Neural Computing and Applications (2022) 34:20149–20164
https://doi.org/10.1007/s00521-022-07569-8

ORIGINAL ARTICLE

Approaching what and how people with mental disorders communicate in social media–Introducing a multi-channel representation

Mario Ezra Aragón[1] · A. Pastor López-Monroy[2] · Luis C. González[3] · Manuel Montes-y-Gómez[1]

Received: 30 December 2021 / Accepted: 27 June 2022 / Published online: 19 July 2022

Abstract
Over the last few years, studies related to the detection of mental disorders in social media have been increasing. The latter because the awareness created by health campaigns that emphasizes the commonness of these disorders among all of us has motivated the creation of new datasets, many of them extracted from social media platforms. In this study, we aim to contribute to the analysis of three major mental disorders that are hitting the world: Anorexia, Depression and Self-harm. To this end, we propose a novel model that, first, extracts three different views, or information channels, from the posts shared by users: thematic interests, writing style, and emotions. Then, it optimally fusions the information from each channel by using a gated multimodal unit. We evaluate the feasibility of our approach in the aforementioned tasks, first by comparing its output against traditional and modern strategies, and later against the best contestants in the eRisk evaluation forum. In both evaluations, our approach clearly outperforms all of its competitors. Through an exhaustive analysis section, we provide evidence of what is being captured by each information channel, then highlighting the importance and robustness of a more holistic view in critical classification tasks.

Keywords Mental disorders · Social media · Multi-channel representation · Deep learning

1 Introduction

Most people believe that mental disorders are uncommon or only happen to people with specific personal profiles, when in fact, they are prevalent and very familiar [1]. Common mental disorders such as anorexia, depression, dementia, post-traumatic stress disorder (PTSD), and schizophrenia affect millions of people around the world [2]. Briefly, a mental disorder is a disease that causes different disturbances in the thinking and behavior of the affected person. These interferences could vary from mild to severe, and could result in an inability to live, respond to ordinary demands or perform routines in daily life. According to the Institute for Health Metrics Evaluation (IHME), about 13% of the global population (971 million people) suffer from some kind of mental disorder [3]. Similarly, a 2018 study of mental disorders in Mexico revealed that 17% of people in the country have at least one mental disorder and one in four will suffer at least one in their lifetime [4]. Nowadays, social media platforms provide the possibility for people to share information, then exposing interests, thoughts, worries and opinions. This presents an opportunity to understand how language is used by people experiencing a mental disorder. Although in general this approach is applied to the masses, it could also offer the opportunity, under very rigorous anonymity

✉ Mario Ezra Aragón
mearagon@inaoep.mx

A. Pastor López-Monroy
pastor.lopez@cimat.mx

Luis C. González
lcgonzalez@uach.mx

Manuel Montes-y-Gómez
mmontesg@inaoep.mx

[1] National Institute of Astrophysics Optics and Electronics(INAOE), Puebla, Mexico

[2] Mathematics Research Center (CIMAT), Guanajuato, Mexico

[3] School of Engineering, Autonomous University of Chihuahua, Chihuahua, Mexico

clauses, to help people who suffer from mental disorders get professional help in a timely manner [5, 6].

Many typical analyses run on the information shared by users and only considers the thematic aspect of the content, simply ignoring important patterns that may be beyond the topics. Thus, the hypothesis of this work is that there are other dimensions of the communication that provide insightful information to characterize users, for example the writing style or even the emotions transmitted in the text. Accordingly, the goal of this study is to present a novel approach that exploits all these different views and obtains a more holistic representation of the users, which we name as multi-channel representation. For this purpose, we define a *channel* as a different property or view from the same modality [7]. In this work, we use the text modality and three channels that will separately focus on different aspects of the users' shared content. The first is the thematic information, the second corresponds to the expressed emotions, and the third is the author writing style. The intuition of our approach is that people that present some mental disorder tend to express differently, at diverse dimensions, regarding the control group. For example, people tend to repeatedly bring up topics related to prior traumas or even sentimental relationships, at the same time they communicate particular emotions such as anger and disgust. In this work, we study how all of these different communication aspects can be captured and combined to offer a more integrated view of the users, providing evidence that although each channel is different they complement each other. Interestingly, all these components are related to the ways humans analyze not only *what* is communicated through messages, but also the manner in *which* it is expressed.

Summarizing, the main contributions of this study are the following:

1. We propose a new representation based on the writing style of the users that allows capturing their writing variability. This representation complements the emotion-based representation described in [8] as well as the traditional thematic-centered representations offered by, for example, GloVe [9] and BERT [10] embeddings.
2. We propose a dynamically weighted late-fusion approach to combine three different information channels: thematic, emotion, and style; Results clearly suggest the feasibility of our approach, even improving state of the art results for the detection of anorexia, depression, and self-harm.
3. We analyze and evaluate in detail these three information channels and the importance of their fusion. By this characterization, we aim to provide evidence of its robustness for the detection of mental disorders in social media.

The remainder of the paper is organized as follows: Sect. 2 presents a brief overview of the detection of mental health disorders using social media data. Sect. 3 describes in detail the creation of these three channels. Sect. 4 presents our multi-channel classification model. Sect. 5 describes in detail our experiments, results, and their analysis. Finally, Sect. 6 presents our main conclusions.

2 Related work

In this section, we present an overview of previous works about the detection of anorexia, depression, and self-harm using social media data. We describe their strengths and opportunities, and contrast the strategies used in our proposal.

Several recent works have taken advantage of social media platforms to study the manifestation of different mental disorders. Most of them used crowd-sourcing strategies to collect data [11]. In general, they identify a group of users who expressed in one of their publications having been clinically diagnosed with a mental disorder, and then download all or part of their posts [12, 13]. Having obtained their data, they apply a variety of methods to find relevant and discriminative patterns from the platform usage behavior, social interactions or language use.

Regarding the use of language, some works have employed traditional classification algorithms combined with the analysis of words or word sequences as features [14–16]. Through this kind of analysis, they aim to compare the data of the most frequent words used by users suffering from a mental disorder and healthy users [17]. The problem with this approach is that the resulting vocabularies from both types of users tend to show a high overlap [18, 19].

Other works have applied sentiment analysis techniques to study the emotional properties of the users' posts [20, 21]. They mainly model the positive, negative, and neutral sentiments that the users express, and explore the relationship between these sentiments and the signs of a mental disorder. These works, as well as a psychological theory that relates the manifestation of feelings and emotions with depression [22], inspired the use of emotions to identify depression [8, 23]. In spite of the interesting results of these analyses, they usually fail in detecting users without a mental disorder who tend to express themselves negatively [24, 25].

Other group of works have used a LIWC-based representation [26], which consists of a set of psychological categories that aim to represent users' posts by features of

social relationships, thinking styles, and individual differences [27]. This strategy clearly allows for a better analysis of the users suffering from a mental disorder, nevertheless, its results are only moderately better than those from word-based approaches [11].

Recently, some works have considered the use of ensemble approaches, which combine the previously mentioned representations with different deep neural models [28]. For example, in this work the authors combine the frequencies of words, user-level linguistic metadata, and neural models with word embeddings; it obtained the best-reported result in the eRisk-2018 shared task on depression detection [29]. On the other hand, [30] shows a neural network architecture consisting of eight different sub-models, followed by a fusion mechanism that concatenates the features and predicts if a user presents signs of anorexia. In that work, the authors concluded that the combination of different models obtained better performance than using them separately, suggesting that the different types of features enrich the users' representation and provides relevant information for the detection of anorexia. In a slightly different direction, [31] models the temporal mood variation using an attention network, and [32] applies an attention model combined with sentiment and topic analysis to detect suicidal ideation. These last two studies show the potential of the attention mechanisms in these types of tasks, as their results outperformed those of other deep neural models.

Despite their good performance, an important limitation of the ensemble approaches is the interpretability of the results, and even more so if the final objective is to create a tool aimed to support health professionals. In this regard, [33, 34] they proposed strategies to face this issue. They mainly proposed methods for visualizing the data and characterizing the users affected by mental disorders.

Based on the good performance shown by the ensemble approaches in the detection of anorexia and depression, and motivated by the design of models that could be easily understood, we decided to implement our multi-channel approach as a new and simple way to combine different views of the information shared by social media users.

