applied sciences

MDPI

Article

How Successful Is Transfer Learning for Detecting Anorexia on Social Media?

Pilar López-Úbeda *, Flor Miriam Plaza-del-Arco, Manuel Carlos Díaz-Galiano and Maria-Teresa Martín-Valdivia

SINAI Group, Campus Las Lagunillas s/n, CEATIC—Universidad de Jaén, E-23071 Jaén, Spain; fmplaza@ujaen.es (F.M.P.-d.-A.); mcdiaz@ujaen.es (M.C.D.-G.); maite@ujaen.es (M.-T.M.-V.)
* Correspondence: plubeda@ujaen.es

Abstract: Anorexia is a mental disorder that involves serious abnormalities in nutritional intake behavior. This behavior leads to significant weight loss, which can lead to severe malnutrition. Specifically, eating disorders exhibit the highest mortality rate of any mental illness. Early identification of anorexia, along with appropriate treatment, improves the speed of recovery in patients. Presently there is a strong and consistent association between social media use and eating concerns. Natural Language Processing, a branch of artificial intelligence, has the potential to contribute towards early anorexia detection in textual data. Currently, there is still a long way to go in the identification of anorexia on social media due to the low number of texts available and in fact, most of these are focused on the treatment of English texts. The main contribution of this paper is the application of transfer learning techniques using Transformer-based models for detecting anorexia in tweets written in Spanish. In particular, we compare the performance between already available multilingual and monolingual models, and we conduct an error analysis to understand the capabilities of these models for Spanish.

Keywords: anorexia detection; transfer learning; BERT; text classification; Natural Language Processing

Citation: López-Úbeda, P.; Plaza-del-Arco, F.M.; Díaz-Galiano, M.C.; Martín-Valdivia, M.-T. How Successful Is Transfer Learning for Detecting Anorexia on Social Media? *Appl. Sci.* **2021**, *11*, 1838. https://doi.org/10.3390/app11041838

Received: 15 January 2021
Accepted: 5 February 2021
Published: 19 February 2021

1. Introduction

Mental illness is a leading cause of disability worldwide and continues to increase, with significant health implications and major social, human rights, and economic consequences on a global scale [1]. Globally, an estimated 264 million people are affected by depression, about 45 million people are suffering from bipolar disorder and about 20 million have schizophrenia [2].

According to the American psychiatric association (https://www.psychiatry.org/patients-families/eating-disorders/what-are-eating-disorders), "eating disorders are illnesses in which the people experience severe disturbances in their eating behaviors and related thoughts and emotions. People with eating disorders typically become pre-occupied with food and their body weight". There are three main types of eating disorders: anorexia nervosa, bulimia nervosa, and eating disorder.

Early detection in eating disorders is important to increase the chances of recovery, and NLP techniques can be used as a tool for detecting and predicting disorders in individuals by helping professionals identify these types of diseases as early as possible. Indeed, the presence of this type of disorder is increasing every day on social media platforms [3].

In this study, we focus on conducting classification experiments based on Machine Learning (ML) to detect anorexia nervosa in Twitter comments. The main contributions of this paper can be summarized as follows:

1. We perform anorexia detection in Spanish tweets due to the limited availability of studies aimed at detecting this type of disorder, most of which have been conducted in English.

2. We employ state-of-the-art systems based on transfer learning – including Transformer-based models such as BERT and XLM (see Section 4.2). To the best of our knowledge, this type of method has not yet been applied to detecting anorexia in informal texts written in Spanish.
3. We compare the results obtained with deep learning systems – including LSTMs and CNNs with Transformer-based models to observe the behavior of this type of system in tweets written in Spanish (see Section 5).
4. We compare the monolingual Transformer-based model (BETO) with two multilingual ones (M-BERT and XLM) to explore whether BETO trained on Spanish texts outperform multilingual models in the Spanish anorexia task.
5. We study the behavior of the algorithms explored in Spanish. Most of the studies for the detection of anorexia focused on English. However, anorexia detection is a global concern involving all languages. In particular, Spanish is the second most spoken language in the world and the third most used in social networks, which shows the need to develop algorithms capable of detecting psychological disorders in social media.
6. We perform an error analysis to identify the weaknesses and capabilities of the systems.

The rest of the paper is structured as follows: In Section 2 some previous related studies are described. The data we used to evaluate our experiments are described in Section 3. The experimental methodology is laid out in Section 4. The evaluation of the results is presented in Section 5. Finally, the analysis of errors is conducted in Section 6, and conclusions are presented in Section 7.

2. Related Work

NLP techniques can be used to make predictions about people with mental health problems by analyzing the messages they write on Facebook, Twitter, and other social media [4,5]. These texts contain informal language in which users give information and express their thoughts, feelings, and moods about their daily lives [6,7]. There are several studies that cover a range of mental health topics including predicting depression diagnosis [8], assessing suicide risk [9,10], schizophrenia [11,12], stress [13] and obsessive-compulsive disorder [14].

As the prediction of the first symptoms of mental illness has received considerable attention, several competitions such as eRisk and CLPsych have recently taken place. In 2017, eRisk (https://erisk.irlab.org/) at CLEF (Conference and Labs of the Evaluation Forum) proposed an author profiling task where the aim is to identify specific mental conditions such as depression [15]. Presently eRisk encompasses the groups interested in this type of challenge, that of presenting computer models that can early identify users with symptoms of depression and anorexia [16]. Another workshop is CLPsych (https://clpsych.org/) (Computational Linguistics and Clinical Psychology) which has been held until 2019 as part of the yearly conference organized by the American Association of Computational Linguistics (NAACL). The participation in this conference shows the advances in the detection of current problems such as depression, Alzheimer's, autism, violence, and aphasia [17]. However, all these clinics and workshops are focused on English texts.

Eating disorders are serious mental illnesses, and anorexia nervosa is associated with the highest mortality rate of any psychiatric illness [18], and for this reason, there is a need to study these types of diseases and detect them early. Wolf et al. [19] conducted a study in which they found that the writings of pro-food-disordered bloggers have a more closed style and are less emotionally expressive. These findings indicate that writing can be analyzed with NLP techniques to detect whether or not a person may have signs of anorexia.

Many studies have been conducted with the aim of detecting mental illness automatically using ML techniques. Conway et al. [20] recently conducted a review of the most widely used methods in NLP to identify cancer, mental health issues, substance abuse, and

so on, and the most commonly used social networks in those studies. Most of the papers that use classical ML approaches [8,21–23], report that the most widely used model is Logistic Regression. Some of these studies employ the LIWC tool [24] to add new features to the model. In addition, it should be noted that they use different methods for word representation such as word embeddings, bag-of-words, and tf-idf.

Regarding neural networks, these have been successfully applied to many different domains and scenarios [25–27]. Specifically, in the detection of mental illness, we have found a small amount of related work. Ive et al. [28] used a Recurrent Neural Network (RNN) with an attention mechanism that improved their baseline for detecting diseases such as bipolarity, addictions, autism, and depression, among others. Convolutional Neural Networks (CNN) have also been applied to the analysis of mental illness [29], obtaining values above 91% accuracy in binary classification and above 62% in multi-class classification.

