Online Social Networks and Media 31 (2022) 100225

Contents lists available at ScienceDirect

Online Social Networks and Media

journal homepage: www.elsevier.com/locate/osnem

Early depression detection in social media based on deep learning and underlying emotions

José Solenir L. Figuerêdo *, Ana Lúcia L.M. Maia, Rodrigo Tripodi Calumby

University of Feira de Santana, Av. Transnordestina, s/n, Novo Horizonte, Feira de Santana, 44036-900, Bahia, Brazil

ARTICLE INFO

Keywords:
Depression
Social media
Deep convolutional neural networks
Ensemble

ABSTRACT

Depression is a challenge to public health, frequently related to disability and one of the reasons that lead to suicide. Many of the ones who suffer depression use social media to obtain information or even to talk about their problem. Some studies have proposed to detect potentially depressive users in these online environments. However, unsatisfactory effectiveness is still a barrier to practical application. Hence, we propose a method of early detection of depression in social media based on a convolutional neural network in combination with context-independent word embeddings and Early and Late Fusion approaches. These approaches are experimentally evaluated, considering the importance of the underlying emotions encoded in the emoticons. The results show that the proposed method was able to detect potentially depressive users, reaching a precision of 0.76 with equivalent or superior effectiveness in relation to many baselines (F_1 (0.71)). In addition, the semantic mapping of emoticons allowed to obtain significantly better results, including higher recall and precision with gains of 46.3% and 32.1%, respectively. Regarding the baseline word embedding approach, the higher recall and precision gains of our semantic mapping of emoticons were 14.5% and 40.8%. In terms of overall effectiveness, this work advanced the state-of-the-art, considering both individual embeddings and the fusion-based approaches. Moreover, it is demonstrated that emotions expressed by depressed people and encoded through emoticons are important suggestive evidence of the problem and a valuable asset for early detection.

1. Introduction

Depression is a psychological disorder related to a combination of genetic, biological, environmental and psychological factors [1,2]. People who suffer from this disorder tend to have a set of symptoms, e.g., loss of energy, changes in appetite, anxiety, reduced concentration, indecision, feelings of worthlessness, guilt or hopelessness. Despite the advances in prevention, diagnosis and treatment, the number of people with depression has continuously grown [2,3]. In fact, depression is the leading cause of health problems and disability worldwide. According to recent estimates by the World Health Organization (WHO), more than 300 million people suffer from it worldwide [2]. It also showed an increase of 18.4% in the number of cases between 2005 and 2015. Depression is also the major contributor to suicide deaths, which number close to 800,000 per year [2]. It highlights the severity of this disease and the need to devise strategies to improve the diagnosis and suppress its progression.

Although there are many methods to treat depression, less than half of those affected in the world receive proper treatment [3]. Numerous factors lead to non-treatment, including the absence of diagnosis and inaccurate or even incorrect assessments. From this scenario, the need for improvements in the process of detecting depression emerges as a relevant challenge. Hence, a promising complementary alternative is the exploitation of data generated in social media.

The emergence of social media platforms such as Facebook, Twitter, and Reddit allowed people to share their personal experiences, ideas, or thoughts directly and comprehensively. In addition to the explicit meaning carried in those messages posted on social media, they also embody implicit information about their authors. Hence, as indicated in [4], social media also makes it possible to find signs that characterize the emergence of depression in individuals.

In this context, social media is a resource that can be explored for detecting people with depression, especially because people who suffer from this tend to hide their health status, which makes it difficult for a specialist to diagnose, but may be latent in the content published in social media.

Thus, exploring the use of language on social media becomes a promising direction. Specifically, with early detection on social media,

* Corresponding author.
E-mail addresses: jslfigueredo@ecomp.uefs.br (J.S.L. Figuerêdo), anamarreiros@gmail.com (A.L.L.M. Maia), rtcalumby@uefs.br (R.T. Calumby).

https://doi.org/10.1016/j.osnem.2022.100225
Received 25 August 2021; Received in revised form 17 June 2022; Accepted 4 July 2022
Available online 21 July 2022

preventive measures can be applied more efficiently. Given the large-scale data analysis task, one can use both direct assessment approaches, carried out by an expert, and/or automated approaches such as using machine learning [5].

Machine learning has achieved impressive results in difficult pattern-oriented problems [6], such as speech recognition, object recognition in images and videos, and natural language processing (NLP). However, regarding the detection of depression, using traditional classification methods may not succeed, due to the difficulty in extracting discriminative features from texts on social media [7]. A recent alternative is the use of advanced techniques such as Deep Learning (DL). DL has achieved surprising results in many applications (e.g., health diagnosis [8], image synthesis [9], and autonomous car driving [10]). In addition, although still incipient, DL has also shown promising results in the early detection of depression [11,12].

Early detection is essential for depression treatment, as it allows preventive measures to be taken to attenuate or mitigate problems, especially considering lives are at risk. Despite this, many proposals consider only a reactive solution, ignoring the temporal issue. However, the time factor must be considered, since detecting it when the subject explicitly demonstrates the problem may be too late, and the treatment may not come in time or the quality of life may have already been seriously affected.

Over the years, different approaches have been developed to tackle the task of depression detection in social media, focusing mainly on the extraction of textual representations [4,13], mostly relying on deep learning based text embedding [12,14]. However, the proposed approaches still do not achieve satisfactory effectiveness based on what is expected to allow deployment in real settings. A sounding alternative to improve these approaches is the use of data fusion strategies, aiming at capturing the best of each system, relying on different views (models) and their complementarity over the data. Nevertheless, for depression detection, few studies explored it and almost none assessed the combination of representation models.

Therefore, this work proposes a method of early depression detection in social media using DL and Early and Late Fusion approaches. It seeks to improve the process of identifying potentially depressed users and is experimentally evaluated using different word embeddings as feature representations. In Summary, the main contributions of this work are:

- We propose a new method that explores the use of early and late fusion to help the early detection of depression in social media users.
- We conduct extensive experiments and demonstrate the proposed method outperforms the baselines in various scenarios.
- We demonstrate the decisive role of emoticons[1] on detection performance as they are a proxy for users' feelings and emotions.

The remainder of this paper is organized as follows. Section 2 describes relevant background concepts. Section 3 presents the related works and Section 4 describes the proposed method. In turn, Section 5 presents the experimental settings. The results and discussions are presented in Section 6. Section 7 addresses challenges and research directions. Finally, Section 8 brings the conclusions and future work.

2. Background

2.1. Deep Learning

DL is a subfield of Machine learning (ML) that achieves great power and flexibility by learning to represent knowledge as a hierarchy of concepts, with each concept defined in relation to simpler concepts and more abstract representations computed in terms of less abstract ones [15]. Unlike traditional learning approaches where a thorough selection of features is performed by a human being, DL embodies a general-purpose learning procedure. Therefore, DL allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction [16].

Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are examples of DL networks. CNNs have brought breakthroughs in image, video, speech and audio processing [16,17], whereas RNNs allowed significant advances in sequential data processing such as text and speech [18]. More recently, some architectures using the so-called transformer approach outperformed RNNs in several tasks [19]. Nevertheless, adaptations to support text processing using CNN-based architectures have also achieved promising effectiveness [20–22] and are the focus of this work.

CNNs are a specialized type of neural network for data processing with a grid topology [15]. Unlike other types of networks, CNNs include at least one layer of convolution filters, which are oriented to local features extraction [23]. CNNs are widely applied in computer vision, but have also been effectively used for text classification [24].

2.2. Word embeddings

Word embeddings became a popular way of textual representation, and have been applied in many tasks, with emphasis on natural language recognition, document classification and sentiment analysis [25]. The embedding methods provide low dimensional vector representations for unstructured text [26]. It has a great generalization power and, following the distributive hypothesis, captures that similar words tend to occur in similar contexts [27]. For example, vectors for "green" is closer to "blue" than "shoe", since the first two words refer to colors. Similarly, "Rome" is as close to "Italy" as "Paris" is to "France", due to their contextual similarity. This should also occur between synonyms (e.g., "dog" and "puppy", "huge" and "enormous"). In summary, word embeddings represent a set of techniques in which individual words are represented as real value vectors in a predefined vector space, usually with tens or hundreds of dimensions [28,29]. In this work, word embeddings were used for feature representation from the users posts on social media.

2.3. Data fusion

Data fusion is used in different contexts, being frequently applied in information classification and retrieval tasks. When properly exploited, fusion has been shown to decisively improve the effectiveness of systems [30]. There are basically three approaches for data fusion: Early Fusion, Late Fusion, and hybrid [31–33]. The Early Fusion process usually corresponds to the aggregation and/or concatenation of a set of features, possibly extracted based on different strategies, before the training or ranking stage. On the other hand, the Late Fusion process performs the "aggregation" of decisions made by a set of trained classifiers. In general, when using Late Fusion, its effectiveness is expected to be superior to the best individual classifier [34]. In addition to the Late and early fusion approaches, there are also intermediate hybrid fusion methods. Among the most widespread methods are the Gaussian Processes and Autoencoders (classical, deep or variational). In this work, we exploit the early and late fusion methods, whereas hybrid ones are left to future specific studies given their usually higher complexity.

