Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 27-38 recibido 30-03-2022 revisado 20-05-2022 aceptado 23-05-2022

Information fusion for mental disorders detection: multimodal BERT against fusioning multiple BERTs

Fusión de información para detección de transtornos mentales: BERT multimodal contra múltiples BERTs fusionados

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

[1]Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, Mexico
[2]Centro de Investigación en Matemáticas A.C., Guanajuato, Mexico
[3]Universidad Autónoma de Chihuahua, Chihuahua, Mexico
{mearagon,mmontesg}@inaoep.mx, pastor.lopez@cimat.mx, lcgonzalez@uach.mx

Abstract: Given the increasing number of modalities that modern classification problems provide, recently a multimodal BERT transformer (MMBT) was proposed. An interesting opportunity to evaluate the effectiveness of such model is posed by the problem of timely detection of mental disorders of social media users. For this problem, a multi-channel perspective involves extracting from each user post different types of information, such as thematic, emotional and stylistic content. This study evaluates the suitability of tackling this problem by the apparently ad-hoc MMBT, moreover, we further evaluate if regular BERT models could be combined or fused in such a way that could have a chance in a multi-channel arena. For the evaluation, we use recent public data sets for three important mental disorders: Depression, Anorexia, and Self-harm. Results suggest that BERT models can get on their own a data representation that could be later fusioned and boost the classification performance by at least 5% in F1 measure, even surpassing the MMBT.
Keywords: Multichannel information, Transformers, Mental disorders.

Resumen: Dado el creciente número de modalidades que ofrecen los problemas de clasificación modernos, recientemente se ha propuesto un transformer BERT multimodal (MMBT). Una oportunidad interesante para evaluar la eficacia de dicho modelo la plantea el problema de la detección oportuna de los trastornos mentales de usuarios de las redes sociales. Para este problema, una perspectiva multicanal implica extraer de cada post de los usuarios diferentes tipos de información, como su contenido temático, emocional y estilístico. Este estudio evalúa la idoneidad de abordar este problema mediante el aparentemente ad-hoc MMBT, además, evaluamos si los modelos BERT regulares podrían combinarse o fusionarse de tal manera que pudieran tener una oportunidad en un escenario multicanal. Para la evaluación, utilizamos conjuntos de datos públicos recientes para tres importantes trastornos mentales: Depresión, Anorexia y Autolesiones. Los resultados sugieren que los modelos BERT pueden obtener por sí solos una representación de los datos que podría fusionarse posteriormente y aumentar el rendimiento de la clasificación en al menos un 5% en la medida F1, superando incluso al MMBT.
Palabras clave: Multicanal, Transformers, Trastornos Mentales.

1 Introduction

Over the last few years, millions of people around the world have been affected by one or more mental disorders, for example, in 2018 a study of mental disorders in Mexico reveals that 17% of people in the country have at least one mental disorder and one in four will suffer a mental disorder at least

ISSN 1135-5948. DOI 10.26342/2022-69-2

once in their life (Renteria-Rodriguez, 2018). Unfortunately, this phenomenon causes interference in their daily life, especially affecting their behavior and thinking. Regularly, the self-awareness of having a mental disorder causes emotional and physical damage that could make people feel fear to the idea of being vulnerable to criticism, judgment, or opposing opinions. Mental disorders may be related to a particular event that generated excessive stress on the affected person or to a series of different stressful events (World Health Organization, 2019). For instance, some of the causes are environmental stress, genetic factors, or different difficult life situations. There are several common mental disorders such as depression, anorexia, or self-harm affecting people worldwide (Kessler et al., 2017).

A reality nowadays, is that for some people their social life does not occur in their surroundings or immediate environment, but takes place in a virtual world created by social media platforms like Facebook, Twitter, or Reddit (Baer, 2021). In other words, social media has become a vital link for some of us. This scenario presents opportunities to study and analyze, given the availability of data, how people communicate, and more specifically, how this communication could be associated to possible mental health issues that people are experimenting, contributing in this way to attract attention to silent disorders.