3 Word representations at the different channels

In order to analyze the information shared by social media users from various views, we propose to represent each word from their posts by using *three different embedding vectors*, aiming to emphasize their thematic, emotional, and writing style contexts, respectively. The following subsections describe each of these vectors.

3.1 Thematic channel

For this channel, since our aim is to capture the semantic or thematic information that is related to each word, we use traditional non-contextual as well as contextual word embeddings. For the first, we employ vanilla GloVe embeddings [9], whereas, for the second, we use the embeddings from BERT [10]. GloVe model is composed of two approaches: a matrix factorization and shallow windows. The main idea of GloVe is to learn word representations in a local context. This model was pre-trained with high-dimensional corpora from Twitter, Common Crawl, and Wikipedia. GloVe presents a lower dimensionality in their word vectors, ranging between 50 and 300 dimensions. GloVe pre-trained with information from Twitter performs well in comparison with other models in tasks of classifying texts from social media networks [9]. On the other hand, BERT is a recent work published by researchers at Google AI Language [10] and stands for Bidirectional Encoder Representations from Transformers. The main innovation from BERT is the bidirectional training of the Transformer's encoder to a language model. This is the main difference from previous works that looked at a text sequence from left-to-right and/or right-to-left during the training. With this technique, the language model has a deeper understanding of language context in comparison with previous models.

For our experiments, we use both types of embeddings separately and evaluate which one contributes the most in the multi-channel representation.

3.2 Emotion channel

In this channel, the representation of the words is done by considering their emotional context. Our intuition is that users with a mental disorder express different and stronger emotions than users without that mental disorder.

In particular, we use the emotion-based word embeddings that are proposed in [8]. In short, to construct these vectors, first, we generate groups of fine-grained emotions for each general emotion that belong to the EmoLEX lexicon [35]. We achieve this by representing each word of the lexicon with its FastText embeddings [36] and then apply a clustering algorithm on them. With the generated groups, we are able to capture different specific topics related to the same emotion. Take, for example, the **Surprise** emotion category, for which we have one group of words that expresses surprise related to art and museums, whereas other groups have words that are related to accidents and disasters. Second, once the groups were generated, we represent each one by the average of the embedding vectors of its words. Each of the resulting

vectors corresponds to a different fine-grained emotion. This whole process creates all computed sub-emotions, where words with similar contexts tend to group together. Finally, we determine the closest fine-grained emotion to each word in the users' posts, and associate them to their respective vectors. This way, the emotion-based word embeddings we are using are represented by fine-grained emotion vectors. Fig. 1 illustrates this whole process to obtain the emotion-based word embeddings.

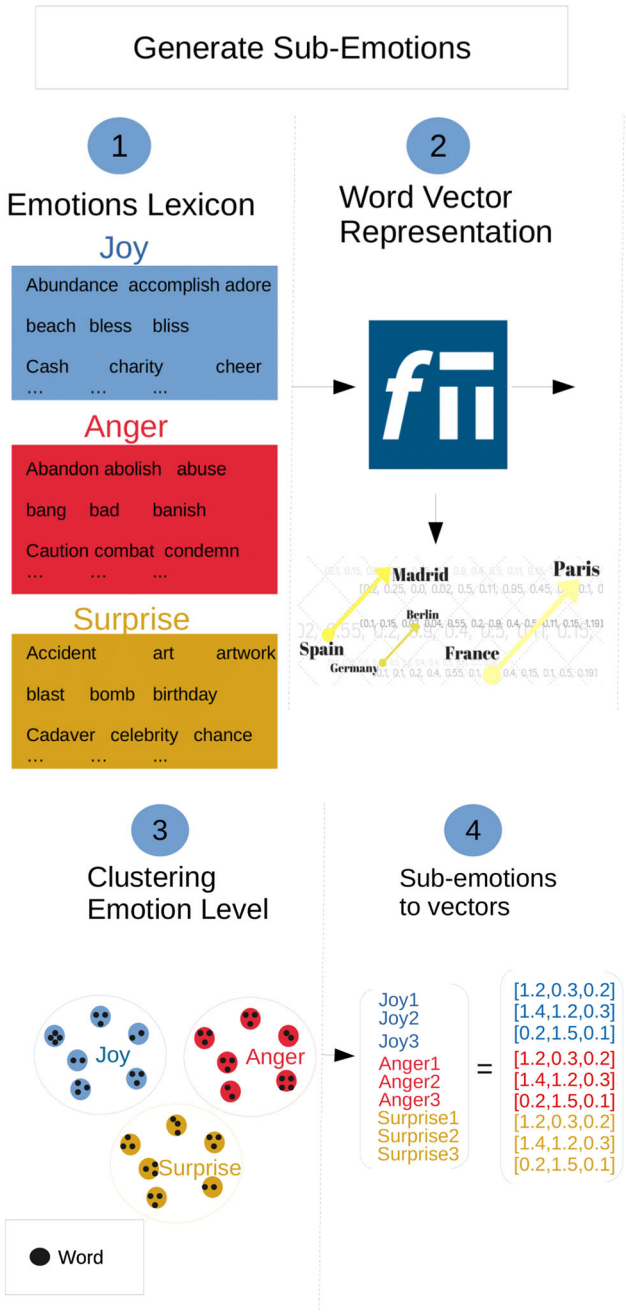

Fig. 1 Diagram of the generation of emotion embeddings

3.3 Style channel

In this channel, the representation of the words aims to capture some aspects of the writing style of social media users. The intuition behind capturing the style is that users with a mental disorder tend to talk more often and differently about the events in the past or the uncertainties in the future than healthy users. In order to capture this kind of information, we devise a new word representation inspired in the FastText vectors [36], where the idea is to weigh the contribution of each char n-gram according to its discriminative value as measured by the chi^2 distribution. Fig. 2 depicts the whole process, which consists of two main modules.

The first module uses the corpus of the task at hand to compute the relevance of each sub-word. To this end, it first divides the users' posts in all their char 3-grams, and then it computes their chi^2 distribution according to the two given classes (depressed and healthy users). For each n-gram (term), we obtain a corresponding chi^2 score that indicates if the document class has influence over the n-gram's frequency. With this approach, we want to capture the most important character n-grams and use them to weight the words.

On the other hand, the second module builds the word embeddings combining the previously extracted char n-grams. It selects each word from the users' posts and divides it into char 3-grams. Then, for each 3-gram, it computes its embedding using FastText. Finally, it obtains the embedding vector of the word by applying a weighted sum of the vectors of its char 3-grams, considering as weights their chi^2 values. We can express this formally as:

$$\mathbf{S_w} = \sum_{i=1}^{n} \mathbf{c_i} \cdot \chi_i^2 \tag{1}$$

where:

- $\mathbf{S_w}$ is the final style vector for each word w.
- $\mathbf{c_i}$ represents the vector of each n-gram.
- χ_i^2 is the chi^2 value of each n-gram.

Take, for example, the word "depression", its style-based embedding is obtained by the weighted sum of the vectors corresponding to its character 3-grams "dep", "epr", ..., "ion".