Recent state-of-the-art studies in NLP have been exploring techniques with pre-training and bidirectional language representation, to understand the semantic complexity of a language [30]. Bidirectional Encoder Representations from Transformers (BERT) [31] was used in a multimodal approach for automatic depression detection [32].

However, we realize that several Transformer-based models have not yet been tested for the detection of anorexia in Spanish tweets. Therefore, our paper differs from the previous studies because: (i) we focus on the detection of anorexia in Spanish tweets (ii) we perform a comparative study of deep learning architectures (LSTMs and CNNs) with novel Transformer-based models, and (iii) we analyze how multilingual and monolingual language models based on Transformer mechanism differ in terms of vocabulary coverage for Spanish.

3. Data

In this section, we introduce the dataset we used to train and evaluate our models. As far as we know, there is only available one dataset annotated with anorexia in Spanish which is called SAD (Spanish Anorexia Dataset) [33]. To build the dataset, the authors collected tweets and annotated them automatically using different hashtags as queries. Specifically, they employ the hashtag #anaymia to label the tweets as anorexia and other hashtags related to the vocabulary of food such as #comidareal #realfood #fitness to annotate the tweets as control. Finally, the dataset is comprised of 5707 tweets, 2707 annotated as positive (anorexia) and 3000 annotated as negative (control). In addition, the authors reveal some interesting statistics about the mood of users and how they express themselves through social media, revealing that users with anorexia disorder tend to use more negative language than users without anorexia.

4. Methodology

In this section, we explore different methods applied to detecting anorexia in Spanish tweets. On the one hand, we show neural network approaches in Section 4.1 and on the other hand, newer methods of transfer learning using Transformer-based models in Section 4.2.

4.1. Deep Neural Networks

Deep learning models are composed of several processing layers, which allows the models to learn more complex representations of the data by taking into consideration multiple levels of abstraction. One of the main differences between neural networks and traditional ML approaches is the ability of neural networks to learn complex feature representations. We discuss three of the most commonly used neural network models that are described below.

4.1.1. Long Short-Term Memory

Hochreiter and Schmidhuber [34] showed a variety of neural networks named Long Short-Term Memory (LSTM). LSTM contains a special hidden unit that acts as a memory cell and uses a gradient-based back-propagation technique that allows the selective retention of relevant information from a previous step while the input sequence is being parsed element by element.

4.1.2. Bidirectional Long Short-Term Memory

Bidirectional Long Short-Term Memory (BiLSTM) is an extension of traditional LSTM that can improve model performance on sequence classification problems [35]. BiLSTM trains two instead of one LSTM on the input sequence, the first one using the normal input sequence, and the second one on a reversed copy of the input sequence. This can provide additional context to the network and make learning faster and more complete about the problem.

4.1.3. Convolutional Neural Networks

Convolutional Neural Networks (CNNs) were originally designed for deep learning computer vision tasks, but they have also proven to be useful for NLP tasks [36]. In CNN, every network layer acts as a detection filter for the presence of specific features or patterns present in the original data. The first layers in a CNN detect large features, later layers detect increasingly smaller sets of more abstract features, and the last layer of the CNN can perform a classification by combining all the specific features detected by the previous layers in the input data.

We tested these networks with different hyperparameter values and the ones that performed best are presented below. We perform 10-fold cross-validation with the training dataset, in this way we get the best hyperparameters over 20 different combinations with Bayesian Optimization. Finally, we used the test set to predict and evaluate the predictions using the best hyperparameters which are presented below. We used 100 units in the case of CNN and 150 units in the case of LSTM. BiLSTM uses 100 units in each LSTM. A batch size of 128 for BiLSTM and CNN and 64 for LSTM was employed. A max-pooling layer was appended to the CNN model. For all neural networks, a dense layer of size 50 with Rectified Linear Unit (ReLU) activation and a dropout function was added to help prevent overfitting. Specifically, we use a dropout rate of 0.25. This was followed by a dense layer with a size of 1 to represent the number of classification classes with a sigmoid function determining the output. Figures 1–3 show the architectures followed for the LSTM, BiLSTM and CNN models, respectively.

For all previous neural networks, we used an embedding layer as input. Word embedding is a form of representing words and documents using a dense vector representation. The position of a word within the vector space is learned from the text and is based on the words that surround the word. This allows words that are used in a similar way to have a similar representation, capturing their meaning.

The vocabulary of the corpus is composed of 12,313 unique words and Table 1 shows the number of words included in the Spanish pre-trained embeddings (https://github.com/dccuchile/spanish-word-embeddings).

Table 1. Analysis of the number of words included in word embeddings.

Corpus	Algorithm	Dimension	Included	Non-Included
Spanish Unannotated Corpora	FastText	300	11,032	1281
Wikipedia Spanish	FastText	300	10,469	1844
Spanish Billion Word Corpus	GloVe	300	10,394	1919