Previous studies have shown that texts shared by users in social networks have different evidence or types (channels or pseudo modalities) of information that may be relevant for the detection of mental disorders (Guntuku et al., 2017), for example, their topics of interest, emotional state, or their writing style. This has motivated us to propose a method that considers these pieces of information and to study how to combine or fusion them. With this in mind, we evaluate the plausibility of using the multimodal BERT (MMBT)(Kiela et al., 2019) for this task, being this the first time. In addition, we state the question: whether MMBT is the best option to handle this sensitive task or if multiple BERTs can better exploit the nature of the data. Thus, we compare the performance of MMBT against different architectures based on early and late fusion approaches of the popular BERT model. In this study we take the text modality and divide it into three channels[1] that focus on different aspects of the users' communication. The first channel captures the thematic information used for contextual analysis. The second channel indicates the manifested emotions, attempting to capture emotional topics related to mental disorders. Finally, the third channel focuses on the writing style, where we want to capture the use of personal expressions and verbs tense, among other aspects. As could be observed, our hypothesis is that people that present some mental disorder tend to express differently, at different dimensions, regarding the control group.

It is widely known that the BERT model (Devlin et al., 2019) has led to important improvements in representation learning for natural language processing and text classification problems. In a recent work (Kiela et al., 2019), the authors demonstrate that supervised bidirectional transformers with unimodal pre-trained components obtain good performance in multimodal fusion. They found that learning to map dense multimodal features to BERT's token embedding space is easy to extend to different modalities. Inspired by their findings, we adapt their Multimodal BERT (MMBT) module with our channels as modalities to create a multichannel contextualized representation. The main idea of our work is to find out the best way to combine the different types of information and see if using multimodal BERT is better than considering a fusion of multiple BERTs, each one specialized in a different channel.

We can summarize the contributions of our work as follows:

1. We adapt a Multimodal BERT (MMBT) for the detection of mental disorders, considering three channels of information: thematic, emotional and stylistic.
2. We explore different strategies to combine these information, considering early and late fusion approaches.
3. We analyze and evaluate in detail these three information channels and the importance of their fusion, then concluding about its feasibility of integration to boost classification performance.

[1]In this work, we define a *channel* as a different property or view from the same modality (Qianli et al., 2017).

2 Related work

2.1 Mental disorders detection

In the last few years, the study of public mental health through social media has increased. This is mainly because these media provide a source of support for those who suffer from a mental health disorder, like for example, a sense of community, relatedness, and understanding (Hilton, 2016; Dyson et al., 2016). In general, for the construction of corpora, researchers 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 (De Choudhury, Counts, and Horvitz, 2013; Wang et al., 2017).

Recent works (Trifan and Oliveira, 2019; Van Rijen et al., 2019), explored the analysis of the posts' content. In these works, the authors consider different features, such as word and char n-grams, and then apply a classification algorithm to make a decision. The shortcoming with that strategy is the high overlap in the vocabulary used by the control and positive users, which generates several missclassifications. Another well-known strategy consists of counting the number of occurrences of positive, negative, and neutral words in texts (Kang, Yoon, and Kim, 2016), or on measuring how similar are their words to some reference negative and positive lexicons (Htait, Fournier, and Bellot, 2017). On the other hand, analyzing sentiments has shown interesting results since it has been found that negative comments are more abundant in people with a declared mental health disorder than in comments generated from a control group (Coopersmith, Dredze, and Harman, 2014; Preotiuc-Pietro et al., 2015). Other works have used a LIWC-based representation (Tausczik and Pennebaker, 2010), which consists of a set of psychological categories that aim to represent users' posts by features of social relationships, thinking styles and individual differences (Coppersmith et al., 2015).