[1] A digital icon or a sequence of keyboard made up of symbols such as punctuation marks, used in text messages, emails, etc. to express a particular emotion.

2.4. Language and depression detection

Depression affects many people, but unfortunately, most of them are unaware of their disease and, therefore, do not seek clinical intervention until symptoms become severe. Therefore, investigating strategies to support this process is of great importance, especially because it directly impacts the overall health condition of the patient. These impacts may be reduced, or even avoided with appropriate interventions, and early detection is a decisive first step. To contribute to this task, the analysis of the language used by the individual became an asset, since there is a relationship between depression and the use of language [35,36].

Several studies based on the language used on social media show that people who suffer from depression tend in general to: (*i*) Talk more about relationships and life; (*ii*) Show personality; (*iii*) Become more concerned with themselves; (*iv*) Use more emoticons, negative emotion, and denial words; (*v*) Use more verbs, adverbs, exclamation and question marks; (*vi*) Frequently use words with strong semantics (e.g., swear); (*vii*) Constantly remembering the past and worrying about the future [37].

Language is a powerful indicator of personality, social or emotional status, as well as of mental health. Several studies indicate that it is possible to predict the person's mental state by examining the use of language, including depression evidences [4,37,38]. Therefore, social media offer a great opportunity to proactively detect these users and refer them as soon as possible to professional help. In fact, considering the large amount and richness of the information available, some works have already relied on social media data to conduct this kind of investigation and are discussed in Section 3.

3. Related works

In recent years, many researchers have worked on detecting mental illnesses on social media, including depression [39,40]. The social platforms created a rich source of text data and social metadata to capture users' behavioral trends and are considered a promising tool for public health [41]. Among these platforms, Twitter is one of the most explored and many other providers have been explored such as Sina Weibo[2] [42,43], Facebook [44,45] and, more recently, Reddit [11,37]. Considering the underlying detection methods, NLP techniques and various classification approaches have been applied to analyze textual data and assess users' mental health through these social media [46].

In this context, many studies were conducted in order to improve depression detection, most relying on supervised learning. Moreover, some works have proposed feature extraction methods to be used with classic machine learning classifiers [4,11,47,48]. In [4], the authors proposed a probabilistic classifier to identify whether an individual is vulnerable to depression based on the analysis of the posts made by this user on Twitter. For this, many features were extracted from posts to represent social engagement, emotions, language and linguistic styles, and references to antidepressant drugs. Using Support Vector Machines (SVM), 70% accuracy was achieved.

Tsugawa et al. [47] investigated user activities on Twitter to estimate the degree of depression. For that, several features were extracted from the users' activity histories (e.g., frequency of words used, rate of the retweet, rate of mentions, proportion of tweets containing URLs, number of users followed and number of followers). Using a SVM classifier, 69% accuracy was achieved. The authors also found that the features extracted by a topic model were useful for this type of investigation.

Twitter has been one of the most used networks in depression detection experiments. However, the work published by Losada et al. [39] using Reddit opened up new development opportunities. The authors argue that the limited number of characters in a tweet reduces the context that can be explored about the writer. In Reddit, there are no limitations regarding the number of characters. The data collected and described in [39] included two groups: a control group with non-depressive users and another one with users identified as depressive. Moreover, for effectiveness evaluation purposes, considering commonly used classic measures (Precision, Recall, and F_1) disregard recognition delay, a measure for early detection was proposed to penalize the delay in detecting depressive users. This data collection was published as part of the *Pilot Task: Early Detection of Depression* - eRisk 2017.[3] In the proposed measure (described in Section 5.3), time (detection delay) is represented by the number of posts that were necessary to be inspected by the system to be confident for a depression detection decision.

Errecalde et al. [48] applied a recent strategy called Concise Semantic Analysis (CSA) in order to deal with early detection of depression on social media. In fact, they developed a CSA variation using Temporal Variation of Terms (TVT), which is based on the use of vocabulary variation throughout different time steps as a conceptual space to represent posts. The authors found that using the TVT approach combined with other representations, such as Bag of Words (BoW), achieved robust effectiveness on early risk detection, ranking among the top performing methods submitted to eRisk 2017.

The top-2 performing methods at eRisk Task 2017 were developed by Trotzek et al. [13]. All classifiers received as features meta linguistic information extracted from each user's texts. In addition, other feature representation methods were assessed such as BoW, paragraph vector, and Latent Semantic Analysis (LSA). For classification, logistic regression and LSTMs were used. The model that obtained the best result used an ensemble of logistic regression classifiers based on BoW with different weightings of terms and *n-grams*, identified as FHDO-BCSGA. The second best model, named FHDO-BCSGB, also used a logistic regression and relied on the vectorization (doc2vec)[4] of documents using paragraph vectors.

The eRisk 2017 data have been explored in some subsequent works. In [50], the authors investigated how to better detect the early risk of depression in social media, aiming at optimizing the classification, without neglecting the temporal dimension. Several algorithms, such as SVM, Random Forest, kNN and logistic regression were used, including their combination through an ensemble, as well as the use of genetic algorithms to optimize the ensemble. The results showed that the application of genetic algorithms and the polarity of the text improved the detection by 16.7% in relation to the baseline. With a deep learning approach, in [11], a CNN based on different word embeddings was evaluated and compared with a classification based on linguistic metadata at the user level using a logistic regression. An ensemble of both approaches reached the state-of-the-art recognition effectiveness in early depression detection using the eRisk 2017 collection.

Although some works have already been developed in the field of early detection of depression, the actual effectiveness is still unsatisfactory in relation to what is expected for a real world application. Hence, this study proposed the assessment of a set of DL-based feature embedding models and classification approaches based on Early and Late fusion methods. The use of these strategies is justified for two main reasons: (*i*) By jointly using different representations, a CNN could cross-capture valuable feature relationships, which would go unnoticed without the fusion of features; and (*ii*) By using a set of CNN models built with different feature representations, complementary points-of-view over the same data are integrated usually allowing outperforming individual predictions.

[2] https://www.weibo.com - As of July 11, 2021.

[3] http://erisk.irlab.org/2017/index.html - As of August 24, 2021 from CLEF 2017 [49].

[4] https://radimrehurek.com/gensim/models/doc2vec.html - As of April 10, 2022.

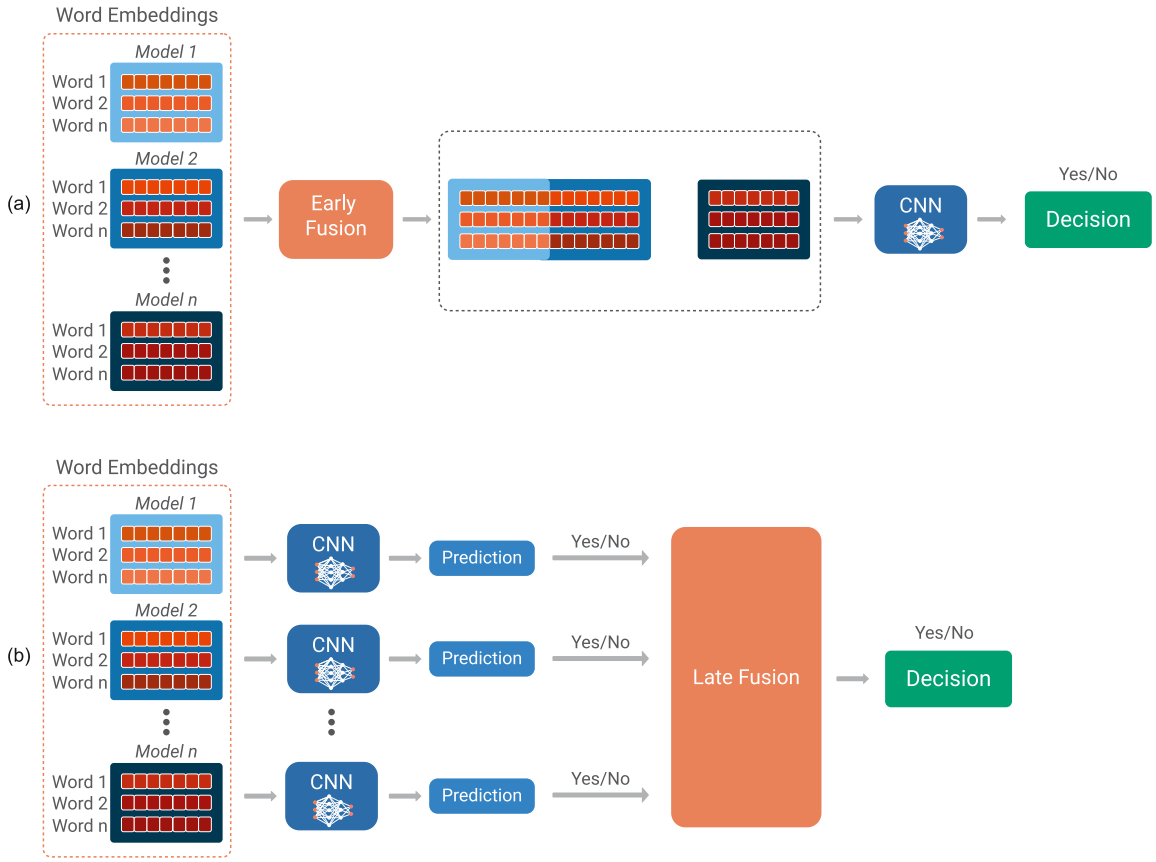

Fig. 1. Fusion Methods: (a) Early Fusion; (b) Late Fusion.