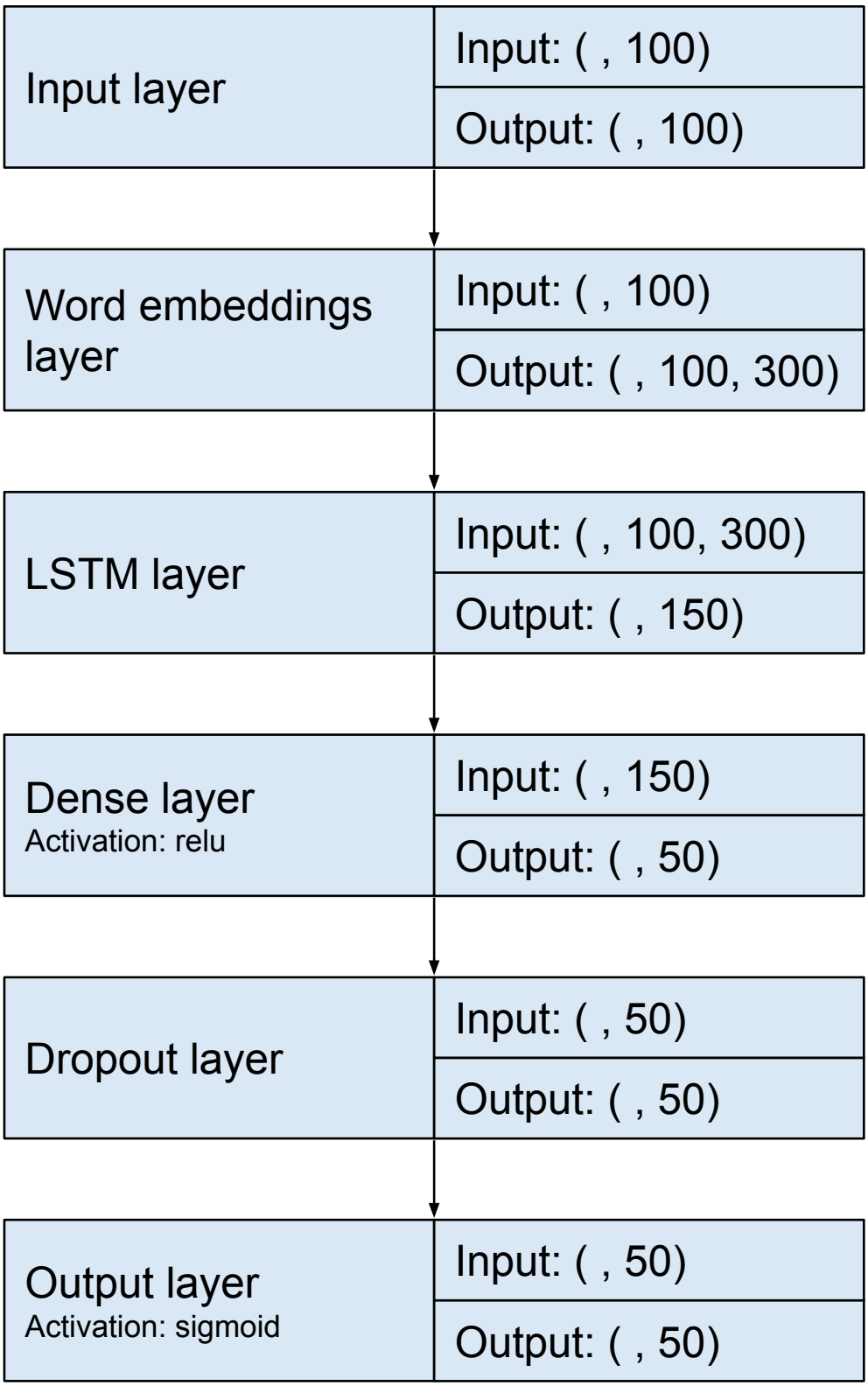

Figure 1. LSTM architecture.

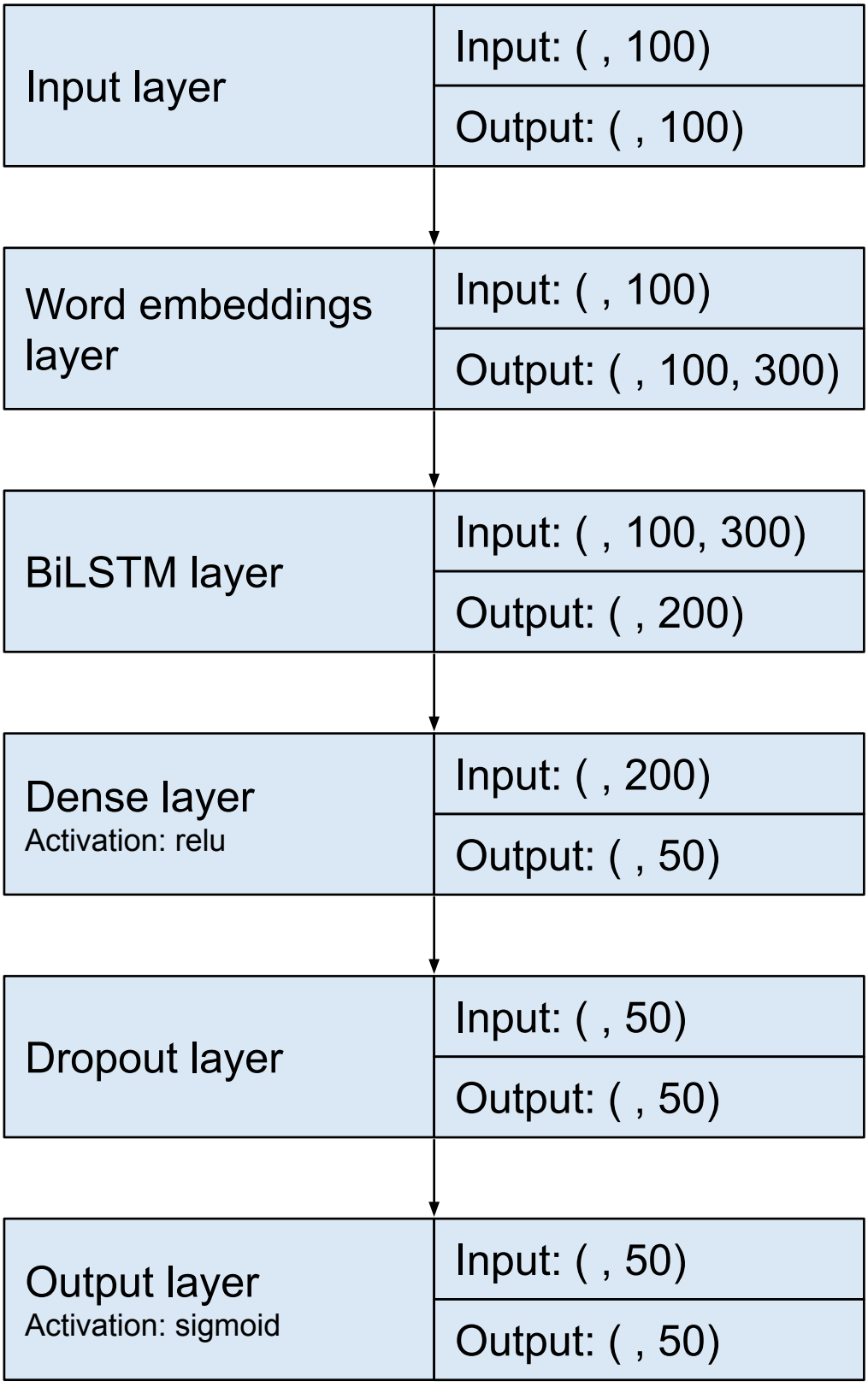

Figure 2. BiLSTM architecture.