It is important to notice that the style embeddings are similar for words that have similar spelling rather than similar meaning. For example, for the word "mental" some of their closest words are "dental", "mentality" and "decremental". For a more detailed discussion of how the style embeddings differentiate from the original semantically oriented embeddings, in Fig. 3 we show the similarity of eight different pairs of words related to mental health. We can observe that style vectors find high cosine

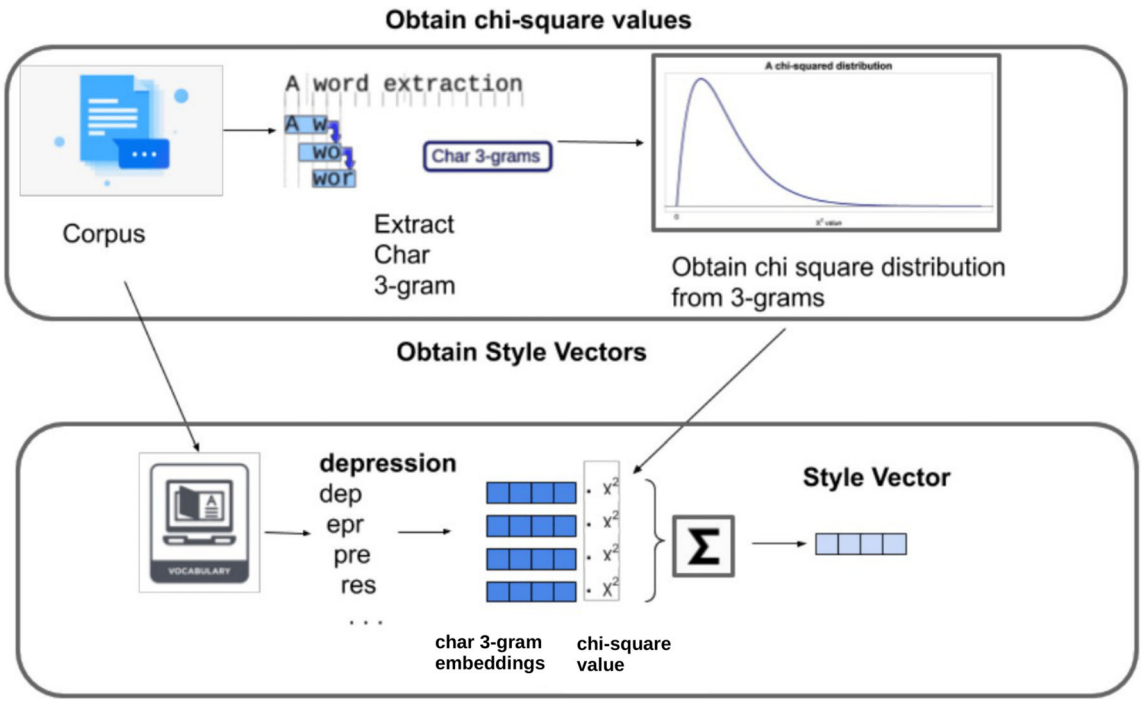

Fig. 2 Diagram of the generation of style embeddings. The first step is to obtain the chi^2 values, then weight the vectors

Depressed - Depressedly original - 0.7649 style - 0.9999	**Depression - Anti-depressant** original - 0.5362 style - 0.9890
Cried - Died original - 0.3750 style - 0.6351	**Tried - Trying** original - 0.7474 style - 0.3269
Illness - Dullness original - 0.4315 style - 0.9997	**Happiness - Sadness** original - 0.6086 style - 0.7202
Eating - Vomiting original - 0.5381 style - 0.8734	**Trying - Crying** original - 0.3954 style - 0.9709

Fig. 3 Similarities of several word pairs using the style and the original embeddings. Style embeddings are similar for words that have similar spelling. For example, words in superlative, or regular verbs in past tense, or words with the same root

similarity between word sharing affixes. Take, for example, the words: "Tried" and "Trying", in the original embeddings they have high similarity since both come from the verb "try", but they have lower similarity in the style embeddings due to the different verb tense. Following this idea, if we analyze the words "Cried" and "Died", these words have high similarity for the style embeddings.

3.4 What does each individual channel capture?

A reasonable question would ponder how different are the word embeddings for the different channels. To offer a glimpse of this, first, we selected some of the words with the highest information gain. Then, we computed their word embeddings for the three channels. Finally, we obtained their closest words using the cosine similarity. Table 1 presents the obtained results. For each query word, we can observe that the three channels offer very different information, that could later be used to improve the detection of users suffering from a mental disorder. For example, for the word "depressed", the thematic channel captures topics related to insecurities or concerns, whereas the style channel retrieves some word variations such as depressing and depressants, and the emotion channel includes some negative adjectives as unhappy and demotivated.

4 The multi-channel classification model

The multi-channel learning paradigm aims to create a new data representation by combining two or more channels of information. Figure 4 shows our multi-channel architecture, which includes two main modules. On the one hand, a Convolutional Neural Network (CNN)[1] for feature extraction. It aligns very well with our hypothesis that in order to identify a user suffering from a mental disorder, it is enough to detect a set of thematic, stylistic or emotional evidences distributed throughout all of the posts. On the other hand, to combine the information we used a new neural network named Gated Multimodal Unit (GMU)

Table 1 Examples of the closest words using cosine similarity for five query words according to the three considered channels

Query word	Thematic	Style	Emotion
Mental	Psychiatric	Dental	Autistics
	Psychological	Mentality	Psychosocial
	Retardation	Mentalism	Psychopathy
	Illness	Mentality	Behavioral
	Disorders	Decremental	Sociological
Depressed	Weak	Depressedly	Unhappy
	Distressed	Depressants	Demotivated
	Insecure	Anti-depressive	Uninterested
	Worried	Depressing	Fatigued
	Disturbed	Depressive	Troubled
Therapy	Treatments	Aromatherapy	Therapeutic
	Therapies	Hydrotherapy	Medicine
	Psychotherapy	Radiotherapy	Hydrotherapy
	Chemotherapy	Immuno-therapy	Clinical
	Medication	Physical-therapy	Pharmacy
Medications	Medicines	Dedications	Anesthesia
	Prescription	Adjudications	Ibuprofen
	Drugs	Meditations	Paracetamol
	Antidepressants	Dedication	Analgesia
	Pills	Education	Promethazine
Compulsions	Obsessions	Envisions	Obsessive
	Phobias	Fusions	Compulsive-disorders
	Compulsion	Inclusions	Obsessive
	Preoccupations	Visions	Paranoid
	Fascinations	Emissions	Personality-disorders

[37], a module that produces a weighted combination of the extracted features from the three channels (we explain the GMU module in detail in the following sub-section). The whole process can be summarized as follows:

1. Represent each word in the user' posts with an embedding vector. This embedding vector corresponds to the channel that is being analyzed.
2. Use a CNN for feature extraction, as described in [38] to extract relevant unigram and bigrams, we employ filters of size equal to 1 and 2, which are represented in Fig. 4 in red and blue respectively.
3. Obtain for each channel different feature maps of each region and concatenate them together to form a single feature vector. This can be interpreted as summarizing the local information to find patterns.
4. Use a GMU module and linear layers to learn the relations between each channel feature vector. Then, apply a sigmoid activation function to classify the final vector with the information of the three channels.

[1] We also evaluated a Recurrent Neural Network (RNN) with an attention mechanism to learn the relation between the channels, but experiments showed a better performance for the CNN alone. We discuss more of this in the analysis of the results section.

The following section details the GMU module, which is the key element for the dynamic combination, defined at user level, of the three information channels considered.

4.1 Gated Multimodal Unit (GMU)

A simple and common idea to combine various types of information (i.e., channels) is to concatenate or add their respective representations into one single vector. Nevertheless, using such strategies assume that all channels have the same relevance, which is usually not the case. In our study, depending on the mental disorder studied, and on the particular user to be analyzed, one or more channels might have different and complementary information.

In a recent work [37], the authors proposed a novel type of hidden unit called Gated Multimodal Unit (GMU). This unit works similarly to the control flow mechanism in gated recurrent units. The gates in the unit let the model regulate the flow of information. The main idea in the GMU is that the unit learns to weight the modalities (channels for us) and fuse them according to their relevance. A GMU works similar to a neural network layer and finds an intermediate representation based on the different modalities.

