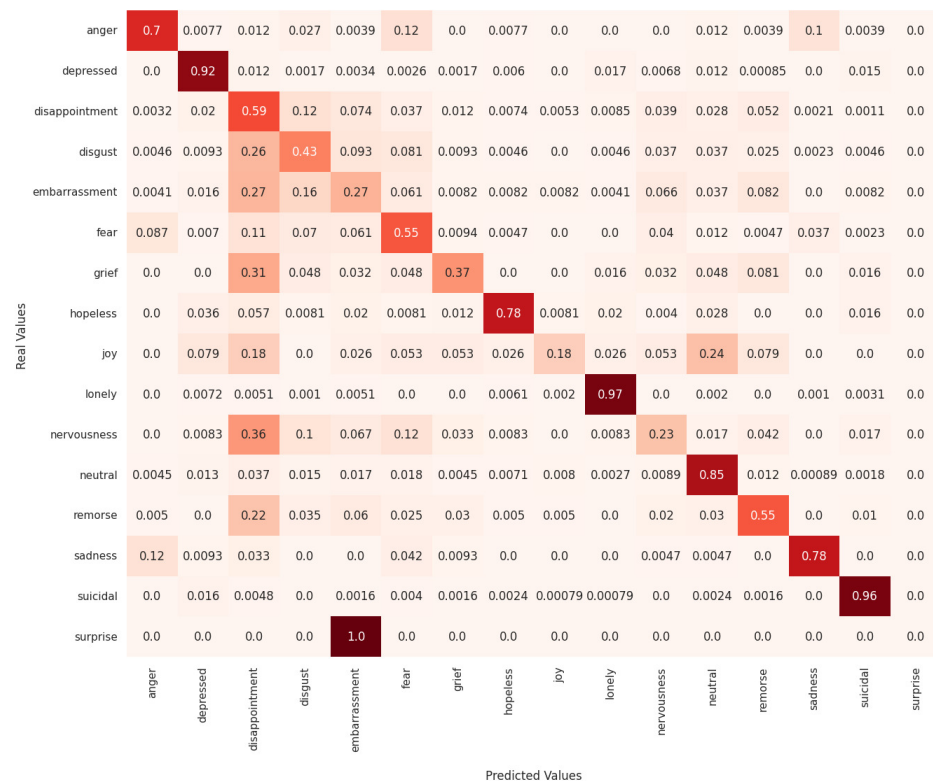
Figure 5. Confusion matrix of the Spanish GPT-2 model.
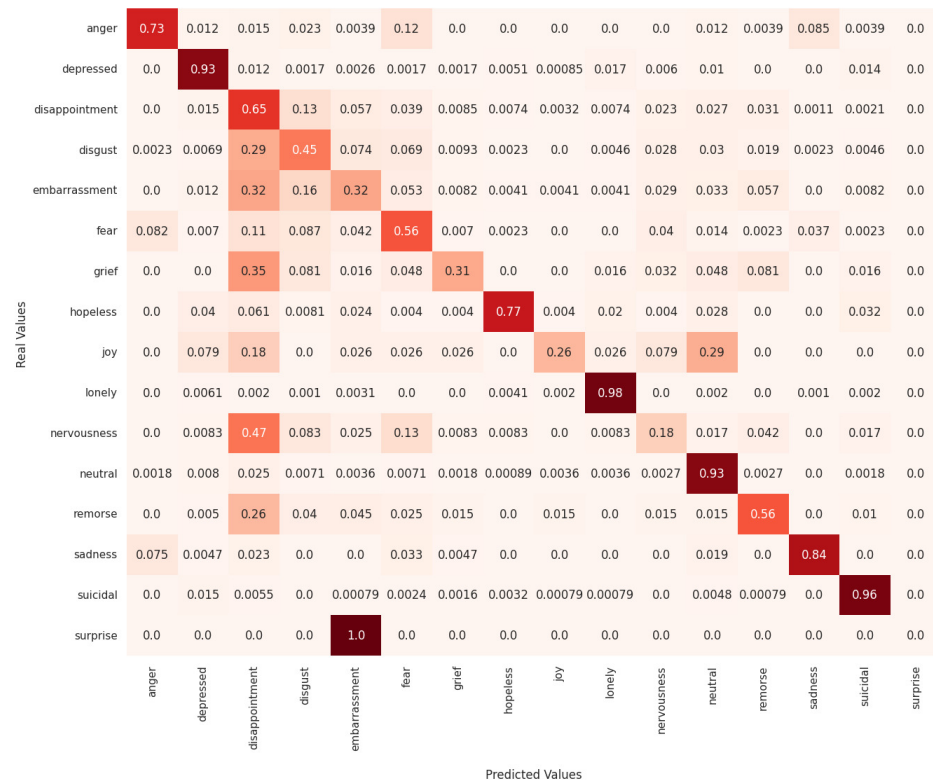
Figure 6. Confusion matrix of the ensemble learning mean-based model.