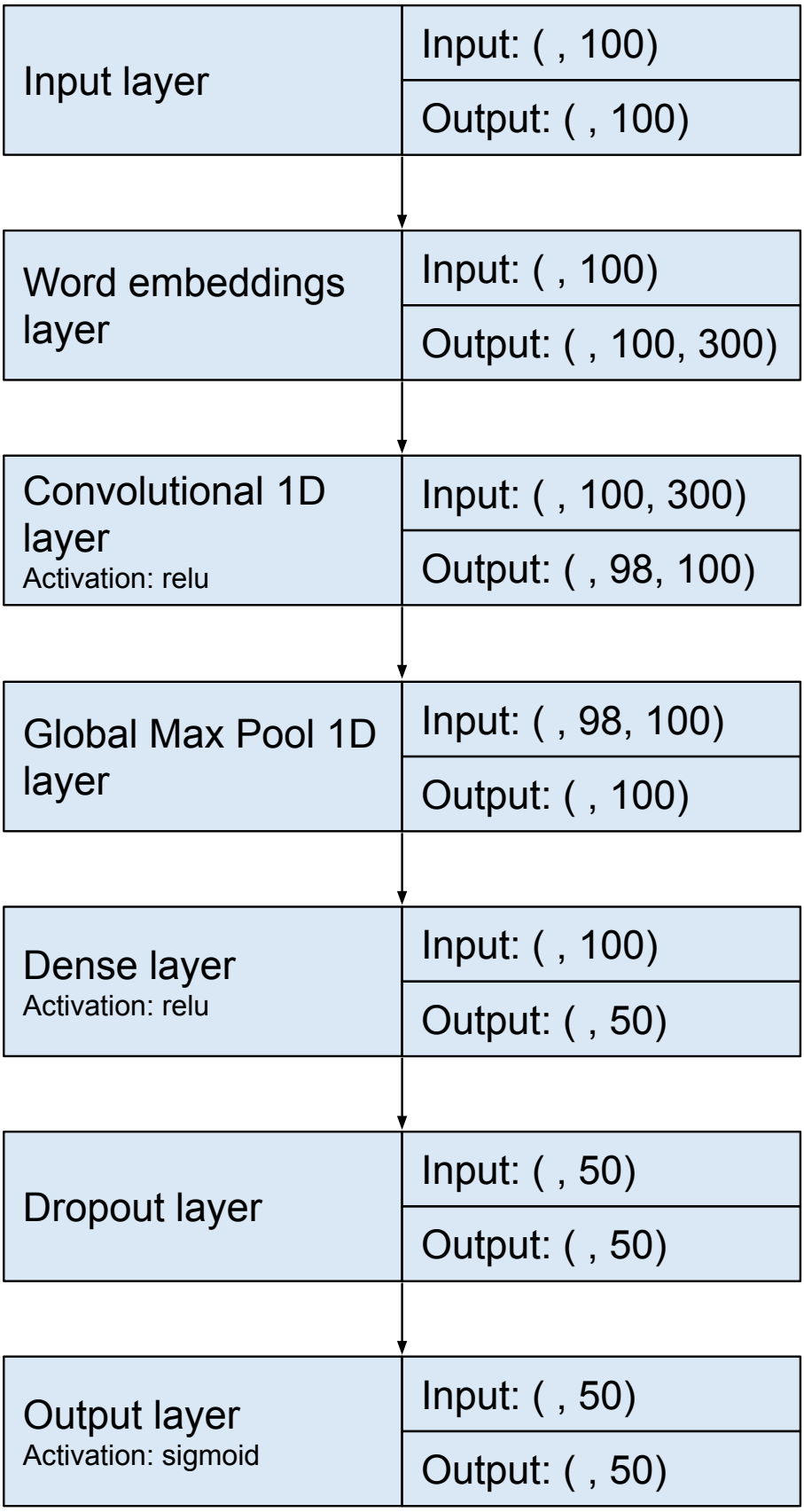

Figure 3. CNN architecture.

After analyzing the existing embeddings in Spanish in our experiments we used the embeddings extracted from the Spanish Unannotated Corpora [37] because this is the model that contains the most words from our vocabulary.

4.2. Transformer-Based Models

Transfer learning refers to the ML paradigm in which an algorithm extracts knowledge from one or more scenarios to help to learn performance in a specific scenario. Transfer learning has emerged as a highly popular technique in developing deep learning models [38]. With this technique the neural network is trained in two stages: (1) pre-training, where the network is generally trained on a large dataset representing a wide diversity of labels or categories; and (2) fine-tuning, where the pre-trained network is further trained on the specific task, which may have fewer labeled examples than the pre-training dataset. The first step helps the network learn general features that can be reused for the target task [39].

Transformer-based models are based on transformers architecture, they introduce an attention mechanism that processes the entire text input simultaneously in order to learn contextual relations between words or sub-words [40]. The Transformer includes two parts: an encoder that reads the text input and generates a vector for each word, and a decoder that produces the translated text from that representation.

In this study, we address the detection of anorexia in Spanish using different Transformer-based models. To this end, we use models that contain or are trained on that language such as BETO and other multilingual models such as BERT [31] and XLM [41]. These models are to be summarized in the following sections.

The BERT model uses a hidden size of 768, 12 Transformer blocks, and 12 self-attention heads. For the optimizer, we leverage the *adam* optimizer which performs well for NLP data and for BERT models in particular. The BERT architecture followed is shown in Figure 4 with an example extracted from the SAD corpus. On the contrary, XLM uses a hidden size of 1280, 16 Transformer blocks, and 16 self-attention heads. For the purposes of fine-tuning, the authors recommend choosing from the following values: batch size, learning rate, max sequence, and several epochs.

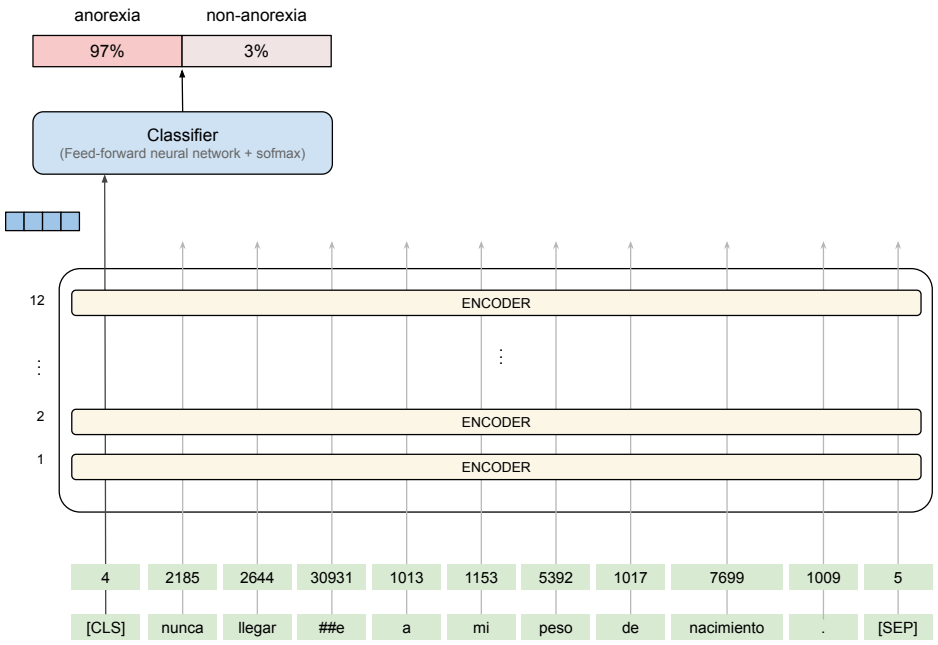

Figure 4. The architecture of BERT.

4.2.1. Spanish BERT

BETO is an initiative to allow the use of BERT pre-trained models for Spanish NLP tasks. This model is trained from a large Spanish corpus (https://github.com/dccuchile/beto) and is similar in size to BERT for English [42].