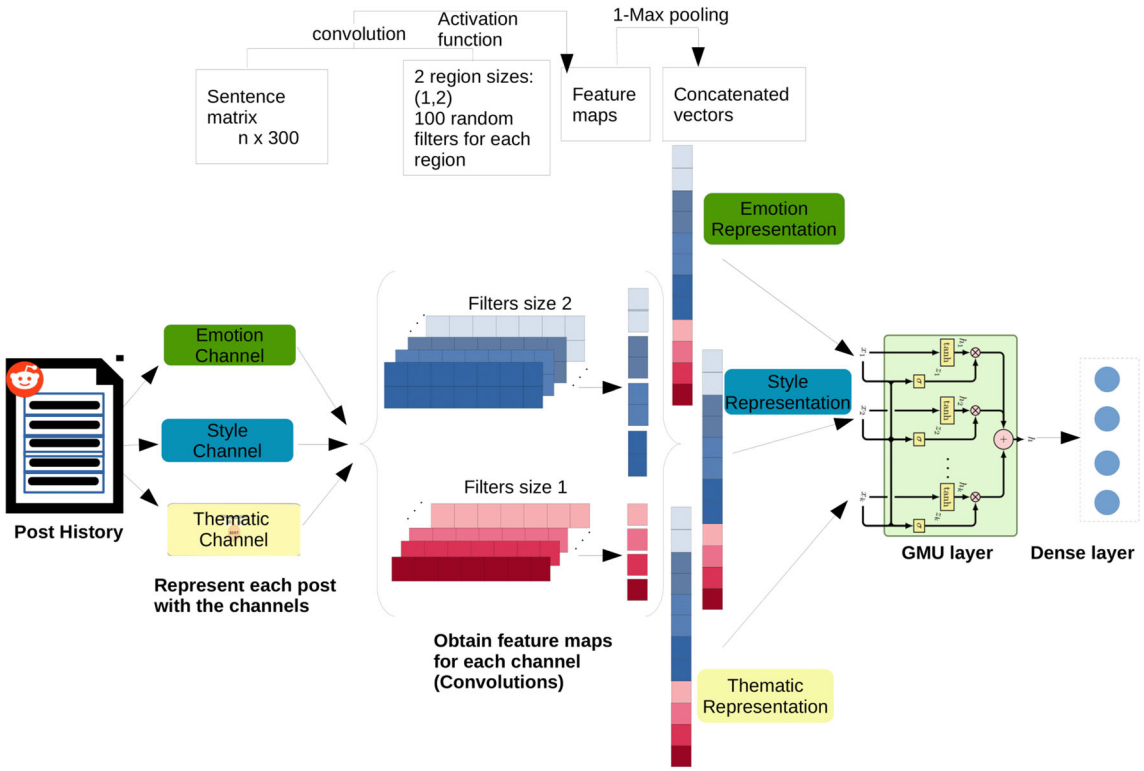

Fig. 4 Diagram of the Convolutional Neural Network Model for the creation of the multi-channel representation. First, we represent each post with the information of each channel. Then, use 100 random filters of size 1 and 2 to extract the local features. Third, use a GMU module and linear layers to learn the relations between each channel feature vector. Finally, apply a sigmoid activation function to classify the final vector

Figure 5 presents a general overview of the GMU module we used, where the x_i inputs represent the feature vectors associated with each modality, and the z_i weights indicate their relevance. Each feature vector enters a neuron with a tanh activation function, and encodes an internal representation based on the modality used. We also have a gate neuron represented by σ, and controls the contribution of each feature from the overall output of the unit. At the end, a final fused representation is obtained, which corresponds to the weighted sum of each modality. Because modality weights are computing at instance level (for each user, in our case), one of the advantages of the GMU is its interpretability; they clearly provide a better understanding of the contribution of the different modalities to the predictions. After training the model, we can visualize the weights z_i and have a better understanding of which modalities had more contribution to the prediction.

In the figure, we can appreciate that x_i is a feature vector associated with a modality i. For each vector, there will be a weight z_i, which controls the contribution of that modality. At the end of the unit, we obtain a final fused representation as to the weighted sum of each modality. These gates will allow the model to decide how each modality affects the unit's output. One of the advantages of

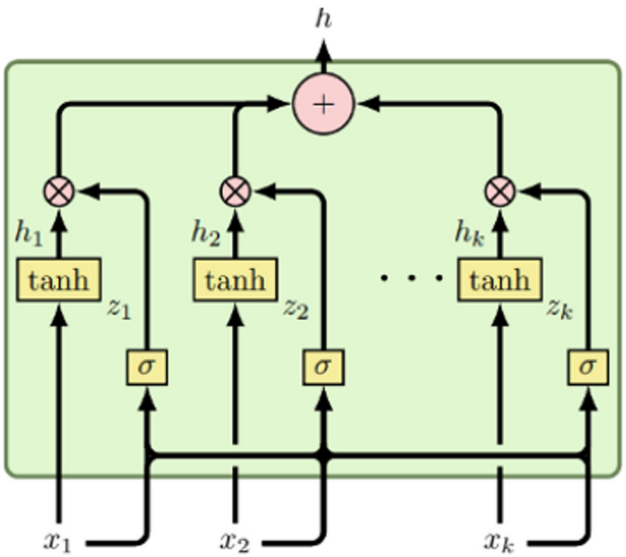

Fig. 5 Overview of GMU module. Where x_i represents the $i-th$ input modality. The final fused representation of all modalities is represented by h at the top (37)

the GMU is its interpretability. After training the model, we can visualize the weights z_i and have a better understanding of which modalities had more contribution to the prediction.

5 Experiments and results

5.1 Data sets

To thoroughly evaluate our proposed approach we use the data sets from the eRisk 2019 and 2020 evaluation tasks [39, 40], which are for the detection of anorexia, depression, and self-harm. These data sets contain the post history of several users from the Reddit platform, an American social network, where users submit content such as text, images, and videos. Posts are organized by subject into boards called subreddits. For each task, the authors explore subreddits related to mental health, select users, and create two categories: 1) positive users, those affected by either anorexia, depression, or self-harm; and 2) the control group, composed of people who do not suffer from any of these mental disorders.

For the depression task, each user completed out a standard Beck's Depression Inventory (BDI) questionnaire [41], which contains 21 questions that allow to assess the level of severity of the depression. In the original task, the organizers asked the participants to predict, for each user, the possible answers to each input of the questionnaire. In contrast, for this study, we exclusively consider a binary prediction task, i.e., to distinguish between positive and control users. In particular, the positive class is composed of users that obtained 21 points or more in the final result of the questionnaire (presence of moderate or severe depression), whereas the control class is formed by the rest of the users, having 20 points or less in their final result. For anorexia and self-harm, the positive class is composed of people who explicitly mentioned that they were diagnosed by a medical specialist with anorexia or that they had committed self-harm. The creators of these data sets mentioned that they discarded users using vague expressions like "I think I have anorexia" during the gathering of data. The control class for both tasks is composed of random users from the Reddit platform. Control group also contain users who often interact in the anorexia, depression, or self-harm threads to add more realism to the data and make the detection of positive users more challenging and closer to reality.

Table 2 shows how classes distribute within these data sets as well as some general information regarding the collections. For the depression task, we used for training the data set from eRisk 2018 [29].

Table 2 Data sets used for experimentation, where P indicates the positive users and C is used for control users

Data set	Train		Test	
	P	C	P	C
Anorexia'19	61	411	73	742
Avg. num. posts	407.8	556.9	241.4	745.1
Avg num. words (post)	37.3	20.9	37.2	21.7
Avg. days	800	650	510	930
Depression'20	214	1493	40	49
Avg. num. posts	440.9	660.8	493.0	543.7
Avg num. words (post)	27.5	22.75	39.2	45.6
Avg. days	686	663	642	1015
Self-harm'20	41	299	104	319
Avg. num. posts	169.0	546.8	112.4	285.6
Avg num. words (post)	24.8	18.8	21.4	11.9
Avg. days	495	500	270	426

5.2 Pre-processing

Users' posts tend to contain a lot of noisy text and irrelevant information for the detection of a mental disorder. Thus, the application of pre-processing techniques is required to allow the classifier to focus on the key information and to be able to obtain reliable results. The first pre-processing step was to normalize the text by lowercasing all words and removing special characters like URLs, emoticons, and #; the stopwords were kept. The next step was to select for each task the words with the highest chi^2 value and remove the rest of the words. For this step, we explored a different numbers of features, it will be described in the evaluation section. It is worth mentioning that when using the GMU module, full texts produced the best results. That was not the case for simpler fusion strategies.

5.3 Experimental settings

Classification & predictions: We separate each post history into N parts. We select the N value empirically, testing recommended sizes of sequences in the literature, i.e., $N = \{25, 35, 50, 100\}$. For training, we process each part of the post history as an individual input and train the model. For the test, each part receives a label of 1 or 0; then, if the majority of the posts are positive, the user is classified as showing a mental disorder. The main idea is to consistently detect the presence of major signs of anorexia, depression, or self-harm through all the user posts.