4. Proposed method

The main objective of this work is to advance the state-of-the-art of early detection of depression in social media. For that, early and late fusion methods are proposed using different word embedding models for feature representation of the user's posts (Fig. 1). Additionally, the CNN architecture proposed in [11] was considered as the classification model.

Fig. 2 presents the CNN architecture, which was integrated into our approach and used in all experiments described in Sections 5 and 6. The architecture consists of a simple convolution layer, with 100 filters of height 2. The width of the filter corresponds to the dimensionality of the input embedding vector. This convolution step generates a 99×1 feature map per filter since no padding is applied and the stride is equal to 1. In the convolution and dense layers, Concatenated Rectified Linear Units (CReLU) are used as activation function [51]. CReLU concatenates the output of a Rectified Linear Unit (ReLU) that considers only the positive part of the activation with the output of another ReLU that considers only the negative part of the activation. As a result, there are twice as many outputs.

To obtain a scalar for each filter, the 1-max polling method is used, resulting in a 100-dimensional vector. Given CReLU is used, this vector becomes 200-dimensional. The output is then propagated through three fully connected (FC) layers. In turn, to reduce overfitting, the dropout regularization technique is applied to the output of the first dense layer. Finally, a softmax layer generates the final output.

In the early fusion process, multiple feature embeddings are concatenated as an integrated representation that is used as the input for the CNN. This process gets a larger set of features that for allowing a better representation (Fig. 1a). For this approach, the input layer of the CNN was modified to support the larger input, i.e., instead of using a 100×300 input layer, a 100×600 layer was used to support the fused representations.

In turn, in the late fusion approach, an ensemble is constructed over multiple classifiers previously trained with independent embeddings. In the late fusion approach, the final decision corresponds to the combination of the results after the individual classification from each model. We apply a majority voting scheme, in which the final decision is made considering the class with the highest number of votes (Fig. 1b).

The proposed method also considers additional aspects related to the importance of emotions encoded by emoticons in the posts. Specifically, we consider two opposite strategies: (*i*) In the first strategy, all emoticons in the posts were discarded; (*ii*) In the second one, the emotions were preserved by mapping the emoticons to representative terms, similarly. For instance, the ":(" symbol was replaced by the term "sad". This approach was assessed to analyze whether emotions, represented by emoticons, impacted the identification of users with depression, as suggested in the literature [37,52]. This procedure relied on a predefined mapping dictionary.[5]

In this work, word embeddings are used as input to the CNN. Unlike some studies [53], in which the embedding models are usually trained from scratch, we rely on a transfer learning strategy with feature extraction models pre-trained with larger datasets, not necessarily related to depression. This transfer learning approach has been a popular solution for dealing with small datasets. The choice of pre-trained models was due to the unavailability of large amounts of labeled data related to depression.

[5] https://en.wikipedia.org/wiki/List_of_emoticons - As of July 11, 2021.

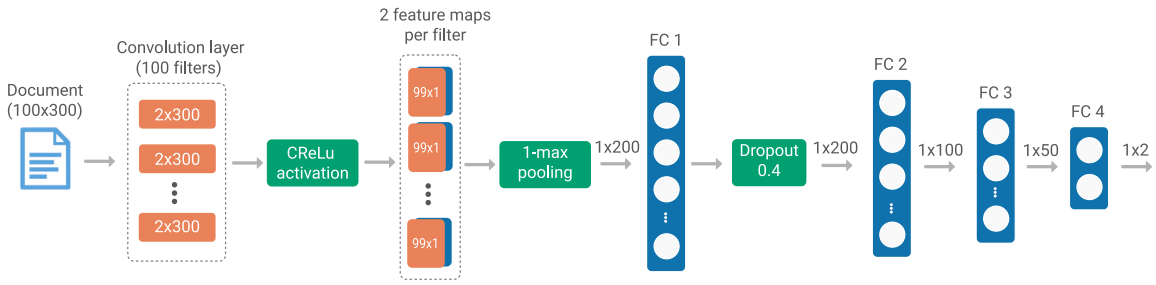

Fig. 2. CNN architecture used in the experiments [11].

Table 1
Main collection and user statistics [49].

	Train		Test	
	Positive	*Control*	*Positive*	*Control*
Number of users	83	403	52	349
Number of posts	30,851	264,172	18,706	217,665
Avg. posts per user	371.7	655.5	359.7	623.7
Avg. days (1st to last post)	572.7	626.6	608.3	623.2
Avg. words per post	27.6	21.3	26.9	22.5

Table 2
Characteristics of word embeddings used in the experiments.

Model	Dimension	Dataset tokens (in billions)	Word vectors (in millions)
FastText Crawl	300	600	2
FastText WN	300	16	1
GloVe Crawl	300	42	2
Glove WN	300	6	0.4

5. Experimental validation

5.1. Dataset

The experiments relied on a data set published as part of the eRisk 2017 [49]. It encompasses a collection of messages in English from Reddit users. The dataset includes the list of posts from users, up to 2000 posts overall, with the following fields: title, date, and text. The title and text fields were concatenated and used in the experiments. The collected messages are organized in chronological order, including data from 887 users (135 labeled as depressive and 752 labeled as non-depressive). As reported in [39], for the construction of the dataset, depressive users were identified by finding posts that clearly mention a diagnosis (e.g., "I was diagnosed with depression"). The dataset was randomly divided into training and test sets. The training set consists of 486 users (83 positive and 403 negative). The test set contains 401 users (52 positive and 349 negative). There is no overlap between training and test users. A summary of the characteristics of the dataset and users is shown in Table 1.

5.2. Configuration

The rationale is to simulate systems that monitor social media and analyze the posts made by users. To simulate this, the data were divided into ten blocks, each containing 10% of each user's messages in chronological order. The first block contains the 10% of the oldest messages, the second block contains the next 10% of the oldest messages, and so on. In this fashion, by joining the first and second blocks, for instance, one gets the 20% oldest messages.

The task is carried out in two phases: the training phase and the testing phase. The testing data are divided into ten blocks. In that one, each block should be processed individually, given a particular characteristic of early risk detection task: support for classification with partial information available at different moments in time. The starting blocks contained the oldest posts, while the ending blocks had the most recent ones. Given the experimental nature, when processing a block the system could take action to three possibilities: classify a user as depressive, non-depressive or postpone the decision, indicating that it is necessary to analyze more data blocks (more posts/time).

Considering the early detection models, a parameter of great relevance to be considered is the minimum confidence threshold (τ) for the prediction. In each block, the classifiers respond with a certain level of confidence, consisting of the estimated probability for the predicted class. The threshold determines whether the model considers the confidence as sufficient to identify the subject as depressive, or else should delay the decision and wait for more data. Hence, the model can be evaluated based on different probability thresholds. Regarding the classification itself, this study used a simple heuristic rule, that is to assign a user to the target class when its associated probability ρ is greater (or equal) than the threshold ($\rho \geq \tau$). Multiple confidence thresholds were evaluated in the interval [0.5,1] with a 0.05 step size.

The training is performed only once with the entire training data set. On the other hand, the 10 test set blocks were processed incrementally. This means that, at each step, for each user, all the posts up to that point, are aggregated and analyzed to support a decision. The decision that a user has no depression is made only after including the last block. In fact, it means the systems did not achieve enough confidence to classify the user as depressive and no more data is available for further analysis. The evaluation process relied on the ground-truth provided with the database. For feature extraction and input to the CNN, many pre-trained embedding models were used, specifically based on fastText[6,7] [54] and GloVe[8] [55]. The fastText 300-dimensional embeddings were trained with data from UMBC webbase corpus (W) and statmt.org news dataset (N) (FastText WN) and also with a Common Crawl (FastText Crawl). The GloVe model was trained with data from Gigaword 5 and Wikipedia 2014 and named, respectively, "GloVe WN" and "Glove Crawl". Further details of these models are presented in Table 2.

Table 2 summarizes the characteristics of the embedding models used in this study. It contains three basic information: the number of dimensions of the vector generated as the embedding for a word; the number of tokens in the database with which the embedding model was trained; and the number of word vectors in each model. Each token represents a symbol (e.g., a word). For instance, considering FastText Crawl model, it can generate an embedding vector for 2 million different words.