For the fine-tuning step, we tested with several optimization hyperparameters and finally we train with a learning rate of 0.00002 and three epochs.

4.2.2. Multilingual BERT

The multilingual method has promoted the state-of-the-art on cross-lingual understanding tasks by jointly pre-training large Transformer models. Multilingual BERT (henceforth, M-BERT), proposed by Devlin et al. [31], is a single language model pre-trained on the concatenation of monolingual Wikipedia corpora from 104 languages including Spanish (https://github.com/google-research/bert) [43]. In this model, we use a batch size of 16, a learning rate of 0.00003, and three epochs.

4.2.3. Cross-Lingual Language Models: XLM

XLM uses a known pre-processing technique named Byte Pair Encoding (BPE) [44] and a dual-language training mechanism with BERT to learn relations between words in different languages. In this study, we show the performance of two XLM models: XLM trained with 17 languages (XLM-17) and XLM trained with 100 languages (XLM-100). In both models, we trained with a batch size of 8, a learning rate of 0.00003, and three epochs.

We use as programming language Python Version 3.8 (Code available at: https://github.com/plubeda/binaryclass_transformers_beto). Moreover, Transformers [45] library by Huggingface (https://huggingface.co) was used to build the BERT network and the tokenizer from available BETO models.

Finally, all experiments (training and evaluation) were performed on a node equipped with two Intel Xeon Silver 4208 CPU at 2.10 GHz, 192 GB RAM, as main processors, and six GPUs NVIDIA GeForce RTX 2080Ti (with 11 GB each).

5. Results

In this section, we report the results we have obtained by evaluating the systems based on ML. To do so we employ the usual metrics in NLP tasks, including Precision (P), Recall (R), F1-score (F1) and macro-average. The metrics are computed as follows:

$$P(c) = \frac{TP}{TP + FP} \tag{1}$$

$$R(c) = \frac{TP}{TP + FN} \tag{2}$$

where c is equal to the anorexia class (0, 1), TP = True Positive, FP = False Positive and FN = False Negative.

$$F1 = \frac{2 * P * R}{P + R} \tag{3}$$

We compare the performance of the different models on the SAD dataset. For this purpose, we evaluate our models using the 10-fold cross-validation technique. Table 2 shows the prediction performance for each model.

Table 2. Results obtained on the SAD corpus.

	Anorexia			Control			Macro-Avg		
Model	**P (%)**	**R (%)**	**F1 (%)**	**P (%)**	**R (%)**	**F1 (%)**	**P (%)**	**R (%)**	**F1 (%)**
LSTM	89.05	92.61	90.75	93.10	89.65	91.31	91.08	91.13	91.03
BiLSTM	87.78	93.54	90.53	93.90	81.16	90.90	90.84	90.85	90.71
CNN	88.85	93.37	90.95	93.89	89.20	91.38	91.37	91.29	91.16
XLM-100	91.96	93.41	92.67	93.98	92.59	93.27	92.98	93.00	92.97
M-BERT	91.77	93.73	92.72	94.21	92.34	93.24	93.00	93.03	92.99
XLM-17	92.51	93.14	92.81	93.77	93.17	93.45	93.14	93.15	93.13
BETO	93.31	94.45	93.87	94.94	93.85	94.39	94.12	94.15	94.13

On the one hand, regarding the deep neural networks models, it can be observed that there are no significant differences in the results. CNN performs slightly better than LSTM and BiLSTM. These results were to be expected since these types of networks need a large amount of training data. One of the advantages of these methods is that features are automatically deduced and optimally tuned to the desired outcome.

On the other hand, the Transformer-based models we have explored outperform neural network models including LSTM, CNN and BiLSTM. In particular, the model with the highest performance is the one trained specifically for Spanish (BETO), followed by the XLM-17 model. These results show the capabilities of pre-trained language models that provide the possibility of evaluating a model without the need to rely on large datasets, while at the same time it is an opportunity in some languages such as Spanish where the resources for this type of task are very limited. As previously mentioned, BETO achieves the highest results since it is trained for a specific language, contrary to the other Transformer-based models that have been trained with multilingual corpora such as XLM-100, M-BERT and XLM-17.

Finally, we compare the performance of our systems with the results obtained by López Úbeda et al. [33]. They employed traditional ML algorithms to evaluate the SAD dataset and the classifiers with the best performance were SVM and MLP, obtaining an F1 score of 91.6%. We would like to highlight that although the results obtained by these classifiers are very good, the Transformer-based models we have explored are more effective. Specifically, BETO obtained an improvement performance of 2.76%, which shows that it is necessary to rely on language model specifically trained for a certain language.

6. Error Analysis

The main purpose of this section is to carry out an error analysis to identify the weaknesses of our system. For this purpose, we conducted three different studies: the first one to obtain the number of words included in the Transformer-based models; afterwards, we created the confusion matrix with the best anorexia detection system; finally, we presented some examples of misclassification.

6.1. Words Included in the Vocabularies of the Pre-Trained Models

Table 3 shows how many words are included for each previously used vocabulary model and how many are not in the SAD dataset. The BETO vocabulary has the greatest coverage in the SAD corpus since it includes over 6000 unique tokens. Specifically, we find words from the corpus that are not contained in other model vocabularies such as *gorda* (fat), *delgada* (thin), *luchar* (fight), *desafío* (challenge), *enfermo* (sick), *ánimo* (courage), *músculo* (muscle), *reducción* (reduction), *desayuno* (breakfast) and *nutrición* (nutrition). All these words have important meanings in the context of anorexia, including descriptions of the physical characteristics, foods, moods, and so on.

Table 3. SAD vocabulary included in the Transformer-based models.

	BETO	M-BERT	XLM-100	XLM-17
Number of words included	92,066	106,631	107,055	83,007
Number of unique words included	6086	4049	3949	1367
Number of words not included	51,153	36,588	36,164	60,212
Number of unique words not included	12,545	14,582	14,682	17,264

6.2. Examples of Misclassified Tweets

In ML, and specifically in the classification task, a confusion matrix is a specific table layout that allows visualization of the performance of an algorithm. Figure 5 shows the confusion matrix with two rows and two columns that reports the number of FP, FN, TP, and TN. It was generated with a 10-fold cross-validation using the BETO model. We can observe that for 292 tweets referring to anorexia, the system predicted that 14 were control (FP), and for 278 tweets regarding control, it predicted that 11 referred to anorexia (FN).