2.2 Fusion approaches for mental disorders detection

How to effectively combine information is challenging and has a long history in machine learning (Baltrusaitis, Ahuja, and Morency, 2019). In particular, some recent works on mental disorders detection have considered the use of ensemble approaches to combine bag of words representations, LIWC features, and different deep neural models (Trotzek, Koitka, and Friedrich, 2018). In (Ragheb et al., 2019), the authors combine the temporal mood variation and Bayesian inference for their detection. The first phase uses an attention-based deep model to construct a representation for the mood variation. Then, in the second phase, the model uses Bayesian inference to detect clear signs of mental disorders and then give a decision based on their combination. In (Ji et al., 2020), the authors apply an attention model combined with sentiment and topic analysis to detect suicidal ideation. A recent work (Uban, Chulvi, and Rosso, 2021), explored the evolution of emotion expression in relation to cognitive styles and found specific patterns in users with mental disorders.

The performance shown by ensemble approaches suggests the suitability of adapting advanced techniques to fusion information from different channels in a more effective way. For that reason, we decided to implement our multi-channel BERT-based approach as a new way to combine the thematic, emotional and stylistic views of the information shared by social media users. This proposed model takes inspiration from multimodal BERT (Kiela et al., 2019) in order to have an adapted version that is able to model individual channels. As our experimental evaluation will show, the proposed strategy improves the classification performance.

3 Information channels representation

The fusion strategies that we are going to explore are implemented on the basis of BERT, therefore, the three channels of information are captured by three different representations of the words. The main idea of this approach is to have different views of the content of the users' posts. We achieve this by generating *three embeddings* for each word in the posts, capturing or emphasizing the thematic, emotional and style information, respectively. In the following subsections, we briefly describe how to generate these three representations.

3.1 Thematic embeddings

In order to capture the thematic content related to each word, we consider vanilla GloVe embeddings (Pennington, Socher, and Man-

ning, 2014). However, since this type of embeddings does not take into account the context of the words, we decided to also use some contextualized word embeddings, in particular the BERT embeddings (Devlin et al., 2019). For our experiments, we used both separately and evaluated which one contributes the most to the final representation.

With contextualized embeddings, words that have similar meaning or show some semantic relation are closer to each other. For example, the words "insecure" and "worried" have similar embeddings, as do the words "therapy" and "treatment".

3.2 Emotion-based embeddings

For this work, we use the emotion-based word embeddings that were originally proposed in (Aragon et al., 2019). In short, to construct these vectors, first, we generated groups of fine-grained emotions for each general emotion that belong to the EmoLEX lexicon (Mohammad and Turney, 2013). We achieved this by representing each word of the lexicon with its FastText embedding (Bojanowski et al., 2016) and then applying a clustering algorithm on them. After obtaining the fine-grained emotions, which are groups of words that capture specific topics related to the same emotion, we represented each of them by means of the average vector of its words. Subsequently, and as the last step, for each word in the vocabulary we measured its cosine similarity with all fine-grained emotions, and assigned to each one of the embedding from its closest fine-grained emotion.

According to the process described above, the words "accident" and "crash" will have the same embedding because both belong to the same subgroup of the Surprise emotion, whereas the word "magician" will have a slightly different embedding since it corresponds to a different subgroup of the same emotion. On the other hand, the words "accomplish" and "achieve" will have a completely different embedding as they belong to the Joy emotion.

3.3 Style-based embeddings

The third representation of the words aims to capture particular characteristics of the writing style of social media users. Its idea is to capture how users with mental disorders tend to talk, for example, referring to past events or to uncertainties about the future. To capture the stylistic information, we propose a new word representation inspired by the successful use of character n-grams in author profiling tasks.

To define the style-based embedding of each word from the users' posts, we carried out the following process:

1. Divide the word into character 3-grams.
2. Compute, for each 3-gram, its embedding using FastText (Bojanowski et al., 2016), as well as its discriminative score according to its chi^2 distribution in the two given classes (positive and control users).
3. Obtain the embedding vector of the word by applying a weighted sum of the vectors of its character 3-grams, considering as weights their chi^2 values.