In the classification phase, the convolutional neural network receives each user post as input, considering the first 100 word vectors of that post (i.e, for each post we select the first 100 words and used its vector representation given by the embedding as input to the CNN). The first 100 words were used to maintain consistency with the main

[6] https://fasttext.cc/docs/en/english-vectors.html - As of October, 2020.
[7] https://fasttext.cc/docs/en/pretrained-vectors.html- As of October, 2020.
[8] https://nlp.stanford.edu/projects/glove/ - As of October, 2020.

baseline, but a reduced number of words could be used, since the average of words per post is 34.58. Before this step, zeros are filled for posts with less than 100 words. Each word has a vector representation attributed by the embedding model, resulting in a 100 × 300 matrix as the input for the neural network. The CNN performs the classification for each post individually, for each user. Thus, since a user can have up to 2000 posts overall, it is necessary to aggregate these results and make a single decision for that specific user. For this, similarly to that performed by Trotzek et al. [11], the 98th percentile of the probabilities of this user being depressed is calculated. This value is then considered for the final decision for that user. The use of the percentile instead of an average probability intends to give more weight (attention) to posts with a greater probability.

The experiments were carried out using the single embeddings, as well as the proposed early and late fusion approaches. The experiments were conducted in the Colab environment.[9] The preprocessing step was performed using the NLTK[10] library and the Keras preprocessing module.[11] The model construction process used the Keras API and Tensorflow framework. For result compatibility purposes, the validation and training process followed the same approach described in [11], our main baseline (see Section 5.3). Regarding the CNN hyperparameters, we also followed the same strategy adopted by the main baseline, except for the number of epochs. In short, the training steps utilized Adam [56] to minimize the cross-entropy loss, using a learning rate of e^{-4}. The models were trained with a batch size of 10.000 posts, for 10 epochs, without hyperparameters optimization.

5.3. Evaluation

The effectiveness evaluation was carried out based on classic machine learning measures such as Precision, Recall and F_1. Additionally, in early detection systems, in addition to the labeling of samples, it is necessary to account for the delay to make the decision. For this reason, we also use the evaluation measure called Early Risk Detection Error ($ERDE_o$) [39], which basically penalizes the late decision making. The delay is measured by the number of distinct posts (k) inspected before making a decision. Considering a d binary decision made by an early risk detection system at k, the $ERDE_o$ is defined as:

$$ERDE_o(d,k) = \begin{cases} c_{fp} & \text{for false positives (FP)} \\ c_{fn} & \text{for false negatives (FN)} \\ lc_o \cdot c_{tp} & \text{for true positives (TP)} \\ 0 & \text{for true negatives (TN)} \end{cases}$$

The $ERDE_o$ function is dependent on o, which controls at what point the cost of a late decision starts to increase more quickly. $ERDE_o$ is defined over three basic terms: c_{fp}, c_{fn} and $lc_o(k) \cdot c_{tp}$. The c_{fp} is the cost of a false positive (prediction that a user has depression when in fact does not). The c_{fn} represents the cost of a false negative (when the system mistakenly identifies a user as non-depressive). Finally, $lc_o \cdot c_{tp}$ indicates the cost of deciding that a user has depression when he/she actually suffers from this condition. In this case, the $lc_o(k)$ factor encodes the cost associated with the delay in detecting true positives, as a way to penalize the late identification. It is worth noting that latency was introduced only for the true positives, because the true negatives are risk-free cases, which, in practice, would not need early intervention. In other words, the function $lc_o(k)$ encodes a cost associated to the delay in detecting true positives and is computed according to Eq. (1).

$$lc_o(k) = 1 - \frac{1}{1 + e^{k-o}} \tag{1}$$

[9] https://colab.research.google.com/ - As of April 19, 2022.
[10] https://www.nltk.org - As of April 19, 2022.
[11] https://keras.io/preprocessing/text - As of April 19, 2022.

Similarly to previous works [11,49,50] the evaluation considered $c_{fn} = c_{tp} = 1$. In turn, c_{fp} was set according to the proportion of positive cases in the test set, that is, 0.1296. As indicated in [49], c_{fn} and c_{tp} were set to 1 because delayed detection can have serious consequences (i.e, late detection is equivalent to not detecting the case).

Given the collection has N individuals, a total of N decisions are made after analyzing the posts of all users. As a result, the general error is calculated by taking the average of the values of $ERDE_o$ for each user. Therefore, the smaller the better. Since all costs are in the interval [0,1], the $ERDE_o$ will also be in the same range. For this study, $ERDE_5$ and $ERDE_{50}$ were used.

For the comparison to the previous methods, the best works from the eRisk 2017 were taken as baselines, specifically, the UNSLA [48], FHDO-BCSGA [13], and FHDO-BCSGB [13]. The UNSLA model corresponds to the method developed in [48], called temporal variation of terms (TVT), combined with Bag of Words (BoW). Specifically, the TVT is an approach for early risk detection based on using the variation of vocabulary along the different time steps as a concept space to represent posts. In turn, the FHDO-BCSGA consists in an ensemble of logistic regression classifiers based on BoW with different weightings of terms and *n-grams*. Finally, the FHDO-BCSGB, also used a logistic regression and relied on the vectorization (doc2vec) to learn vector representations from the user posts.

We also considered additional methods proposed later on, specifically TVT-NB [57], TVT-RF [57] and Trotzek et al. [11]. The TVT-NB and TVT-RF methods were the same used by UNSLA, except for using only TVT without considering BoW representation. The NB and RF indicate the algorithms used, Naive Bayes and Random forest, respectively. In [11], the main baseline, the authors present a CNN, the same used in our work, based on different word embeddings. They also present a logistic regression based on linguistic metadata at the user level (e.g., the average number of the term "I", possessive pronouns, and personal pronouns in posts, frequency of use of the expression "my depression" in posts, frequency of use of words describing medicines for "treatment" of depression, among others). An ensemble of both approaches is also presented. Further details about this baseline are described in the results section.

6. Results and discussion

This section presents both the results obtained in the experiments using the methods based on single embedding models (no embedding fusion), as well as the results achieved by the proposed fusion methods. For clarity purposes, the results are discussed in four sections. Initially, an evaluation is conducted for the available embedding models regarding the preprocessing approaches for emoticons. The second analysis aims at comparing the top performing configurations with their analogue described in [11]. In the third one, the proposed fusion approaches are analyzed in contrast to the baselines, including the state-of-the-art model proposed in [11]. Finally, we analyze the impact of the detection confidence threshold in this early-detection task.

6.1. Importance of underlying emotions

The results achieved by the single embedding models are presented in Table 3. These models were evaluated considering the importance of the underlying emotions encoded in the emoticons. In the first strategy, all emoticons were discarded. In the second one, the meanings of emoticons were preserved through the mapping to semantically equivalent words.

The assessment with multiple measures revealed that using the mapping of emoticons generally allowed better results. For instance, considering general effectiveness in terms of F_1, three out of four models achieved superior effectiveness with the mapping. The only exception was for the *FastText WN* model, which still obtained equivalent F_1. Moreover, the mapping also allowed the best F_1 (0.66), precision

Table 3
Results for individual word embedding models. Comparison of models with emoticon removal and mapping.

Embedding	Emoticons	τ	$ERDE_5$	$ERDE_{50}$	F_1	*Precision*	*Recall*
FastText Crawl	Removal	0.8	12.91	9.07	0.59	0.56	**0.62**
FastText Crawl	Mapping	0.6	**12.40 (4.0%)**	**8.47 (6.6%)**	**0.66 (11.9%)**	**0.74 (32.1%)**	0.60 (−3.2%)
FastText WN	Removal	0.5	13.04	**8.78**	0.62	0.55	**0.69**
FastText WN	Mapping	0.8	**12.78 (2.0%)**	8.98 (−2.3%)	0.62 (0%)	**0.69 (25.5%)**	0.56 (−18.8%)
GloVe Crawl	Removal	0.65	**12.77**	9.63	0.57	**0.60**	0.54
GloVe Crawl	Mapping	0.8	13.38 (−4.8%)	**8.76 (9.0%)**	**0.62 (8.8%)**	0.51 (−15.0%)	**0.79 (46.3%)**
GloVe WN	Removal	0.55	**12.65**	9.06	0.63	0.65	0.62
GloVe WN	Mapping	0.7	12.89 (−1.9%)	**8.84 (2.4%)**	**0.65 (3.2%)**	**0.67 (3.1%)**	**0.63 (1.6%)**

Table 4
Results achieved by our emoticon mapping strategy.

Embedding	Approach	τ	$ERDE_5$	$ERDE_{50}$	F_1	*Precision*	*Recall*
FastTextL Crawl	[11]	0.6	13.01	8.60	0.64	0.60	0.67
FastText Crawl	+Mapping	0.6	**12.40 (4.7%)**	8.47 (1.5%)	**0.66 (11.9%)**	0.74 (23.3%)	0.60 (−3.2%)
FastText WN	[11]	0.55	13.11	7.95	0.60	0.49	0.77
FastText WN	+Mapping	0.8	12.78 (2.5%)	8.98 (−13.0%)	0.62 (3.3%)	**0.69 (40.8%)**	0.56 (−27.3%)
GloVe Crawl	[11]	0.7	12.98	8.59	0.63	0.58	0.69
GloVe Crawl	+Mapping	0.8	13.38 (−3.8%)	8.76 (−2.0%)	0.62 (−1.6%)	0.51 (−12.1%)	**0.79 (14.5%)**
GloVe WN	[11]	0.5	12.95	**7.57**	0.63	0.56	0.56
GloVe WN	+Mapping	0.7	12.89 (0.5%)	8.84 (−16.8%)	0.65 (3.2%)	0.67 (19.6%)	0.63 (12.3%)

(0.74), and recall (0.79), $ERDE_5$ (12.40), and $ERDE_{50}$ (8.47). Table 3 also shows the relative gains achieved by the models with expressive improvements with the mapping. For instance, the best recall and precision represent 46.3% and 32.1% gains. While not as expressive, $ERDE_5$ and $ERDE_{50}$ are also considerably better with the mapping.