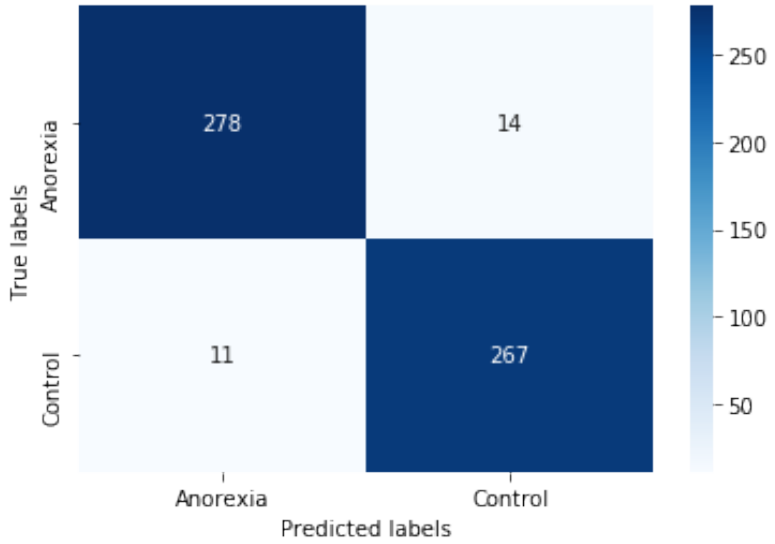

Figure 5. Confusion matrix.

On the other hand, to better understand the tweets mislabeled by BETO we performed a manual inspection on a subset of the data and recorded some of them in Table 4 and the corresponding English translation is in Table 5.

On the one hand, regarding the FN whose IDs are 564, 5297 and 1076, we found that our systems misclassify them because for example, in 564 and 1076 IDs users talk about excesses with food at Christmas time, but the system does not classify them as anorexia. Tweet 5297 contains the word *Anitas* which is the diminutive of Ana commonly used on social media to refer to anorexia, but our model cannot capture the anorexia content and therefore misclassifies it. On the other hand, concerning the FP whose IDs are 362, 478, 634 and 4650, we found that our system misclassifies them because they contain words normally used in the context of anorexia such as *pastilla* (pill), *drogas* (drugs), *objetivo* (goal), *éxito* (success) and *dolor* (pain).

Table 4. Misclassified samples from SAD in Spanish.

ID	Tweet	Annotated	Predicted
564	Aquí yo haciendo mi mejor esfuerzo por sacar de mi cuerpo la cena navideña discretamente	1	0
5297	MOTIVACION para mis queridas Anitas	1	0
1076	¡Y que arranquen los juegos del hambre! (también conocidos como la dieta culpas post festividades)	1	0
362	¡Menos pastilla y más zapatilla!	0	1
478	Esta vida que llevo me está matando. Debería darme a las drogas.	0	1
634	Don't stop. Cada día más cerca del objetivo!!	0	1
4650	Vamos que ya es Miércoles! Sin dolor no hay éxito...	0	1

Table 5. Misclassified samples from SAD in English.

ID	Tweet	Annotated	Predicted
564	Here I am doing my best to get the Christmas dinner out of my body discreetly	1	0
5297	MOTIVATION for my dear Anitas	1	0
1076	And let them start the hunger games! (also known as the post-holiday guilt diet)	1	0
362	Less pill, more shoe!	0	1
478	This life I lead is killing me. I should take drugs.	0	1
634	Don't stop. Every day closer to the goal!!	0	1
4650	Come on, it's already Wednesday! Without pain there is no success...	0	1

7. Discussion

The data collection provided by the eRisk 2019 challenge [16] has served as a starting point for the scientific community interested in anorexia detection using NLP methods. In this challenge, approaches based on neural networks have been previously explored by [46,47] with satisfactory results. Moreover, Mohammadi et al. [48] developed an ensemble approach that employs several attention-based neural sub-models to extract features and predict class probabilities. These features were later used as input features to an SVM that made the final estimation. They achieved the best F1-score (71%). Nivre [49] used a CNN incorporating different pre-trained word embeddings such as GloVe and ELMo. Subsequently, they conducted another study in which they used the CNN-ELMo architecture to improve their results since the word embeddings obtained from ELMo take into account the context of the words and therefore obtained better results (82% F1-score) [50].

Traditional machine learning approaches have also been commonly used in the task of detecting mental illness or more specifically in detecting signs of anorexia [8,20,51]. Specifically, using the SAD corpus, so far the only one available in Spanish and annotated with anorexia, López Úbeda et al. [33] have shown great effectiveness in detecting anorexia using traditional machine learning algorithms such as SVM and MLP reaching both an F1-score of 91.6%. However, these algorithms are not flexible enough to capture more complex relationships naturally and are not suitable for large datasets.

Until now, most of the research presented has been focused on English; however, with this study we have focused on the second most spoken language in the world (with over 470 million native speakers). With respect to the methods proposed, few previous studies

have explored Transformer-based models to detect anorexia or other mental disorders early. These models have been of great interest in the NLP community as they are pre-trained on large corpora. During the pre-training phase, the models learn about the words, structure, morphology, grammar, and other linguistic features of the language that are beneficial for performing language representation and subsequently classify more efficiently. To carry out the experimentation, we have used the SAD corpus and we have compared three neural network architectures (LSTM, BiLSTM and CNN) and three models based on transformers (XLM, M-BERT and BETO) to study the performance of the novel transfer learning methods. The results obtained show that the monolingual pre-trained model BETO outperformed M-BERT and XLM. Thus, one of the major observations in this paper is that this model can more accurately modulate the vocabulary of the corpus because it is trained in Spanish, as opposed to models trained on multilingual corpora.

Finally, we have observed some limitations during the development of the NLP system and its validation. On the one hand, the SAD corpus contains about 5000 annotated tweets, which is a low number of examples to train neural network-based systems and our results show that the data we currently have is not enough to benefit from the potentials of these architectures. On the other hand, all Transformer-based models, although they obtain top performance, have not been trained using a vocabulary as specific as the one used in social networks, which means that they have not been able to learn from the informal language used in social networks. Therefore, it can be seen that there is still room for improvement in the detection of mental disorder comments from social media by applying state-of-the-art methods.

8. Conclusions and Future Work

Recently, due to the large amount of data generated on social media, there is increasing concern about users who promote eating disordered behaviors through the diffusion of messages and images that encourage thinness and harmful weight loss control practices. In this paper, we assess the detection of anorexia in tweets written in Spanish and provide a deeper understanding of the capabilities of new techniques applied to the detection of this kind of disorder.

As far as we know, there is only one dataset labeled with anorexia in Spanish called SAD, which is the one we employ to evaluate our models. The models that we have chosen reflect the state-of-the-art in NLP classification tasks including neural networks such as LSTM, CNN and Transformer-based models based on transfer learning. Moreover, we compare the performance of multilingual and monolingual pre-trained models including M-BERT, BETO and XLM.

The results we have obtained are promising since as opposed to neural networks, Transformer-based models are most effective over small amounts of data. The best performance was achieved by the BETO model with an F1-score of 94.1%, outperforming the best result obtained by López Úbeda et al. [33]. In the error analysis we studied the vocabulary covered in the different pre-trained models and noticed that some words related to anorexia such as *gorda* (fat) or *delgada* (thin) were not found in the multilingual models' vocabulary but were included in the monolingual model BETO, showing the importance of trained a model for a specific language.