To shed some light on what is being detected by the approach proposed, Fig. 6 shows the distribution of the decisions in the post history of some users. Taking as a reference the answers to the BDI questionnaire used for depression detection, we selected the users with the three highest scores, the three lowest, and three users with borderline scores to be depressive, which correspond to severe

depression, normal status, and borderline clinical depression, respectively. Then, we represent each post of their history with a dark blue color if the post obtains a high probability of belonging to a depressive user, and with a white color if the probability was closer to a control user. We can appreciate how the post history of the users with a high score of depression is darker than the users with a low score. It is interesting to notice that the borderline users present different distribution on their post history. This difference could be due to the diversity in the topics they write or express in their posts.

Baselines. As main baseline, we employed a traditional Bag-of-Words representation, where it described the text by the occurrence of words within a document, considering words as well as word n-grams (sequences of words). We also consider a bag of character trigrams, a common approach for style analysis. For these two approaches, we selected the features using a tf-idf weighting and the chi^2 distribution X_k^2. In addition, we consider some baselines based on deep learning approaches, using a CNN, a Bi-LSTM, and a state-of-the-art approach for text classification based on a Bi-LSTM with an attention layer. All of these neural networks used 100 neurons, an ADAM optimizer, and word2vec and Glove embeddings with a dimension of 300. For the CNN, we used 100 random filters of sizes 1, 2, and 3. We also added a BERT model with a fine-tuning over the training data set. Additionally, the obtained results are compared against the top-three participants of the eRisk evaluation tasks. For all these comparisons, we considered the F_1 score over the positive class, which was suggested as the golden standard by the organizers of eRisk [29].

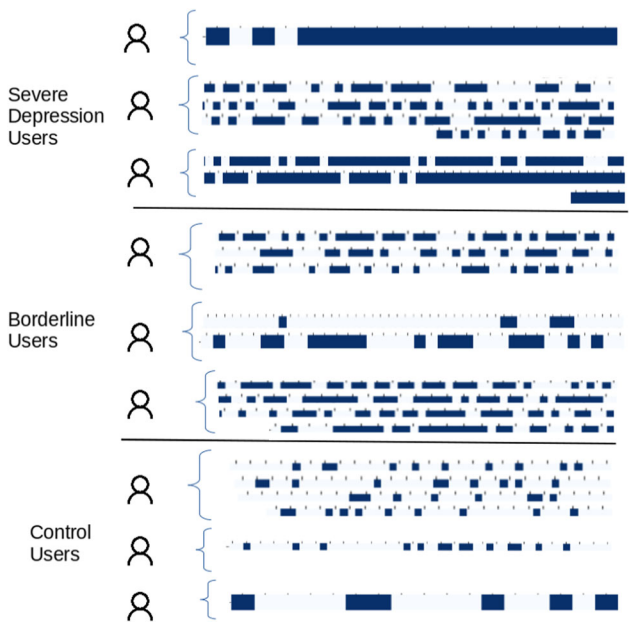

Fig. 6 Output of the different posts of being a positive class. A dark square is used to mark a post as positive and gray as negative. We can appreciate how the post history of the users with a high score of depression is darker than the users with a low score

5.4 Evaluation

Table 3 presents the results in terms of F_1 score over the positive class to detect Anorexia (eRisk'19), Depression (eRisk'20) and Self-harm (eRisk'20). We organize the results in three groups: baseline methods, our proposal but limited to only one channel, and our original proposal using all information channels.

From this evaluation, we observed that most of our proposals outperformed the baseline results. First, some single-channel representations obtain a considerable improvement in comparison with baselines, in particular those based on style and emotion information. Surprisingly, the performance of deep learning models applied over word-based representations is somehow poor and closer to traditional approaches like BoW; we presume this could be attributable to the small size of the data sets in conjunction with their large thematic diversity. The full-channel representation is clearly the performant approach in this comparison, then suggesting the pertinence of combining different types of information. Interestingly we noticed that CNN networks obtain better performance than RNN networks. The latter could be due to the fact that CNN networks search for the presence of specific local information important for the detection of these disorders. In addition, using the GMU module improved the results obtained by the simple concatenation strategy in the tasks of anorexia and depression detection, but not in the self-harm detection task, where it only obtained competitive results.

From the first round of experiments, we highlight the following observations:

1. Most single-channel representations outperformed the baselines, especially noting that style and emotional information are more relevant for the detection of mental disorders in online environments than the thematic aspect without the contextual information.
2. The use of a multi-channel representation improves the results than only using one type of information. This result shows that learning the fusion is very relevant to capture signs of mental disorders in users.
3. Using a GMU improves the results of anorexia and depression detection in comparison with a simple vector concatenation strategy.

Table 3 F1 results over the positive class in three eRisk's tasks

Method	Anorexia	Depression	Self-harm
Baselines			
BoW-unigrams	0.67	0.58	0.50
BoW-Ngrams	0.66	0.57	0.50
Bag of char 3grams	0.67	0.58	0.53
RNN-word2vec	0.65	0.57	0.55
CNN-word2vec	0.66	0.60	0.56
RNN-Attention	0.66	0.50	0.58
BERT	0.77	0.62	0.67
Our methods: Single-channel			
RNN $Thematic_G$	0.65	0.58	0.57
CNN $Thematic_G$	0.67	0.61	0.57
CNN $Thematic_B$	0.77	0.64	0.60
RNN Style	0.73	0.63	0.64
CNN Style	0.70	0.64	0.65
RNN Emotion	0.69	0.62	0.64
CNN Emotion	0.76	0.59	0.65
Our methods: Multi-channel (simple concatenation)			
RNN St+Em (Entry)	0.75	0.65	0.66
RNN St+Em (BA)	0.76	0.48	0.56
RNN St+Em (AA)	0.76	0.55	0.56
CNN St+Em	0.78	0.65	0.72
CNN St+Em+Th_G	0.78	0.67	0.71
CNN+RNN St+Em	0.71	0.64	0.71
C+R St+Em+Th_G	0.72	0.66	0.70
CNN St+Em+Th_B	0.78	0.66	**0.75**
Our methods: Multi-channel (CNN + GMU combination)			
GMU St+Em+Th_G	0.79	0.67	0.73
GMU St+Em+Th_B	**0.82**	**0.70**	0.73

St = style channel, Em = emotional channel and Th = Thematic channel. For the RNN notation: Entry = combination of channels at the input layer; BA = combination of the channels before the attention layer, and AA = combination of the channels after the attention layer

5.4.1 Comparison against the eRisk participants

To add a context regarding this shared task, consider that a total of 54 models were submitted to the anorexia detection task and 57 to the self-harm detection task in eRisk-19 and 20 editions [39, 40]. It is important to mention that the participants focused on obtaining early and accurate predictions of the users, while our approach focuses exclusively on determining accurate classifications.

Table 4 shows how our best approach (i.e., the CNN model with 2 and 3 channels) compares against the top places at the eRisk 2019 and 2020 evaluation tasks. We observe that our approach achieves competitive results in both tasks, first place for Anorexia and tied in first place for Self-harm. For the depression task, organizers changed the evaluation task, and thus, we cannot directly compare our results against the participants[2].

Table 4 F_1, Precision and Recall results over the positive class

Task	Anorexia 2019			Self-harm 2020		
Metric	F1	P	R	F1	P	R
1st place	0.71	0.64	**0.79**	**0.75**	**0.82**	0.69
2nd place	0.68	0.77	0.60	0.62	0.62	0.62
3rd place	0.68	0.67	0.68	0.62	0.59	0.65
Our best	**0.82**	**0.88**	0.76	**0.75**	0.69	**0.82**

Bold indicates for each metric the highest result

For further analysis of these results, Fig. 7 presents a boxplot of the F_1, precision, and recall scores of all participants from both tasks. The green **X** represents our best result of the combination of channels. In the figure, we appreciate that our results are in the highest quartile for both tasks. These results indicate that our multi-channel representation obtains competitive results in comparison with the participants of the anorexia and self-harm detection tasks.