These findings corroborate the literature [37,42,52] by showing that the emotions expressed by depressed people are important suggestive evidences of the problem. The results also show a trade-off between higher gains in precision or recall, which is further discussed in Section 6.4.

6.2. Impact of emoticon semantics

Table 4 shows the results of previous embedding-based methods (without emoticon mapping) [11] and variations with semantic extension through emoticon mapping. Considering the emotions, while in the baselines emoticons' embedding is directly performed along with the text, in our approach, the emoticons were previously mapped to their representative words and kept their position. Then, the resulting mapped text was submitted to the embedding procedure. In general, the semantic mapping of emoticons improved the results, with expressive gains in some measures. Moreover, the emoticon mapping allowed overall best results (highlighted in bold) in all measures, except $ERDE_{50}$.

The general best results ($F_1 = 0.66$) and better early detection ($ERDE_5 = 12.40$) from the semantic extension bring decisive improvements for the task, as it would allow earlier intervention. Considering a highly unbalanced dataset, these are valuable results, as it indicates that the system better learned to classify, although exposed to only a few positive depressive cases. In summary, it highlights the impact of exploiting the emoticon semantics for identifying potentially depressive users.

6.3. Fusion effectiveness analysis

Table 5 presents the results achieved by the proposed fusion methods, as well as the results of the baselines. The first part presents the best results achieved in eRisk 2017. The second and third parts show the results obtained afterward, which include the main baseline (FastText Wiki + Meta LR) (best F_1 and $ERDE_5$). The fourth one brings the results obtained by the individual embedding models with the integration of emoticon mapping. Finally, the fifth part brings the results obtained by the proposed early and late fusion approaches. For the late fusion, only the 5 top-performing configurations (best F_1) are reported. The base models used in the fusion methods are the ones with emoticon mapping presented in part III.

Specifically, regarding the models from Trotzek et al. [11], the ones named based on a word embedding were trained using the same CNN considered in our experiments. Moreover, "Meta LR" refers to models based on logistic regression trained using the selected Linguistic Inquiry and Word Count (LIWC) features in combination with metadata features, e.g., average usage of the word "I" in posts, average of possessive pronouns, average of personal pronouns, frequency of use of the expression "my depression" in posts, frequency of use of words describing drugs to "treat" depression, among others).

The early fusion method, EF1 and EF2 specifically, outperformed most of the baselines according to different effectiveness measures. In addition, considering F_1, one of the main measures used in the eRisk 2017, EF1 outperformed all the methods that rely only on individual embeddings, including the proposed extensions with emoticon mapping. These results suggest the use of early fusion methods are a promising solution. As an extension, additional representation models could be included in this process to possibly enrich the cross-model feature integration.

Still considering the early fusion approach, in terms of F_1, when the fusion was performed using GVWN, there was no improvement against the proposed extensions with emoticon mapping. As GVWN was the embedding that had the lowest number of word vectors (0.4 million), this may have influenced the result, given that some words might not have a vector representation in the embedding. However, considering the Recall measure, there was a significant improvement with the EF6, compared to the individual models, i.e, FTWN (gain of 30.4%) and GVWN (gain of 15.9%).

The *late fusion* approach was similar to or outperformed the baselines in multiple measures. In terms of F_1, it allowed results ≥ 0.68 for all evaluated models (Late Fusion 1 to 5), with this minimum value outperforming most of its component models. Specifically, considering Recall, the proposed method achieved higher values than several baselines, being only lower than the baseline (FastText Wiki + Meta LR). However, the superiority of this baseline came at a very high cost, as it ended up generating a large number of false positives. On the other hand, our method presents a better balance between these measures. A more detailed discussion about this trade-off is presented in Section 6.4.

Table 5
Overall results for the proposed methods and the baselines. The first three parts are the baselines. The other two ones are our results. The best results are highlighted in bold.

Part	Method	Model	τ	$ERDE_5$	$ERDE_{50}$	F_1	P	R
I	Best eRisk models [49]	UNSLA	–	13.66	**9.68**	0.59	0.48	**0.79**
		FHDO-BCSGA	–	12.82	9.69	**0.64**	0.61	0.64
		FHDO-BCSGB	–	**12.70**	10.39	0.55	**0.69**	0.46
II	Villegas *et al.* [57]	TVT-NB	–	13.13	**8.17**	0.54	0.42	**0.73**
		TVT-RF	–	**12.30**	8.95	**0.56**	**0.54**	0.58
III	Trotzek *et al.* [11]	Glove WN	0.5	12.95	7.57	0.63	0.56	0.73
		GloVe Crawl	0.7	12.98	8.59	0.63	0.58	0.69
		FastText Wiki	0.6	13.06	8.17	0.57	0.47	0.71
		FastText WN	0.55	13.11	7.95	0.60	0.49	0.77
		FastText Crawl	0.6	13.01	8.60	0.64	0.60	0.67
		FastText reddit	0.7	13.52	8.04	0.62	0.51	0.79
		FastText reddit	0.8	12.71	9.23	0.56	0.63	0.50
		Meta LR	0.35	12.65	8.57	0.66	0.59	0.73
		Meta LR	0.55	12.35	9.86	0.65	0.72	0.60
		Glove WN + Meta LR	0.45	12.34	8.93	0.71	**0.72**	0.69
		FastText Wiki + Meta LR	0.35	13.52	**7.29**	0.55	0.41	**0.85**
		FastText Wiki + Meta LR	0.5	**12.13**	8.77	**0.71**	0.71	0.71
		FastText reddit+ Meta LR	0.55	12.46	8.77	0.67	0.69	0.65
IV	Proposed (Semantic mapping)	FastText Crawl (FTC)	0.6	**12.40**	**8.47**	**0.66**	**0.74**	0.60
		FastText WN (FTWN)	0.8	12.78	8.98	0.62	0.69	0.56
		GloVe Crawl (GVC)	0.8	13.38	8.76	0.62	0.51	**0.79**
		GloVe WN (GVWN)	0.7	12.89	8.84	0.65	0.67	0.63
V	Fusion approaches	Early Fusion 1 (EF1) (FTC + GVC)	0.7	12.71	9.03	0.67	0.73	0.62
		Early Fusion 2 (FTWN + GVC)	0.75	12.87	9.10	0.66	0.67	0.65
		Early Fusion 3 (FTC + FTWN)	0.65	12.67	9.04	0.62	0.59	0.65
		Early Fusion 4 (FTC + GVWN)	0.75	12.68	9.35	0.62	0.72	0.54
		Early Fusion 5 (GVC + GVWN)	0.8	**12.64**	9.33	0.60	0.64	0.56
		Early Fusion 6 (FTWN + GVWN)	0.6	13.26	**8.71**	0.60	0.51	**0.73**
		Late Fusion 1 (FTC + FTWN + GVWN)	–	13.29	9.52	0.68	0.63	**0.73**
		Late Fusion 2 (FTC + FTWN + GVWN + GVC)	–	12.99	8.94	0.68	0.73	0.63
		Late Fusion 3 (EF1 + FTC + GVWN)	–	13.05	9.29	0.70	0.71	0.69
		Late Fusion 4 (EF1 + FTC + GVWN + GVC)	–	13.00	9.23	**0.71**	0.73	0.69
		Late Fusion 5 (EF1 + FTC + FTWN + GVC)	–	12.79	8.91	0.70	**0.76**	0.65

In addition, it also maintained competitive results in terms of $ERDE_o$, which was adapted to assess *late fusion* results. For this, rigorously, $ERDE_o$ was calculated using the k value of the system, among those participating in the ensemble, which needed to evaluate the largest number of posts for the final decision.

6.4. Detection threshold and effectiveness trade-off

Considering the importance of maximizing the detection of depressive users, recall is highlighted as an important effectiveness measure. In this context, the best models reported in [11] achieved high values (up to 0.85) but were followed with low precision (0.41), indicating a large number of false positives. This trade-off is depicted in Fig. 4 through a recall vs. precision scatter plot. All models from Table 5 are presented, and most of them allowed higher precision or recall at the expense of the other. The best trade-off is achieved by the models based on meta features or that relied on the proposed fusion approaches. Moreover, the fusion allowed similar F_1, even without using the meta features, but only the textual information from users' posts.

The detection confidence threshold has a direct impact on the results. Lower threshold values allow positive detection even with low confidence, which results in the higher recall. Nevertheless, it is usually followed by false positives and, consequently, lower precision. Fig. 3 illustrates this trade-off for the GVWN model from Table 5. Similar to FastText Wiki + Meta LR, this model is able to achieve high recall (0.88) with proper adjustment of the threshold. However, this result comes at the expense of quite low precision. This could make the use of this system impractical in the real world, as many users could be subjected to unnecessary interventions. A similar impact is also observed over the $ERDE_o$ measure. This analysis emphasizes the importance of simultaneously considering multiple measures, such as F_1, precision and recall for the effectiveness assessment of such critical detection methods.