achieved the best result at 60.48%.

Finally, Table 7 shows examples of some common misclassifications made by the MarIA model. It is noticeable that the model often confuses emotions such as *embarrassment*, *grief*, and *nervousness* with the emotion of *disappointment*, as shown in Figure 4. This is because distinguishing between these emotions based on text alone is a challenging task, as shown in Table 7. Also, the emotion of disappointment has more examples in the training set compared to other emotions.

Table 7. Error Analysis with some examples of misclassifications made by MarIA. We include the text and a literal translation of the text into English using the Google Translation Service.

#	Text	Truth	Prediction
1	¡Gracias por explicarlo! No estoy seguro de por qué mi comentario fue rechazado Sin contexto, no tiene sentido (*Thanks for explaining it! Not sure why my comment was downvoted Without context, it doesn't make sense.*)	embarrassment	disappointment
2	Maldita sea, realmente necesitas una novia, amigo (*Damn, you really need a girlfriend, dude*)	embarrassment	disappointment
3	Maldita sea, [NOMBRE] se dejó llevar (*Damn, [NAME] got carried away*)	grief	disappointment
4	Qué pobre mujer que solo quería decirle a alguien qué hacer (*What a poor woman who just wanted to tell someone what to do.*)	grief	disappointment
5	No tengo confianza para impulsar (*I don't have the confidence to push*)	nervousness	disappointment
6	Pero en serio, probablemente te vas a sentir muy mal (*But seriously, you're probably going to feel really bad.*)	nervousness	disappointment
7	Pero eventualmente ese mismo amigo más cercano apareció en mi puerta y se negó a irse (*But eventually that same closest friend showed up at my door and refused to leave.*)	joy	disappointment

5.4 Evaluation with the EmoEvalES 2021 dataset

To measure the robustness of our pipeline, we have included an evaluation of the performance of both strategies with the EmoEvalEs 2021 dataset. Table 8 shows the results obtained. According to Plaza-Del-Arco et al. (2021), the reference metric used for the evaluation is accuracy. Our best results, the fine-tuning approach of MarIA and bloom-3b have surpassed the best previous result for this task (GSI-UPM with an accuracy of 72.77%), achieving an accuracy of 73.5507% and 73.1280% respectively. It draws our attention that in the EmoEvalES 2021 shared task, the GSI-UPM team achieved their best result by fine-tuning XLM-RoBERTa, but our results with this architecture are very limited (both in test and in custom validation). After reviewing the working notes of the GSI-UPM team, we noticed that they used a version of XLM, but pre-trained with millions of tweets (Barbieri et al., 2022). It is possible that the limited results are related to the complexity of the model, as it contains 16 layers, 16 attention heads, and 1280 hidden states, but further research should be done to see the real cause of the limited results of this model with the EmoEvalES 2021 dataset.

Table 8. Benchmark of different encoder-decoder models for the EmoEvalEs dataset compared to the best approach in the official leaderboard. Metrics reported for each model and dataset include accuracy.