5.5 Analysis of results

5.5.1 Contribution of each information channel

One of the most important aspects of our proposal is to understand the way the weighting mechanism is dynamically learning the relevant information. The GMU units are one of the key elements to weight and highlight each channel. In order to show this, we focus on analyzing the gates (z_i) that determine how relevant is each modality. For this purpose, we fed the module using the posts in the test set and average the gate outputs per channel. Note that the z_i value represents the contribution of the feature calculated from x_i to the overall output of the unit. Fig. 8 shows the results for the three channels, where each row already takes into account the average of all posts per mental disorder.

It is worth noting how the activations for each channel are different depending on the mental disorder. For example, for depression and anorexia tasks, the thematic channel (BERT) has the highest value, however for self-harm the style channel has the highest value. It is also interesting to mention that, the thematic channel is the one with the highest variation. For example, the highest value is in anorexia and the lowest value is in self-harm. This variation indicates that the posts of users who suffer from anorexia are presumably more homogeneous than those who suffer from self-harm.

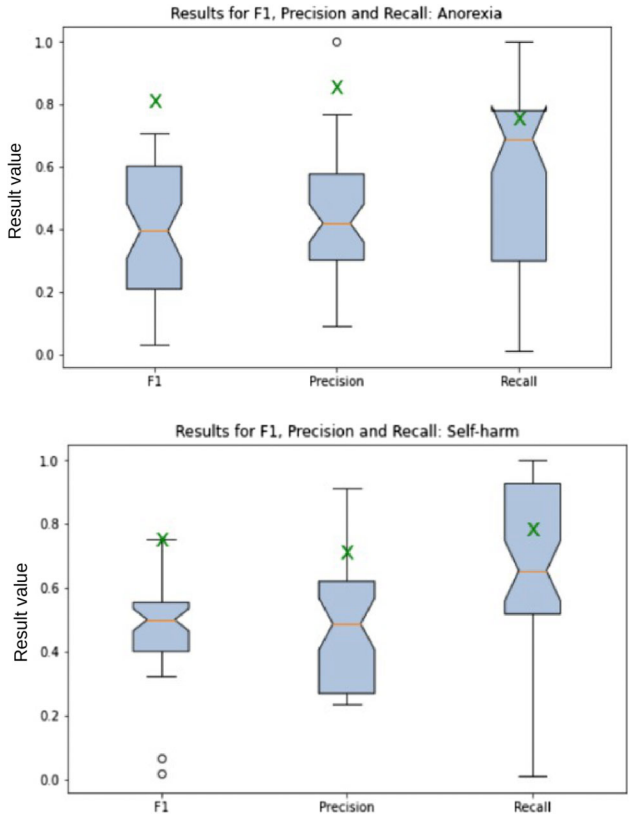

Fig. 7 Boxplot of the F1 scores for anorexia (upper part) and self-harm (bottom part), where the green X represents our best approach

One additional analysis of the GMU activations is shown in Table 5. In this table, we show the posts with the highest z_i value for each mental disorder and each channel. We can appreciate that even when the topics are not directly related to mental disorders, the posts are related to personal opinions and concerns.

5.5.2 Qualitative analysis of each channel

This analysis aims to investigate to what extent each information channel captures different information. For visualizing this, we used a strategy inspired by the back-propagation in vision. This strategy named as saliency, measures how much each input contributes to the final decision, which is obtained by using its first derivative. More formally, we analyze the output of our model, and computed the saliency as:

$$\mathcal{S}_j = \sum_{x_l \in \mathbf{x}_j} \left\| \frac{\partial \tilde{y}_i}{\partial x_l} \right\| \tag{2}$$

of the three channels with different sample texts that were extracted from users with a mental disorder. We define the

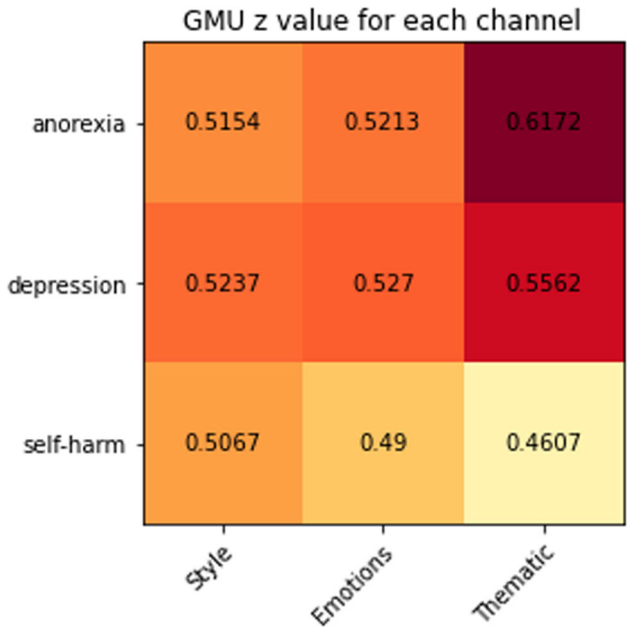

Fig. 8 Average proportion of GMU unit activations for the channels over the test set. The Figure presents the average z_i value for each channel and mental disorder

saliency $\mathcal{S}_j$ of a specific word as the average of the magnitude of the gradient of each component in the embedded representation [42]. We present these saliency maps in Fig. 9, where the red color indicates a high value and the bluer color represents a lower value. We notice that depending on the channel and the context, the saliency is higher for different words. For example, the word "healthier" has high saliency for the emotion channel, but it is low for the style channel. For the word "thoughts", the saliency is high for the emotion and style channel but is low for the thematic channel. See how the word "confidence" gets high saliency for the three channels (at different level of importance), but the rest of the words in the context have different saliency.

To further analyze the saliency, we compute for each task the average saliency for each word. Then, we select the words with the highest value, avoiding words with less than ten occurrences in the documents (we want to avoid words that appear few times but obtain high saliency and do not generalize the task). In Table 6 we show these words (see next page). Note that for each channel, the words are different, but they have a close relationship with each task. For example, for self-harm, the style words are disorders, addicted or tension, while for the emotion channel the highest words are killed, bother, or cutting. We can conclude that the channels contribute individual and

[2] To clarify this point, our approach focuses on a binary classification task, i.e., to discriminate between users suffering from depression and control users, while, on the other hand, the eRisk task considered the assessment of the level of depression severity for each user.

Table 5 Posts with highest z_i value for each mental disorder and channel

Channel	Post
Anorexia	
Style	*"...produced in a manner to sell more rather than staying true to what the lore is..."*
Emotion	*"... actually made to portray stories that are in touch with the issues that the current world is going through however when such changes are happening so often they do not feel that they fit in..."*
Thematic	*"...they fit in with they are trying to achieve because of this. Is trying to push social justice..."*
Depression	
Style	*"I use to try and humanize myself. In their eyes is fake nightmares and say that thinking of talking to them stopped the nightmare..."*
Emotion	*"I think I know who they're playing, but who knows. Sent to losers by eliminated."*
Thematic	*"... civil war losers, top gg. its all gaming after this, losers..."*
Self-harm	
Style	*"... thank you I appreciate it very much and I hope to be able to serve to the best of my ability no matter what others think..."*
Emotion	*"...me Im usually fine, Im a christian and love that show supernatural too also lucifer. What the hell guys you re giving me a bad name..."*
Thematic	*"...she yelled out loud, it scared me so bad that the knife slipped and I ended up slicing into my thumb."*