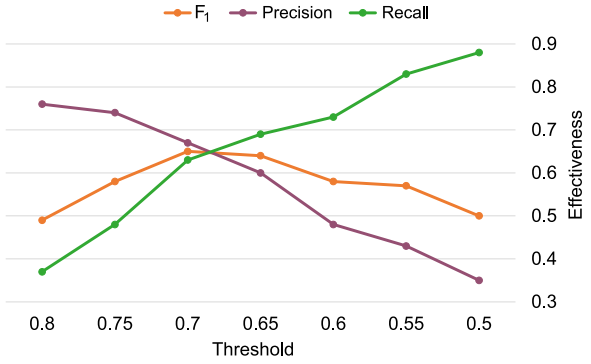

Fig. 3. Detection threshold and effectiveness trade-off.

7. Challenges discussion and research directions

The results of this study revealed the importance of social media in terms of detecting depression. The language used by users proved to be a valuable resource to accomplish this task. However, despite the advances, there are still many challenges and open questions. In fact, many subjects still demand deeper investigations such as data collection and usage, ethical aspects, investigation o novel techniques to face the problem, as well as the definition of interventions that could be carried out after detection.

7.1. Prediction methods and feature representation

Over the years, many approaches have been applied to analyze data from social media, from techniques based on the analysis of

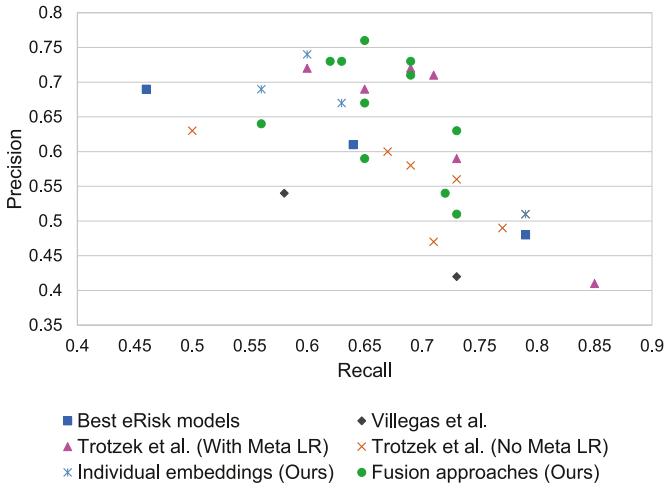

Fig. 4. Scatter plot between precision and recall for all models considered in the Table 5.

feelings [58] to the use of language modeling [59]. The search for more effective models has led researchers to explore promising techniques and, in some cases, their combination. Nevertheless, other techniques have not been widely investigated and may attract research focus in the near future.

Similar to our work, some studies have relied on static word embedding models for feature representations, but considering those models individually. In turn, we use Early and Late Fusion approaches. However, considering that the models used are static, that is, each word always has the same vector representation, the use of contextualized language representation models, such as BERT [60] and XLNet [61] deserves to be investigated in depth. Another direction worth exploring regards the use of transfer learning from models that have achieved high effectiveness in different tasks. Hence, it may allow using both pre-trained embeddings from an alternative database, as well as using data related to the problem for fine-tuning. Regarding the fusion methods, other hybrid fusion techniques could be investigated in-depth, for example, Autoencoders and Tf–idf.

Another approach to be investigated concerns the combination of data already retrieved from social media with data from another source, such as data related to sleep patterns, physical activities, neurotransmitters, as well as information about food and habits. With this, it is expected to improve the model's effectiveness and reliability, since depression patients commonly present symptoms related to these and other characteristics [62]. However, for this to be accomplished, it is important to overcome the ethical issues involved, notably the ethical issues of research with human beings and data privacy on social media that, in general, are not yet fully understood by the ethics councils and the general public [63].

An alternative not broadly investigated in this context is the Reinforcement Learning (RL) technique. In this sense, new approaches must be developed to introduce RL in one of the stages of the generation of the predictive model. For example, the authors in [64] applied this strategy to select the posts that are relevant to the task of detecting depression. Furthermore, it may also contribute to the selection process of representative features to be used in the training of predictive models.

In this work, the proposed method was evaluated using textual messages from social media. However, the proposed method may be used for different types of text sequences, for instance, e-mails or instant messages. Nevertheless, it should be properly evaluated and optimized for the specific context considering the inherent characteristics of such media, e.g., the text size, writing styles, and vocabulary.

7.2. Interventions and public policy

New proposals related to healthcare informatics, such as detecting depression, commonly lead people to discuss the practical uses of this type of system, for example, its use in the intervention process. Regarding the practical use of this type of system, but in the context of intervention for anorexic people, De Choudhury [65] argues that the design considerations in this space need to ensure that the benefits obtained by the intervention exceed the risks. In this same study, the author indicates that the intervention can be carried out by communicating the risks directly to the individuals or reliable social or clinical contacts. The use of such an approach could be investigated in the context of depression detection systems.

With respect to eating disorders, some platforms, such as Tumblr,[12] have already offered basic intervention measures to help vulnerable people [65]. Therefore, this type of initiative could be expanded to other social media and health conditions. In addition, it is important to include other disorders, especially depression. However, it is important that interventions are carried out in a non-intrusive way, so that the user's problem does not worsen, or lead to other problems. For this, it is important to rely on a multidisciplinary team to outline how this type of strategy could be effectively put into practice.

Interventions can be carried out in several ways, from a pop-up message that directs the user to a psychological support entity or an anonymous private message to that user, or even by contacting a person close to this user who can contribute to improving his/her well-being. Additionally, the platform itself could send motivational messages or even advertisements that encourage that person to seek help from an expert. Notice that this type of approach must be performed in a way that the user does not feel violated. For these researches and actions to be developed, it is necessary that effective public policies are created. If effective results were achieved, the whole society would benefit.

7.3. Ethics and legal issues

In a research process, beyond the many biases involved, it is necessary to be aware of the ethical aspects and legal challenges related to the theme. Considering studies related to depression, the researcher needs to understand that the user-generated content presents sensitive content, and these people, in general, are not open to talk about it conventionally. Thus, researchers need to be aware of the ethical and regulatory challenges that surround AI in the healthcare field, as a way to avoid disparities or even negatively impact the user's health status, which may already be affected.

Beyond the issues related to the process of obtaining data, it is important that system is committed to justice. In this sense, researchers must keep attention to the biases that these intelligent systems could embed, such as racial and gender issues. Scientific societies and regulatory agencies should develop best practices to recognize and minimize the effects of biased training data sets [66]. Finally, one of the most important challenges is related to the interpretable nature of some algorithms, often called a black box. The difficulty of explaining an automated decision can end up reducing the reliability of the system. In this sense, it is important that the development of these systems follow guidelines or produce information that facilitates the understanding of why the system made a certain decision [67]. Therefore, some questions are still worth debating such as: should users be aware of the use of an automated system to promote their diagnosis? or else, given that a prediction was made incorrectly, who will be responsible?.

Although some ethical discussions on these subjects have already been carried out, there is still a need for a more solid understanding about the automation of healthcare decisions and their impacts. It demands a multidisciplinary discussion, involving multiple science professionals and society representatives. The guidelines that may arise from these discussions must be duly supervised, to ensure that they are being followed.

[12] https://www.tumblr.com/ - As of April 10, 2022.

8. Conclusion

Mental illnesses are one of the most prevalent public health problems worldwide. Among these, depression stands out due to the numerous problems, such as suicide. Thus, we propose an early detection method using a CNN, in combination with Early and Late Fusion strategies. A set of pre-trained embeddings and their fusion were evaluated with the proposed method as a way of representing textual features. Our findings suggest that the pre-trained embeddings are able to build a good representation of the language used by the users. Comparing with the literature, the proposed models achieved better results.

Similar to the results obtained for the embeddings individually, the proposed method also obtained promising results. In all experiments performed, it showed equivalent values or numerical superiority, compared to individual embeddings trained with CNN, as well as in relation to baselines. These results were achieved by preserving a trade-off between all effectiveness criteria. In addition, even with a high imbalance of the database, the models managed to achieve superior effectiveness. Beyond it, the emoticons mapping allowed the best results, which ratifies that emotions are an important way to characterize people with depression.

As future work, one may focus on: (*i*) evaluate the use of alternative language modeling methods such as BERT and XLNet; (*ii*) assessing the impact of pre-trained embeddings against training from scratch; (*iii*) evaluate the proposed method in a novel, possibly larger, databases. In addition, we intend to apply intermediate fusion methods to face the task of early detection of depression and to use others hybrid fusion techniques. We also intend to use different embeddings for a post together along the third dimension (like the R, G and B channels of an RGB image) and training a network with those samples; (*iv*) include other relevant information in the model generation, such as gender and age, and assess the impact of this on its performance; and (*v*) Conduct further experiments with hyperparameter tuning with large datasets and more robust validation strategies.