As future work, we plan to apply techniques based on sentiment analysis or emotion detection to detect anorexia when there is no implicit information about the mental illness vocabulary. We would also like to test the effectiveness of pre-trained models in detecting other mental disorders such as depression, anxiety, or stress. Regarding language modeling, we plan to use Spanish word embeddings trained on Twitter, both the existing embeddings [52] and self-trained embedding.

Author Contributions: Conceptualization, M.-T.M.-V. and M.C.D.-G.; methodology, M.-T.M.-V. and M.C.D.-G.; software, P.L.-Ú. and F.M.P.-A.; validation, M.-T.M.-V. and M.C.D.-G.; formal analysis, P.L.-Ú., F.M.P.-d.-A. and M.C.D.-G.; investigation, F.M.P.-d.-A. and M.C.D.-G.; resources, P.L.-Ú. and F.M.P.d.-A.; data curation, P.L.-Ú. and F.M.P.-d.-A.; writing—original draft preparation, P.L.-Ú., F.M.P.-d.-A. and M.C.D.-G.; writing—review and editing, M.-T.M.-V. and M.C.D.-G.; visualization, P.L.-Ú. and F.M.P.-d.-A.; supervision, M.-T.M.-V. and M.C.D.-G.; project administration, M.-T.M.-V. and M.C.D.-G.; funding acquisition, M.-T.M.-V. and M.C.D.-G. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been partially supported by a grant from European Regional Development Fund (ERDF), LIVING-LANG project [RTI2018-094653-B-C21], and the Ministry of Science, Innovation, and Universities (scholarship [FPI-PRE2019-089310]) from the Spanish Government.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this study is available at: https://github.com/plubeda/SAD.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Vigo, D.; Thornicroft, G.; Atun, R. Estimating the true global burden of mental illness. *Lancet Psychiatry* **2016**, *3*, 171–178. [CrossRef]
2. James, S.L.; Abate, D.; Abate, K.H.; Abay, S.M.; Abbafati, C.; Abbasi, N.; Abbastabar, H.; Abd-Allah, F.; Abdela, J.; Abdelalim, A.; et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: A systematic analysis for the Global Burden of Disease Study 2017. *Lancet* **2018**, *392*, 1789–1858. [CrossRef]
3. Sidani, J.E.; Shensa, A.; Hoffman, B.; Hanmer, J.; Primack, B.A. The association between social media use and eating concerns among US young adults. *J. Acad. Nutr. Diet.* **2016**, *116*, 1465–1472. [CrossRef]
4. Calvo, R.A.; Milne, D.N.; Hussain, M.S.; Christensen, H. Natural language processing in mental health applications using non-clinical texts. *Nat. Lang. Eng.* **2017**, *23*, 649–685. [CrossRef]
5. Cavazos-Rehg, P.A.; Krauss, M.J.; Costello, S.J.; Kaiser, N.; Cahn, E.S.; Fitzsimmons-Craft, E.E.; Wilfley, D.E. "I just want to be skinny": A content analysis of tweets expressing eating disorder symptoms. *PLoS ONE* **2019**, *14*, e0207506. [CrossRef] [PubMed]
6. Dredze, M. How social media will change public health. *IEEE Intell. Syst.* **2012**, *27*, 81–84. [CrossRef]
7. Srivastava, S.; Pant, M.; Nagar, A. Yuva: An e-health model for dealing with psychological issues of adolescents. *J. Comput. Sci.* **2017**, *21*, 150–163. [CrossRef]
8. Eichstaedt, J.C.; Smith, R.J.; Merchant, R.M.; Ungar, L.H.; Crutchley, P.; Preoţiuc-Pietro, D.; Asch, D.A.; Schwartz, H.A. Facebook language predicts depression in medical records. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 11203–11208. [CrossRef]
9. Aladağ, A.E.; Muderrisoglu, S.; Akbas, N.B.; Zahmacioglu, O.; Bingol, H.O. Detecting suicidal ideation on forums: Proof-of-concept study. *J. Med. Internet Res.* **2018**, *20*, e215. [CrossRef]
10. Coppersmith, G.; Leary, R.; Crutchley, P.; Fine, A. Natural language processing of social media as screening for suicide risk. *Biomed. Inform. Insights* **2018**, *10*, 1178222618792860. [CrossRef] [PubMed]
11. Birnbaum, M.L.; Ernala, S.K.; Rizvi, A.F.; De Choudhury, M.; Kane, J.M. A collaborative approach to identifying social media markers of schizophrenia by employing machine learning and clinical appraisals. *J. Med. Internet Res.* **2017**, *19*, e289. [CrossRef] [PubMed]
12. Ernala, S.K.; Labetoulle, T.; Bane, F.; Birnbaum, M.L.; Rizvi, A.F.; Kane, J.M.; De Choudhury, M. Characterizing audience engagement and assessing its impact on social media disclosures of mental illnesses. In Proceedings of the Twelfth International AAAI Conference on Web and Social Media, Stanford, USA, 25–28 June 2018.
13. Loveys, K.; Crutchley, P.; Wyatt, E.; Coppersmith, G. Small but mighty: Affective micropatterns for quantifying mental health from social media language. In Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality, Vancouver, Canada, 3 August 2017; pp. 85–95.
14. Shrivastava, A.; Tripathy, A.K.; Dalal, P.K. A SVM-based classification approach for obsessive compulsive disorder by oxidative stress biomarkers. *J. Comput. Sci.* **2019**, *36*, 101023. [CrossRef]
15. Losada, D.E.; Crestani, F.; Parapar, J. eRISK 2017: CLEF lab on early risk prediction on the internet: Experimental foundations. In *Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages, Dublin, Ireland, 11–14 September 2017*; Springer: Berlin, Germany, 2017; pp. 346–360.
16. Losada, D.E.; Crestani, F.; Parapar, J. Overview of eRisk 2019 Early Risk Prediction on the Internet. In *International Conference of the Cross-Language Evaluation Forum for European Languages*; Springer: Lugano, Switzerland, 2019; pp. 340–357.