The posts are related to personal opinions and concerns

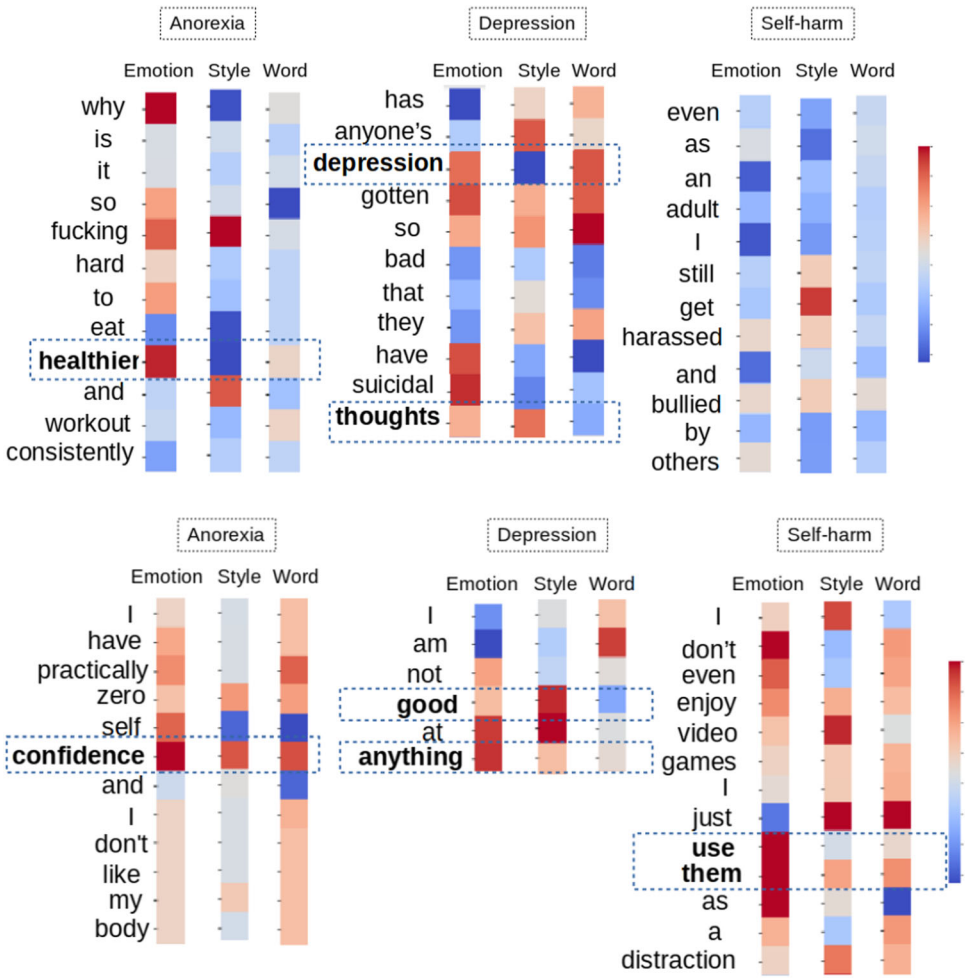

Fig. 9 Saliency obtained with the different type of channels for the positive class. The red color indicates a high value and the bluer color represents a lower value

Table 6 Words with the highest saliency for each task and each channel

Task	Thematic	Style	Emotion
Anorexia	Mouth	Stereotype	Frustrating
	Dicyclomine	Stigmatized	Likeness
	Young	Promise	Mean
	Stomach	Lunch	Jealous
	Meal	Anonymous	Far
Depression	Pills	Mysteries	Frustrating
	Need	Borderline	Likeness
	Together	Professional	Pathological
	Suicide	Ditsy	Mean
	Depressive	Stigmatized	Life-call
Self-harm	Disinterest	Addicted	Killed
	Shyness	Tensions	Bother
	Accused	Disorders	Codependent
	Dies	Homework	Cutting
	Friendzone	Crime	Boyfriend

The words are different, but they have a close relationship with each task

contrasting information between each other, and this information helps us improve the detection of mental disorders in online environments.

5.5.3 On the predicted posts' probabilities

As we previously mentioned, decisions about users are generated by combining the predictions made for each post. To better understand this process, Fig. 10 presents the distributions of the posts' prediction values for the three tasks considered. In this case, the prediction values are nothing other than the probabilities of the posts belonging to the positive class in accordance to the classifier used.

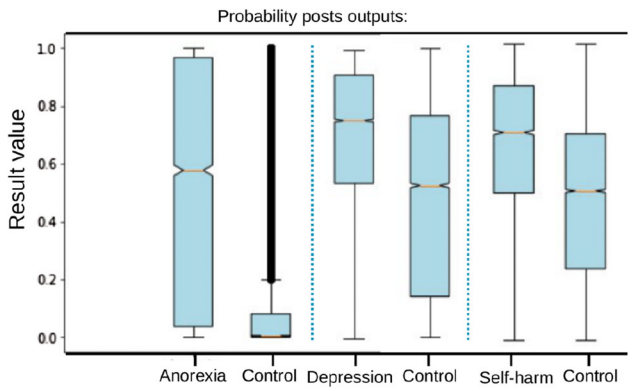

Fig. 10 Distribution of the posts' predictions. The prediction values are the probabilities of the posts to belong to the positive class

Figure 10 shows some interesting information. For the self-harm and depression tasks it is possible to observe that most of the prediction values for the positive users' posts are higher than 0.6, thus suggesting the suitability of our approach to detect evidence of the presence of those mental disorders. Nevertheless, control users' posts also show some high probabilities, which may indicate that their topics, emotions and style overlap to some degree with those from the positive users. On the other hand, for the anorexia task it can be observed that control users are clearly distinguishable from positive users, since most of their posts have little or no probability of belonging to the positive class. However, in this case not all posts from positive users show high probabilities, perhaps due to their greater thematic and style diversity. In summary, the figure shows that for the detection of depression and self-harm false positives are the main concern, while for the detection of anorexia false negatives are the key issue.

5.5.4 Complementary analysis of the channels

In closing this analysis, we investigate how diverse yet complementary these channels are in terms of the information they capture.

To measure their complementarity, we used the Maximum Possible F1 (MPF) metric. This measure is a variation of the Maximum Possible Accuracy (MPA), which is defined as the quotient of the correctly classified instances over the total number of test instances. For this analysis, we considered an instance as correctly if at least one of the channels classified it correctly.

Table 7 presents the MPF scores for each task, measured over the positive class. For the MPF values, we can appreciate an improvement in comparison with our best reported results (last row of table). These results indicate that the channels are complementary to each other. Analyzing the obtained insights results, it is clear that there is still room for improvement, but to achieve it, it will be necessary to explore more channels as well as other fusion strategies.

Table 7 MPF results in the three tasks, measured over the positive class

Metric	Anorexia	Depression	Self-harm
MPF	0.8872	0.7568	0.8442
Our best F1	0.82	0.70	0.75

5.6 Limitations and ethical concerns

This study presents some limitations, mainly because these data sets are observational studies and we do not have access to the personal and medical information that is often considered in risk assessment studies. There are also some limitations given to the nature of the data, as it may differ from users at risk who do not have an online account or decided to not make their profiles public. In addition, in the data sets of anorexia and self-harm, it is not guaranteed that the users annotated as positive are actually at risk because the annotation was performed after reading just a few posts.

We believe that it is important to mention that the analysis of data provided by social networks to detect health problems and assist clinicians is an open issue, not uncontroversial. When we analyze social media content, we respect concerns regarding individual privacy or certain ethical considerations. Given the personal behavior and emotional health of the users, these concerns could appear due to the usage of sensitive information. The experiments we perform for this work and the usage of the data sets are for research and analysis only, and the mishandling or misuse of the information is prohibited.

6 Conclusion and future work

In this study, we explored the detection of Anorexia, Depression, and Self-harm in users of social media environments by means of a novel multi-channel representation. Each information channel focuses on extracting information that correspond to the users' writing style, emotions and thematic interests. Our proposal can automatically show how to combine these features and extract the most relevant information from each channel. Results clearly suggest that combining different types of information helps in the detection of users with mental disorders, outperforms traditional and state-of-the-art baselines, and is strongly competitive with the performance of top eRisk participants. The analysis of our method yields that the complementarity in the types of information is important in getting a better picture and understanding of the posts written by the users. We can also highlight the importance and robustness of our holistic view in critical classification tasks, such as mental disorders detection. The use of a multi-channel representation improves the results than only using one type of information. This result confirms that learning to combine different types of information is very relevant to capturing signs of mental disorders in users. Our analysis yields that the complementarity in the types of information is important to get a better picture and understanding of the posts written by the users.

In the future work, we want to explore more sophisticated combination techniques that could improve the results and understanding of mental disorders detection. Also we noted that most of the analysis of mental disorders has been made for the English language; therefore, we are interested in expanding this study to the Spanish language.

Acknowledgements Aragon thanks for doctoral scholarship CONACyT-Mexico 654803.