CRediT authorship contribution statement

José Solenir L. Figuerêdo: Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing, Investigation. **Ana Lúcia L.M. Maia:** Conceptualization, Methodology, Validation, Resources, Writing – review & editing. **Rodrigo Tripodi Calumby:** Supervision, Conceptualization, Methodology, Validation, Resources, Project administration, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] N.I.M. Health, Depression basics, 2016, https://www.nimh.nih.gov/health/publications/depression/index.shtml. (Accessed 24 August 2021).

[2] W.H. Organization, Depression and other common mental disorders: Global health estimates, 2017, https://apps.who.int/iris/bitstream/handle/10665/254610/WHO-MSD-MER-2017.2-eng.pdf. (Accessed 24 August 2021).

[3] W.H. Organization, Depression, 2018, https://www.who.int/en/news-room/fact-sheets/detail/depression. (Accessed 24 August 2021).

[4] M.D. Choudhury, M. Gamon, S. Counts, E. Horvitz, Predicting depression via social media, in: Proceedings of the 18th ICWSM, Cambridge, Massachusetts, USA, July 8-11, 2013, URL https://www.microsoft.com/en-us/research/publication/predicting-depression-via-social-media/. (Accessed 24 August 2021).

[5] T.M.H. Li, M. Chau, P.W.C. Wong, P.S.F. Yip, A hybrid system for online detection of emotional distress, in: Proceedings of the PAISI, Kuala Lumpur, Malaysia, May 29, 2012, pp. 73–80, http://dx.doi.org/10.1007/978-3-642-30428-6_6.

[6] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, G.-Z. Yang, Deep learning for health informatics, IEEE J. BHI 21 (1) (2016) 4–21, http://dx.doi.org/10.1109/JBHI.2016.2636665.

[7] Q. Cheng, T.M. Li, C.-L. Kwok, T. Zhu, P.S. Yip, Assessing suicide risk and emotional distress in Chinese social media: A text mining and machine learning study, JMIR 19 (7) (2017) e243, http://dx.doi.org/10.2196/jmir.7276.

[8] A.B. Menegotto, C.D.L. Becker, S.C. Cazella, Computer-aided diagnosis of hepatocellular carcinoma fusing imaging and structured health data, Health Inf. Sci. Syst. 9 (1) (2021) 20, http://dx.doi.org/10.1007/s13755-021-00151-x.

[9] A. Brock, J. Donahue, K. Simonyan, Large scale GAN training for high fidelity natural image synthesis, 2018, CoRR, arXiv:1809.11096.

[10] Y. Tian, K. Pei, S. Jana, B. Ray, DeepTest: automated testing of deep-neural-network-driven autonomous cars, in: Proceedings of the 40th ICSE, Gothenburg, Sweden, May 27 - June 03, 2018, pp. 303–314, http://dx.doi.org/10.1145/3180155.3180220.

[11] M. Trotzek, S. Koitka, C.M. Friedrich, Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences, IEEE TKDE 32 (3) (2020) 588–601, http://dx.doi.org/10.1109/TKDE.2018.2885515.

[12] A.H. Orabi, P. Buddhitha, M.H. Orabi, D. Inkpen, Deep learning for depression detection of Twitter users, in: Proceedings of the 5th CLPsych@NAACL-HTL, New Orleans, la, USA, 2018, pp. 88–97, URL https://aclanthology.org/W18-0609/. (Accessed 24 August 2021).

[13] M. Trotzek, S. Koitka, C.M. Friedrich, Linguistic metadata augmented classifiers at the CLEF 2017 task for early detection of depression, in: Working Notes of CLEF, Dublin, Ireland, 2017, URL http://ceur-ws.org/Vol-1866/paper_54.pdf. (Accessed 24 August 2021).

[14] M.Y. Wu, C.-Y. Shen, E.T. Wang, A.L.P. Chen, A deep architecture for depression detection using posting, behavior, and living environment data, JIIS (2018) http://dx.doi.org/10.1007/s10844-018-0533-4.

[15] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016.

[16] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444, http://dx.doi.org/10.1038/nature14539.

[17] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, T. Chen, Recent advances in convolutional neural networks, Pattern Recognit. 77 (2018) 354–377, http://dx.doi.org/10.1016/j.patcog.2017.10.013.

[18] H. Salehinejad, J. Baarbe, S. Sankar, J. Barfett, E. Colak, S. Valaee, Recent advances in recurrent neural networks, 2018, CoRR, arXiv:1801.01078. (Accessed 24 August 2021).

[19] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N.E.Y. Soplin, R. Yamamoto, X. Wang, S. Watanabe, T. Yoshimura, W. Zhang, A comparative study on transformer vs RNN in speech applications, in: 2019 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU, 2019, pp. 449–456, http://dx.doi.org/10.1109/ASRU46091.2019.9003750.

[20] Y. Zhang, B.C. Wallace, A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification, in: Proceedings of the 8th IJCNLP, Taipei, Taiwan, November 27 - December 1, Volume 1: Long Papers, 2017, pp. 253–263, URL https://aclanthology.org/I17-1026/. (Accessed 24 August 2021).

[21] J. Du, L. Gui, R. Xu, Y. He, A convolutional attention model for text classification, in: X. Huang, J. Jiang, D. Zhao, Y. Feng, Y. Hong (Eds.), Natural Language Processing and Chinese Computing, Springer International Publishing, Cham, 2018, pp. 183–195, http://dx.doi.org/10.1007/978-3-319-73618-1_16.

[22] H. Wang, J. He, X. Zhang, S. Liu, A short text classification method based on N-gram and CNN, Chinese J. Electron. 29 (2) (2020) 248–254, http://dx.doi.org/10.1049/cje.2020.01.001.

[23] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, et al., Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324, http://dx.doi.org/10.1109/5.726791.

[24] S.V. Georgakopoulos, S.K. Tasoulis, A.G. Vrahatis, V.P. Plagianakos, Convolutional neural networks for toxic comment classification, in: Proceedings of the 10th SETN, Patras, Greece, July 09-12, 2018, ACM, 2018, pp. 35:1–35:6, http://dx.doi.org/10.1145/3200947.3208069.

[25] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, B. Qin, Learning sentiment-specific word embedding for Twitter sentiment classification, in: Proceedings of the 52nd Annual Meeting of the ACL, June 22-27, 2014, Baltimore, MD, USA, Vol. 1, 2014, pp. 1555–1565, http://dx.doi.org/10.3115/v1/p14-1146.

[26] Y. Goldberg, Neural Network Methods for Natural Language Processing, in: Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers, 2017, http://dx.doi.org/10.2200/S00762ED1V01Y201703HLT037.

[27] H. Rubenstein, J.B. Goodenough, Contextual correlates of synonymy, Commun. ACM 8 (10) (1965) 627–633, http://dx.doi.org/10.1145/365628.365657.

[28] C. Wang, P. Nulty, D. Lillis, A comparative study on word embeddings in deep learning for text classification, in: NLPIR 2020, Association for Computing Machinery, New York, NY, USA, 2020, pp. 37–46, http://dx.doi.org/10.1145/3443279.3443304.

[29] S. Wang, W. Zhou, C. Jiang, A survey of word embeddings based on deep learning, Computing 102 (3) (2020) 717–740, http://dx.doi.org/10.1007/s00607-019-00768-7.

[30] Z. Liu, C. Li, X. Gao, G. Wang, J. Yang, Ensemble-based depression detection in speech, in: IEEE BIBM, Kansas City, MO, USA, November 13-16, 2017, pp. 975–980, http://dx.doi.org/10.1109/BIBM.2017.8217789.

[31] M. Ebersbach, R. Herms, M. Eibl, Fusion methods for ICD10 code classification of death certificates in multilingual corpora, in: L. Cappellato, N. Ferro, L. Goeuriot, T. Mandl (Eds.), Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017, in: CEUR Workshop Proceedings, vol. 1866, CEUR-WS.org, 2017, URL http://ceur-ws.org/Vol-1866/paper_66.pdf. (Accessed 24 August 2021).

[32] S. Nemati, R. Rohani, M.E. Basiri, M. Abdar, N.Y. Yen, V. Makarenkov, A hybrid latent space data fusion method for multimodal emotion recognition, IEEE Access 7 (2019) 172948–172964, http://dx.doi.org/10.1109/ACCESS.2019.2955637.

[33] K. Bayoudh, R. Knani, F. Hamdaoui, A. Mtibaa, A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets, Vis. Comput. (2021) http://dx.doi.org/10.1007/s00371-021-02166-7.

[34] J.A. Vargas Muñoz, R. da Silva Torres, M.A. Gonçalves, A soft computing approach for learning to aggregate rankings, in: CIKM '15, Association for Computing Machinery, New York, NY, USA, 2015, pp. 83–92, http://dx.doi.org/10.1145/2806416.2806478.

[35] J.W. Pennebaker, M.R. Mehl, K.G. Niederhoffer, Psychological aspects of natural language use: Our words, our selves, Ann. Rev. Psychol. 54 (1) (2003) 547–577, http://dx.doi.org/10.1146/annurev.psych.54.101601.145041.