Strategy	Approach	Accuracy	% Difference
GSI-UPM	XLM	72.7657	-
Encoder-only	ALBETO	70.8333	-1.9324
	BERTIN	68.4179	-4.3478
	BETO	70.3502	-2.4155
	Distilled BETO	70.5314	-2.2343
	MarIA	73.5507	+0.7850
	mBERT	66.4251	-6.3406
	Distilled mBERT	64.1908	-8.5749
	TwHIN	71.6787	-1.0870
	XLM	18.5990	-54.1667
Encoder-decoder	Spanish GPT-2	69.6256	-3.1401
	BLOOM-3b	73.1280	+0.3623
	Llama-2	70.1087	-2.657
	mBART	64.4968	-8.2689

After this analysis, we come to the following results regarding the proposed RQs. Regarding RQ1 about measuring the reliability of identifying negative fine-grained emotions, the results indicate that both precision and recall are usually good for some negative emotions (all the emotions in the dataset except *neutral*, *joy* and *surprise*). However, the developed EA models often fail to classify documents labeled as *disappointment*, confusing these emotions with others such as *disgust*, *embarrassment*, our *nervousness*. Regarding RQ2, to determine what is the best approach to face EA using the text, the best macro average F1 score is achieved with the fine-tuning of MarIA, a monolingual Spanish LLM based on the RoBERTa architecture. This strategy slightly outperformed other approaches based on fine-tuning generative language models. However, the ensemble learning strategy showed better results than MarIA in the negative emotions: (1) *depressed*, (2) *disappointment*, (3) *embarrassment*, (4) *grief*, (5) *nervousness*, (6) *sadness*, and (7) *suicidal*. In this sense, we also answer RQ3, which asks whether generative models are effective for EA, since they achieve similar performance as the fine-tuned models. However, the combination of the results outperforms the results obtained by the individual models, compared to our experiment where we combined all the fine-tuned LLMs using a knowledge integration strategy, in which we observed a degradation of performance compared to MarIA.

6 CONCLUSIONS AND FURTHER WORK

Mental health issues are a major global public health concern. Social media platforms are often one of the means by which users or individuals express their difficulties and find emotional support. Previous computational studies have consistently shown that individuals with mental disorders exhibit changes

577 in their language and behavior, including a higher prevalence of negative emotions and a more intense
578 self-focus. Therefore, clues to an individual's altered mental state may be evident in their online posts,
579 which often convey negativity.

580   This work presents two significant contributions to the field of EA for the detection of mental disorders
581 in Spanish. To this end, we have created a novel corpus of 16 different emotions and performed an
582 in-depth evaluation of several feature sets including linguistic features and transformer-based models
583 based on encoders and encoder-decoder. We have also tested different techniques for feature integration,
584 such as knowledge integration and ensemble learning, to see if they improve the performance of the
585 models. Our results show that the fine-tuning approach of the encoder-only MarIA model has achieved
586 the best result, with a macro F1 score of 60.48%.

587   As a limitation, statistical tests should be conducted to perform a better comparison of the models,
588 since using a fixed validation split can bias some decisions about the best performing models. In this
589 sense, we propose to extend this work using nested cross-validation for a better comparison. Moreover, as
590 commented during the corpus compilation and annotation process, we split long documents because of
591 the maximum length limitations of Transformers. In this sense, we will evaluate other strategies to handle
592 long documents, such as pooling the sentence embeddings or using Longformers.

593   As for future lines, we plan to expand the dataset to include less common emotions such as *surprise*,
594 *grief*, and *nervousness* in texts from different platforms, and to reduce the bias toward negative emotions.
595 In addition, we plan to explore different data augmentation techniques and investigate how emotions are
596 expressed within entities and their relationships in order to discover the emotions expressed by users.