Author contributions MEA helped in conceptualization, methodology, investigation, formal analysis, writing–original draft preparation. APLM contributed to methodology, validation, investigation, writing–review and editing. LCGG contributed to supervision, visualization, writing–review and editing. MMG helped in conceptualization, supervision, project administration, resources, writing–review and editing.

Data availability The data that support the findings of this study are available from https://erisk.irlab.org/. Restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of eRisk organizers.

Declarations

Conflict of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

References

1. Mathers C, Loncar D (2006) Projections of global mortality and burden of disease from 2002 to 2030. Public library of science, PLOS Medicine
2. Kessler R, Bromet E, Jonge P, Shahly V, and Marsha (2017) The burden of depressive illness. Public health perspectives on depressive disorders 40–66
3. Guardian News and Media (2019) Mental illness: Is there really a global epidemic? The guardian
4. Renteria-Rodriguez M (2018) Salud mental en mexico. NOTA-INCyTU NÚMERO 007
5. Guntuku SC, Yaden D, Kern M, Ungar L, Eichstaedt J (2017) Detecting depression and mental illness on social media: an integrative review. current opin behavioral Sci 18:43–49
6. Pestian JP, Nasrallah H, Matykiewicz P, Bennett A, and Leenaars AA (2010) Suicide note classification using natural language processing: a content analysislin heidelberg. biomed inform insights
7. Qianli MA, Lifeng S, Enhuan C, Shuai T, Jiabing W, and Garrison C (2017) Walking walking walking: action recognition from action echoes. Twenty-Sixth International Joint Conference on Artificial Intelligence
8. Aragón M., López-Monroy AP, González-Gurrola LC, and Montes-y Gómez M (2019) Detecting depression in social media using fine-grained emotions. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, Vol 1 (Long and Short Papers)
9. Pennington J, Socher R, and Manning C (2014) GloVe: Global vectors for word representation. In: Proceedings of the 2014

conference on empirical methods in natural language processing (EMNLP)
10. Devlin J, Chang M, Lee K, and Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. NAACL
11. De Choudhury M, Gamon M, Counts S, and Horvitz E (2003) Predicting depression via social media. In: Proceedings of the 7th international AAAI conference on weblogs and social media
12. De Choudhury Munmun, Counts Scott, and Horvitz Eric (2013) Social media as a measurement tool of depression in populations. In :Proceedings of the 5th annual ACM web science conference
13. Wang Tao, Brede Markus, Ianni Antonella, and Mentzakis Emmanouil (2017) Detecting and characterizing eating-disorder communities on social media. In: Proceedings of the tenth ACM international conference on web search and data mining
14. Tsugawa S, Kikuchi Y, Kishino F, Nakajima K, Itoh Y, and Ohsaki H (2015) Recognizing depression from twitter activity. In: Proceedings of the 33rd annual ACM conference on human factors in computing systems, pp 3187–3196
15. Schwartz HA, Eichstaedt J, Kern M, Park G, Sap M, Stillwell D, Kosinski M, and Ungar L (2014) 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
16. Liu N, Zhou Z, Xin K, and Ren F (2018) Tua1 at erisk.In: Proceedings of the 9th international conference of the CLEF association, CLEF 2018, Avignon, France
17. Coppersmith G, Harman C, and Dredze M (2014) Measuring post traumatic stress disorder in twitter. In: Proceedings of the Eighth international AAAI conference on weblogs and social media
18. Trifan A, and Oliveira JL (2019) Bioinfo@uavr at erisk 2019: delving into social media texts for the early detection of mental and food disorders. In: Proceedings of the 10th international conference of the CLEF association, CLEF 2019, Lugano, Switzerland
19. Van Rijen P, Teodoro D, Naderi N, Mottin L, Knafou J, Jeffryes M, and Ruch P (2019) A data-driven approach for measuring the severity of the signs of depression using reddit posts. In: Proceedings of the 10th international conference of the CLEF association, CLEF 2019, Lugano, Switzerland
20. Ramírez-Cifuentes D, and Freire A (2018) Upf's participation at the clef erisk 2018: Early risk prediction on the internet. In: Proceedings of the 9th international conference of the CLEF association, CLEF 2018, Avignon, France
21. Preotiuc-Pietro D, Eichstaedt J, Park G, Sap M, Smith L, Tobolsky V, Schwartz HA, and Ungar L (2015) The role of personality, age and gender in tweeting about mental illnesses. In: Proceedings of the 2nd workshop on computational linguistics and clinical psychology
22. American Psychiatric Association (2013) Diagnostic and statistical manual of mental disorders (5th ed.). American psychiatric association
23. Xuetong C, Martin DS, Thomas WJ, and Suzanne E (2018) What about mood swings? identifying depression on twitter with temporal measures of emotions. Companion proceedings of the the web conference 2018, international world wide web conferences steering committee, 1653–1660
24. Coopersmith G, Dredze M, and Harman C (2014) Quantifying mental health signals in twitter. workshop on computational linguistics and clinical psychology
25. Coppersmith G, Ngo K, Leary R, and Wood A (2016) Exploratory analysis of social media prior to a suicide attempt. In: Proceedings of the third workshop on computational linguistics and clinical psychology
26. Tausczik YR, Pennebaker JW (2010) The psychological meaning of words: Liwc and computerized text analysis methods. J Language Soc Psychol 29:24–54
27. Coppersmith G, Dredze M, Harman C, and Hollingshead K (2015) From adhd to sad: analyzing the language of mental health on twitter through self-reported diagnoses. In :Proceedings of the 2nd workshop on computational linguistics and clinical psychology
28. Trotzek M, Koitka S, and Friedrich CM (2018) Word embeddings and linguistic metadata at the clef 2018 tasks for early detection of depression and anorexia. In: Proceedings of the 9th international conference of the CLEF association, CLEF 2018, Avignon, France
29. Losada DE, Crestani F, and Parapar J (2018) Overview of erisk 2018: Early risk prediction on the internet (extended lab overview). In: Proceedings of the 9th international conference of the CLEF association, CLEF 2018, Avignon, France
30. Mohammadi E, Amini H, and Kosseim L (2019) Quick and (maybe not so) easy detection of anorexia in social media posts. Experimental IR Meets Multilinguality, Multimodality, and Interaction. 10th International conference of the CLEF association, CLEF 2019, Lugano, Switzerland
31. Ragheb W, Aze J, Bringay S, and Servajean M (2019) Attentive multi-stage learning for early risk detection of signs of anorexia and self-harm on social media. In: Proceedings of the 10th international conference of the CLEF association, CLEF 2019, Lugano, Switzerland
32. Ji S, Li X, Huang Z, and Cambria E (2020) Suicidal ideation and mental disorder detection with attentive relation networks. arXiv: 2004.07601
33. Ríssola E, and Aliannejadi M, and Crestani F (2020) Beyond modelling: Understanding mental disorders in online social media. Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal
34. Burdisso S, Errecalde M, Montes-y Gómez M (2019) A text classification framework for simple and effective early depression detection over social media streams. Expert Syst Appl 133:182–197
35. Mohammad SM, Turney PD (2013) Crowdsourcing a word-emotion association lexicon. Comput Intell 29:436–465
36. Bojanowski P, Grave E, Joulin A, and Mikolov T (2016) Enriching word vectors with subword information. Transactions of the association for computational linguistics
37. Arevalo J, Solorio T, Montes-y Gómez M, González FA (2019) Gated multimodal networks. Neural Comput Appl 32(14):10209–10228
38. Kim Y (2014) Convolutional neural networks for sentence classification. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)
39. Losada DE, Crestani F, and Parapar J (2019) Overview of erisk 2019: Early risk prediction on the internet. Experimental IR Meets Multilinguality, Multimodality, and Interaction. 10th international conference of the CLEF association, CLEF 2019, Lugano, Switzerland
40. Losada DE, Crestani F, and Parapar J (2020) Overview of eRisk 2020: Early Risk Prediction on the Internet. Experimental IR Meets Multilinguality, Multimodality, and Interaction proceedings of the Eleventh International conference of the CLEF association (CLEF 2020)
41. Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J (1961) An inventory for measuring depression. JAMA Psychiatry 4(6):561–571
42. Li J, Chen X, Hovy EH, and Jurafsky D (2016) Visualizing and understanding neural models in nlp. HLT-NAACL