[36] S. Rude, E.-M. Gortner, J. Pennebaker, Language use of depressed and depression-vulnerable college students, Cogn. Emot. 18 (8) (2004) 1121–1133.

[37] F. Benamara, V. Moriceau, J. Mothe, F. Ramiandrisoa, Z. He, Automatic detection of depressive users in social media, in: Proceedings of the French Information Retrieval Conference, Rennes, France, 2018, http://dx.doi.org/10.24348/coria.2018.paper4.

[38] P. Burnap, G. Colombo, R. Amery, A. Hodorog, J. Scourfield, Multi-class machine classification of suicide-related communication on Twitter, Online Soc. Netw. Media 2 (2017) 32–44, http://dx.doi.org/10.1016/j.osnem.2017.08.001.

[39] D.E. Losada, F. Crestani, A test collection for research on depression and language use, in: Proceedings of CLEF, Évora, Portugal, 2016, pp. 28–39, http://dx.doi.org/10.1007/978-3-319-44564-9_3.

[40] C. Yang, P. Srinivasan, Life satisfaction and the pursuit of happiness on Twitter, PLoS One 11 (3) (2016) e0150881, http://dx.doi.org/10.1371/journal.pone.0150881.

[41] M.D. Choudhury, S. Counts, E. Horvitz, Predicting postpartum changes in emotion and behavior via social media, in: ACM SIGCHI, Paris, France, April 27 - May 2, 2013, 2013, pp. 3267–3276, http://dx.doi.org/10.1145/2470654.2466447.

[42] X. Wang, C. Zhang, Y. Ji, L. Sun, L. Wu, Z. Bao, A depression detection model based on sentiment analysis in micro-blog social network, in: Trends and Applications in Knowledge Discovery and Data Mining - PAKDD 2013 International Workshops: DMApps, DANTH, QIMIE, BDM, CDA, CloudSD, Gold Coast, QLD, Australia, April 14-17, 2013, Revised Selected Papers, 2013, pp. 201–213, http://dx.doi.org/10.1007/978-3-642-40319-4_18.

[43] Q. Hu, A. Li, F. Heng, J. Li, T. Zhu, Predicting depression of social media user on different observation windows, in: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT 2015, Singapore, December 6-9, 2015 - Vol. I, 2015, pp. 361–364, http://dx.doi.org/10.1109/WI-IAT.2015.166.

[44] H.A. Schwartz, J. Eichstaedt, M.L. Kern, G. Park, M. Sap, D. Stillwell, M. Kosinski, L. Ungar, Towards assessing changes in degree of depression through facebook, in: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal To Clinical Reality, 2014, pp. 118–125, URL https://aclanthology.org/W14-3214/. (Accessed 24 August 2021).

[45] M.R. Islam, M.A. Kabir, A. Ahmed, A.R.M. Kamal, H. Wang, A. Ulhaq, Depression detection from social network data using machine learning techniques, Health Inf. Sci. Syst. 6 (1) (2018) 8, http://dx.doi.org/10.1007/s13755-018-0046-0.

[46] M.M. Tadesse, H. Lin, B. Xu, L. Yang, Detection of depression-related posts in reddit social media forum, IEEE Access 7 (2019) 44883–44893, http://dx.doi.org/10.1109/ACCESS.2019.2909180.

[47] S. Tsugawa, Y. Kikuchi, F. Kishino, K. Nakajima, Y. Itoh, H. Ohsaki, Recognizing depression from Twitter activity, in: Proceedings of ACM CHI, Seoul, Republic of Korea, April 18-23, 2015, pp. 3187–3196, http://dx.doi.org/10.1145/2702123.2702280.

[48] M.L. Errecalde, M.P. Villegas, D.G. Funez, M.J.G. Ucelay, L.C. Cagnina, Temporal variation of terms as concept space for early risk prediction, in: L. Cappellato, N. Ferro, L. Goeuriot, T. Mandl (Eds.), Working Notes of CLEF, Dublin, Ireland, Vol. 1866, 2017, URL http://ceur-ws.org/Vol-1866/paper_103.pdf. (Accessed 24 August 2021).

[49] D.E. Losada, F. Crestani, J. Parapar, eRISK 2017: CLEF lab on early risk prediction on the internet: Experimental foundations, in: Proceedings of CLEF 2017, Dublin, Ireland, 2017, pp. 346–360, http://dx.doi.org/10.1007/978-3-319-65813-1_30.

[50] V. Leiva, A. Freire, Towards suicide prevention: Early detection of depression on social media, in: Proceedings of the 4th INSCI, Thessaloniki, Greece, November 22-24, 2017, pp. 428–436, http://dx.doi.org/10.1007/978-3-319-70284-1_34.

[51] W. Shang, K. Sohn, D. Almeida, H. Lee, Understanding and improving convolutional neural networks via concatenated rectified linear units, in: Proceedings of the 33nd ICML, New York City, NY, USA, June 19-24, 2016, pp. 2217–2225, http://dx.doi.org/10.5555/3045390.3045624.

[52] N. Vedula, S. Parthasarathy, Emotional and linguistic cues of depression from social media, in: Proceedings of the 2017 ICDHT, London, United Kingdom, July 2-5, 2017, 2017, pp. 127–136, http://dx.doi.org/10.1145/3079452.3079465.

[53] C. Lin, P. Hu, H. Su, S. Li, J. Mei, J. Zhou, H. Leung, SenseMood: Depression detection on social media, in: Proceedings of the 2020 International Conference on Multimedia Retrieval, in: ICMR '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 407–411, http://dx.doi.org/10.1145/3372278.3391932.

[54] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, A. Joulin, Advances in pre-training distributed word representations, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018.

[55] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: Proceedings of the EMNLP, October 25-29, Doha, Qatar, 2014, pp. 1532–1543, http://dx.doi.org/10.3115/v1/d14-1162.

[56] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015, URL http://arxiv.org/abs/1412.6980.

[57] M.P. Villegas, D.G. Funez, M.J.G. Ucelay, L.C. Cagnina, M.L. Errecalde, LIDIC - unsl's participation at erisk 2017: Pilot task on early detection of depression, in: Working Notes of CLEF, Dublin, Ireland, 2017, URL http://ceur-ws.org/Vol-1866/paper_107.pdf. (Accessed 24 August 2021).

[58] X. Wang, C. Zhang, Y. Ji, L. Sun, L. Wu, Z. Bao, A depression detection model based on sentiment analysis in micro-blog social network, in: Trends and Applications in Knowledge Discovery and Data Mining - PAKDD 2013 International Workshops: DMApps, DANTH, QIMIE, BDM, CDA, CloudSD, Gold Coast, QLD, Australia, April 14-17, 2013, Revised Selected Papers, 2013, pp. 201–213, http://dx.doi.org/10.1007/978-3-642-40319-4_18.

[59] M. Trotzek, S. Koitka, C.M. Friedrich, Word embeddings and linguistic metadata at the CLEF 2018 tasks for early detection of depression and anorexia, in: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018, 2018.

[60] J. Devlin, M. Chang, K. Lee, K. Toutanova, 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, Vol. 1, Long and Short Papers, 2019, pp. 4171–4186.

[61] Z. Yang, Z. Dai, Y. Yang, J.G. Carbonell, R. Salakhutdinov, Q.V. Le, XLNet: Generalized autoregressive pretraining for language understanding, 2019, CoRR, arXiv:1906.08237.

[62] A.P. Association, Diagnostic and Statistical Manual of Mental Disorders : DSM-5, fifth ed., American Psychiatric Association Arlington, VA, 2013.

[63] A. Wongkoblap, M.A. Vadillo, V. Curcin, Researching mental health disorders in the era of social media: Systematic review, J. Med. Internet Res. 19 (6) (2017) e228, http://dx.doi.org/10.2196/jmir.7215.

[64] T. Gui, Q. Zhang, L. Zhu, X. Zhou, M. Peng, X. Huang, Depression detection on social media with reinforcement learning, in: M. Sun, X. Huang, H. Ji, Z. Liu, Y. Liu (Eds.), Chinese Computational Linguistics - 18th China National Conference, CCL 2019, Kunming, China, October 18-20, 2019, Proceedings, in: Lecture Notes in Computer Science, vol. 11856, Springer, 2019, pp. 613–624, http://dx.doi.org/10.1007/978-3-030-32381-3_49.

[65] M.D. Choudhury, Anorexia on tumblr: A characterization study, in: P. Kostkova, F. Grasso (Eds.), Proceedings of the 5th International Conference on Digital Health 2015, Florence, Italy, May 18-20, 2015, ACM, 2015, pp. 43–50, http://dx.doi.org/10.1145/2750511.2750515.

[66] E. Vayena, A. Blasimme, I.G. Cohen, Machine learning in medicine: Addressing ethical challenges, PloS Med. 15 (11) (2018) 1–4, http://dx.doi.org/10.1371/journal.pmed.1002689.

[67] B. Mittelstadt, C. Russell, S. Wachter, Explaining explanations in AI, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, in: FAT* '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 279–288, http://dx.doi.org/10.1145/3287560.3287574.