597 ## REFERENCES

598 Acheampong, F. A., Nunoo-Mensah, H., and Chen, W. (2021). Transformer models for text-based emotion
599   detection: a review of bert-based approaches. *Artificial Intelligence Review*, pages 1–41.
600 Barbieri, F., Espinosa Anke, L., and Camacho-Collados, J. (2022). XLM-T: Multilingual language models
601   in Twitter for sentiment analysis and beyond. In Calzolari, N., Béchet, F., Blache, P., Choukri, K.,
602   Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and
603   Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*,
604   pages 258–266, Marseille, France. European Language Resources Association.
605 Batbaatar, E., Li, M., and Ryu, K. H. (2019). Semantic-emotion neural network for emotion recognition
606   from text. *IEEE Access*, 7:111866–111878.
607 Becker, K., Moreira, V. P., and dos Santos, A. G. (2017). Multilingual emotion classification using
608   supervised learning: Comparative experiments. *Information Processing and Management*, 53(3):684–
609   704.
610 Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P.,
611   Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A.,
612   Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark,
613   J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020a). Language models
614   are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors,
615   *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates,
616   Inc.
617 Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P.,
618   Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh,
619   A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B.,
620   Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020b). Language
621   models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
622 Canales, L. and Martínez-Barco, P. (2014). Emotion detection from text: A survey. In *Proceedings of
623   the workshop on natural language processing in the 5th information systems research working days
624   (JISIC)*, pages 37–43.
625 Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., and Pérez, J. (2020). Spanish pre-trained bert
626   model and evaluation data. In *PML4DC at ICLR 2020*, pages 1–10.
627 Cañete, J., Donoso, S., Bravo-Marquez, F., Carvallo, A., and Araujo, V. (2022). ALBETO and DistilBETO:
628   Lightweight spanish language models. In *Proceedings of the Thirteenth Language Resources and
629   Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 4291–4298. European
630   Language Resources Association.

Compare, A., Zarbo, C., Shonin, E., Van Gordon, W., and Marconi, C. (2014). Emotional regulation and depression: A potential mediator between heart and mind. *Cardiovascular psychiatry and neurology*, 2014.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Cowen, A. S. and Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, 114:E7900 – E7909.

De Arriba, A., Oriol, M., and Franch, X. (2021). Merging datasets for emotion analysis. In *2021 36th IEEE/ACM International Conference on Automated Software Engineering Workshops (ASEW)*, pages 227–231. IEEE.

De Choudhury, M., Counts, S., Horvitz, E. J., and Hoff, A. (2014). Characterizing and predicting postpartum depression from shared facebook data. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 626–638.

de la Rosa, J., Ponferrada, E. G., Romero, M., Villegas, P., González de Prado Salas, P., and Grandury, M. (2022). BERTIN: efficient pre-training of a spanish language model using perplexity sampling. *Proces. del Leng. Natural*, 68:13–23.

Dejonckheere, E., Kalokerinos, E. K., Bastian, B., and Kuppens, P. (2019). Poor emotion regulation ability mediates the link between depressive symptoms and affective bipolarity. *Cognition and Emotion*, 33(5):1076–1083.

Dejonckheere, E., Mestdagh, M., Houben, M., Erbas, Y., Pe, M., Koval, P., Brose, A., Bastian, B., and Kuppens, P. (2018). The bipolarity of affect and depressive symptoms. *Journal of personality and social psychology*, 114(2):323.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., and Ravi, S. (2020). GoEmotions: A Dataset of Fine-Grained Emotions. In *58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Ekman, P. and Davidson, R. J. (1993). Voluntary smiling changes regional brain activity. *Psychological Science*, 4(5):342–345.

Fatima, A., Li, Y., Hills, T. T., and Stella, M. (2021). Dasentimental: Detecting depression, anxiety, and stress in texts via emotional recall, cognitive networks, and machine learning. *Big Data and Cognitive Computing*, 5(4).