17. Loveys, K.; Niederhoffer, K.; Prud'hommeaux, E.; Resnik, R.; Resnik, P. In Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, New Orleans, LO, USA, 5 June 2018.
18. Klump, K.L.; Bulik, C.M.; Kaye, W.H.; Treasure, J.; Tyson, E. Academy for eating disorders position paper: Eating disorders are serious mental illnesses. *Int. J. Eat. Disord.* **2009**, *42*, 97–103. [CrossRef]
19. Wolf, M.; Theis, F.; Kordy, H. Language use in eating disorder blogs: Psychological implications of social online activity. *J. Lang. Soc. Psychol.* **2013**, *32*, 212–226. [CrossRef]
20. Conway, M.; Hu, M.; Chapman, W.W. Recent Advances in Using Natural Language Processing to Address Public Health Research Questions Using Social Media and Consumer Generated Data. *Yearb. Med. Inform.* **2019**, *28*, 208–217.
21. Yan, H.; Fitzsimmons-Craft, E.E.; Goodman, M.; Krauss, M.; Das, S.; Cavazos-Rehg, P. Automatic detection of eating disorder-related social media posts that could benefit from a mental health intervention. *Int. J. Eat. Disord.* **2019**, *52*, 1150–1156. [CrossRef]
22. Moessner, M.; Feldhege, J.; Wolf, M.; Bauer, S. Analyzing big data in social media: Text and network analyses of an eating disorder forum. *Int. J. Eat. Disord.* **2018**, *51*, 656–667. [CrossRef] [PubMed]
23. Sharma, E.; De Choudhury, M. Mental health support and its relationship to linguistic accommodation in online communities. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal, Canada, 21–26 April 2018; pp. 1–13.
24. Tausczik, Y.R.; Pennebaker, J.W. The psychological meaning of words: LIWC and computerized text analysis methods. *J. Lang. Soc. Psychol.* **2010**, *29*, 24–54. [CrossRef]
25. Kebria, P.M.; Khosravi, A.; Salaken, S.M.; Nahavandi, S. Deep imitation learning for autonomous vehicles based on convolutional neural networks. *IEEE/CAA J. Autom. Sin.* **2019**, *7*, 82–95. [CrossRef]
26. Chen, L.; Hu, X.; Tian, W.; Wang, H.; Cao, D.; Wang, F.Y. Parallel planning: A new motion planning framework for autonomous driving. *IEEE/CAA J. Autom. Sin.* **2018**, *6*, 236–246. [CrossRef]
27. Wang, S.; Cai, J.; Lin, Q.; Guo, W. An overview of unsupervised deep feature representation for text categorization. *IEEE Trans. Comput. Soc. Syst.* **2019**, *6*, 504–517. [CrossRef]
28. Ive, J.; Gkotsis, G.; Dutta, R.; Stewart, R.; Velupillai, S. Hierarchical neural model with attention mechanisms for the classification of social media text related to mental health. In Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, New Orleans, LO, USA, 5 June 2018; pp. 69–77.
29. Gkotsis, G.; Oellrich, A.; Velupillai, S.; Liakata, M.; Hubbard, T.J.; Dobson, R.J.; Dutta, R. Characterisation of mental health conditions in social media using Informed Deep Learning. *Sci. Rep.* **2017**, *7*, 45141. [CrossRef]
30. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, LO, USA, 1–6 June 2018; pp. 2227–2237 [CrossRef]
31. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
32. Rodrigues Makiuchi, M.; Warnita, T.; Uto, K.; Shinoda, K. Multimodal Fusion of BERT-CNN and Gated CNN Representations for Depression Detection. In Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop, Nice, France, 21–25 October 2019; pp. 55–63.
33. López Úbeda, P.; Plaza del Arco, F.M.; Díaz Galiano, M.C.; Urena Lopez, L.A.; Martin, M. Detecting Anorexia in Spanish Tweets. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), Varna, Bulgaria, 2–4 September 2019; pp. 655–663. [CrossRef]
34. Hochreiter, S.; Schmidhuber, J. LSTM can solve hard long time lag problems. In *Advances in Neural Information Processing Systems*; The MIT Press: Denver, CO, USA, 1996; pp. 473–479.
35. Schuster, M.; Paliwal, K.K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681. [CrossRef]
36. Kalchbrenner, N.; Grefenstette, E.; Blunsom, P. A Convolutional Neural Network for Modelling Sentences. *arXiv* **2014**, arXiv:1404.2188.
37. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching Word Vectors with Subword Information. *arXiv* **2016**, arXiv:1607.04606.
38. Wang, G.; Qiao, J.; Bi, J.; Li, W.; Zhou, M. TL-GDBN: Growing deep belief network with transfer learning. *IEEE Trans. Autom. Sci. Eng.* **2018**, *16*, 874–885. [CrossRef]
39. Yang, Q.; Zhang, Y.; Dai, W.; Pan, S.J. *Transfer Learning*; Cambridge University Press: Cambridge, MA, USA, 2020. [CrossRef]
40. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 5998–6008.
41. Lample, G.; Conneau, A. Cross-lingual Language Model Pretraining. *arXiv* **2019**, arXiv:1901.07291.
42. Cui, Y.; Che, W.; Liu, T.; Qin, B.; Yang, Z.; Wang, S.; Hu, G. Pre-Training with Whole Word Masking for Chinese BERT. *arXiv* **2019**, arXiv:1906.08101.
43. Pires, T.; Schlinger, E.; Garrette, D. How multilingual is Multilingual BERT? *arXiv* **2019**, arXiv:1906.01502.
44. Sennrich, R.; Haddow, B.; Birch, A. Neural Machine Translation of Rare Words with Subword Units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, 7–12 August 2016; pp. 1715–1725. [CrossRef]

45. Wolf, T.; Chaumond, J.; Debut, L.; Sanh, V.; Delangue, C.; Moi, A.; Cistac, P.; Funtowicz, M.; Davison, J.; Shleifer, S.; et al. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Vienna, Austria, 16–20 November 2020; pp. 38–45.
46. Ragheb, W.; Azé, J.; Bringay, S.; Servajean, M. Attentive Multi-Stage Learning for Early Risk Detection of Signs of Anorexia and Self-Harm on Social Media. In *CLEF (Working Notes)*; CCSD: Lugano, Switzerland, 2019.
47. Masood, R.; Ramiandrisoa, F.; Aker, A. UDE at eRisk 2019: Early Risk Prediction on the Internet. In *Conference and Labs of the Evaluation Forum, Living Labs (CLEF 2019)*; CCSD: Lugano, Switzerland, 2019; pp. 1–9.
48. Mohammadi, E.; Amini, H.; Kosseim, L. Quick and (Maybe Not So) Easy Detection of Anorexia in Social Media Posts. In *CLEF (Working Notes)*; Concordia University: Montreal, QC, Canada, 2019.
49. Nivre, J. *Uppsala University and Gavagai at CLEF eRISK: Comparing Word Embedding Models*; Springer: Cham, Switzerland, 2019.
50. Amini, H.; Kosseim, L. Towards Explainability in Using Deep Learning for the Detection of Anorexia in Social Media. In *International Conference on Applications of Natural Language to Information Systems*; Springer: Cham, Switzerland, 2020; pp. 225–235.
51. Plaza-del Arco, F.M.; López-Úbeda, P.; Dıaz-Galiano, M.C.; Urena-López, L.A.; Martın-Valdivia, M.T. *Integrating UMLS for Early Detection of Sings of Anorexia*; Universidad de Jaen, Campus Las Lagunillas: Jaen, Spain, 2019.
52. Cieliebak, M.; Deriu, J.M.; Egger, D.; Uzdilli, F. A twitter corpus and benchmark resources for german sentiment analysis. In Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, Valencia, Spain, 3 April 2017; pp. 45–51.