Garcıa-Dıaz, J. A., Colomo-Palacios, R., and Valencia-Garcıa, R. (2021). Umuteam at emoevales 2021: Emotion analysis for spanish based on explainable linguistic features and transformers. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, volume 2943 of *CEUR Workshop Proceedings*, pages 59–71. CEUR-WS.org.

García-Díaz, J. A., Vicente, P. J. V., Almela, Á., and Valencia-García, R. (2022). Umutextstats: A linguistic feature extraction tool for spanish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 6035–6044. European Language Resources Association.

Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., and Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49. Big data in the behavioural sciences.

Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pàmies, M., Llop-Palao, J., Silveira-Ocampo, J., Carrino, C. P., Armentano-Oller, C., Penagos, C. R., Gonzalez-Agirre, A., and Villegas, M. (2022). MarIA: Spanish language models. *Proces. del Leng. Natural*, 68:39–60.

He, P., Gao, J., and Chen, W. (2021). DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. *CoRR*, abs/2111.09543.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models. *CoRR*.

Joormann, J. and Gotlib, I. H. (2010). Emotion regulation in depression: Relation to cognitive inhibition. *Cognition and Emotion*, 24(2):281–298.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2017). Bag of tricks for efficient text classification. In Lapata, M., Blunsom, P., and Koller, A., editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Lample, G. and Conneau, A. (2019). Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Lazarus, R. S. and Lazarus, B. N. (1994). *Passion and Reason: Making Sense of Our Emotions*. Oxford University Press USA.

Leis, A., Ronzano, F., Mayer, M. A., Furlong, L. I., and Sanz, F. (2019). Detecting signs of depression in tweets in spanish: behavioral and linguistic analysis. *Journal of medical Internet research*, 21(6):e14199.

Leiva, V. and Freire, A. (2017). Towards suicide prevention: early detection of depression on social media. In *Internet Science: 4th International Conference, INSCI 2017, Thessaloniki, Greece, November 22-24, 2017, Proceedings 4*, pages 428–436. Springer.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Medhat, W., Hassan, A., and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In Bengio, Y. and LeCun, Y., editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Mohsin, M. A. and Beltiukov, A. (2019). Summarizing emotions from text using plutchik's wheel of emotions. In *7th Scientific Conference on Information Technologies for Intelligent Decision Making Support (ITIDS 2019)*, pages 291–294. Atlantis Press.

Munezero, M., Montero, C. S., Sutinen, E., and Pajunen, J. (2014). Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE transactions on affective computing*, 5(2):101–111.

Murthy, A. R. and Kumar, K. M. A. (2021). A review of different approaches for detecting emotion from text. *IOP Conference Series: Materials Science and Engineering*, 1110(1):012009.

Nabeel, Z., Mehmood, M., Baqir, A., and Amjad, A. (2021). Classifying emotions in roman urdu posts using machine learning. In *2021 Mohammad Ali Jinnah University International Conference on Computing (MAJICC)*, pages 1–7. IEEE.

Nandwani, P. and Verma, R. (2021a). A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11.

Nandwani, P. and Verma, R. (2021b). A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11(1):81.

Pan, R., Díaz, J., and Valencia-García, R. (2023a). Umuteam at erisk clef 2023 shared task: transformer models for early detection of pathological gambling, depression, and eating disorder. *Working Notes of CLEF*, pages 18–21.

Pan, R., Garcıa-Dıaz, J., and Valencia-Garcıa, R. (2023b). Umuteam at mental-riskes2023 iberlef: Transformer and ensemble learning models for early detection of eating disorders and depression. In *IberLEF (Working Notes). CEUR Workshop Proceedings*.

Parapar, J., Martín-Rodilla, P., Losada, D. E., and Crestani, F. (2023). Overview of erisk 2023: Early risk prediction on the internet. In Arampatzis, A., Kanoulas, E., Tsikrika, T., Vrochidis, S., Giachanou, A., Li, D., Aliannejadi, M., Vlachos, M., Faggioli, G., and Ferro, N., editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 294–315, Cham. Springer Nature Switzerland.

Park, S.-H., Bae, B.-C., and Cheong, Y.-G. (2020). Emotion recognition from text stories using an emotion embedding model. In *2020 IEEE international conference on big data and smart computing (BigComp)*, pages 579–583. IEEE.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In Walker, M., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Plaza-del Arco, F. M., Jiménez-Zafra, S. M., Montejo-Ráez, A., Molina-González, M. D., Ureña-López, L. A., and Martín-Valdivia, M. T. (2021). Overview of the emoevales task on emotion detection for spanish at iberlef 2021. *Procesamiento del Lenguaje Natural*, 67:155–161.

Plaza-del Arco, F. M., Nozza, D., and Hovy, D. (2023). Leveraging label variation in large language models for zero-shot text classification. *arXiv preprint arXiv:2307.12973*.

Plaza-Del-Arco, F. M., Zafra, S. M. J., Ráez, A. M., González, M., López, L. A. U., and Valdivia, M. T. M. (2021). Overview of the emoevales task on emotion detection for spanish at iberlef 2021. *Proces. del Leng. Natural*, 67:155–161.

Plutchik, R. (1980). Chapter 1 - a general psychoevolutionary theory of emotion. In Plutchik, R. and Kellerman, H., editors, *Theories of Emotion*, pages 3–33. Academic Press.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. Technical report, OpenAI.

Rehm, J. and Shield, K. D. (2019). Global burden of disease and the impact of mental and addictive disorders. *Current Psychiatry Reports*, 21:1–7.

Saffar, A. H., Mann, T. K., and Ofoghi, B. (2023). Textual emotion detection in health: Advances and applications. *Journal of Biomedical Informatics*, 137:104258.

Sangeetha, K. and Dhandayudam, P. (2021). Sentiment analysis of student feedback using multi-head attention fusion model of word and context embedding for lstm. *Journal of Ambient Intelligence and Humanized Computing*, 12.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilic, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., Tow, J., Rush, A. M., Biderman, S., Webson, A., Ammanamanchi, P. S., Wang, T., Sagot, B., Muennighoff, N., del Moral, A. V., Ruwase, O., Bawden, R., Bekman, S., McMillan-Major, A., Beltagy, I., Nguyen, H., Saulnier, L., Tan, S., Suarez, P. O., Sanh, V., Laurençon, H., Jernite, Y., Launay, J., Mitchell, M., Raffel, C., Gokaslan, A., Simhi, A., Soroa, A., Aji, A. F., Alfassy, A., Rogers, A., Nitzav, A. K., Xu, C., Mou, C., Emezue, C., Klamm, C., Leong, C., van Strien, D., and Adelani, D. I. (2022). Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Singh, M., Jakhar, A. K., and Pandey, S. (2021). Sentiment analysis on the impact of coronavirus in social life using the bert model. *Social Network Analysis and Mining*, 11(1):33.

Skaik, R. and Inkpen, D. (2020). Using social media for mental health surveillance: a review. *ACM Computing Surveys (CSUR)*, 53(6):1–31.

Strapparava, C. and Valitutti, A. (2004). Wordnet affect: an affective extension of wordnet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal*, volume 4, pages 1083–1086. Lisbon, Portugal.

Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Canton-Ferrer, C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., and Hajishirzi, H. (2023). Self-instruct: Aligning language models with self-generated instructions. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

World Health Organization and others (2022). Mental health and covid-19: early evidence of the pandemic's impact: scientific brief, 2 march 2022. Technical report, World Health Organization.

Zhang, T., Yang, K., Ji, S., and Ananiadou, S. (2023a). Emotion fusion for mental illness detection from social media: A survey. *Information Fusion*, 92:231–246.

Zhang, X., Malkov, Y., Florez, O., Park, S., McWilliams, B., Han, J., and El-Kishky, A. (2023b). TwHIN-BERT: A socially-enriched pre-trained language model for multilingual tweet representations. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5597–5607.