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-based embeddings are similar for words that have similar spelling rather than meaning. For example, words in superlative resemble each other, as well as regular verbs in past tense, or words with the same root. Take for instance the word "mental" some of their closest words would be "dental", "mentality" or "decremental".

4 On the fusion of the three channels

The objective of our work is to compare different ways of combining information from different channels. This is a key stage in the classification process, and have the intuitive idea of learning the relevance of each channel in an automatic way. We use two main strategies, firstly, one based on multimodal BERT whose idea is to learn a joint representation of the three types of information, and secondly, different architectures that treat each channel separately and apply different early and late fusion techniques.

4.1 Multimodal BERT (MMBT)

It is a recently proposed supervised multimodal bitransformer model for classifying images and text (Kiela et al., 2019). The MMBT model starts with pre-trained BERT

weights, and takes their contextual embeddings as input. These contextual embeddings are obtained as the sum of the segment, position, and token embeddings of each word. Then, the model weights them and project each of the embeddings to a token input. In Figure 1, we can appreciate the components of the architecture of the MMBT model. Although proposed for only two modalities, this architecture can be generalized to any number of modalities, assigning a different segment id to each of them.

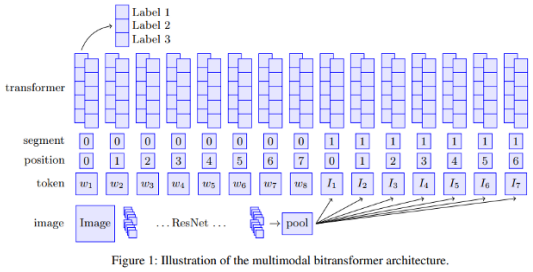

Figure 1: Illustration of the multimodal bitransformer architecture.

Figure 1: Multimodal bitransformer architecture proposed in (Kiela et al., 2019).

In the original work, the authors took contextual embeddings as input, learned the weights, and projected each image's embeddings to a dimensional token input. Instead of image embeddings, we used the sequence of the words for the fine-tuning with our different channels as embeddings. Once we fine-tuned the model with the channels, we took advantage of the contextual information learned for classifying the users' posts. For this purpose, we used the first output of the final MMBT layer as input to a Convolutional Neural Network (CNN) for feature extraction, and then add a dense layer for achieving the classification. Figure 2 presents the general architecture of this approach.

4.2 Combining information using multiple BERTs

The authors of MMBT (Kiela et al., 2019) noted that their method is compatible with scenarios where not every modality is present and can be generalized to an arbitrary number of modalities. Then, for the second model, we decided to train individual BERTs and fine-tuning them with each channel separately. After the training, similar to the first approach, we used the first output of the final layer of each BERT and concatenate them as input for a CNN layer, we will refer to

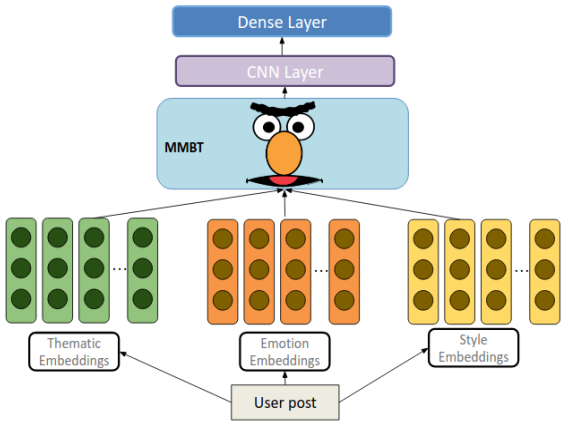

Figure 2: General diagram of the Model 1: Multimodal BERT with vectors of three channels, then a CNN layer, and a classification layer.

this model as BERT-CNN. In Figure 3, we present the general diagram for this process.

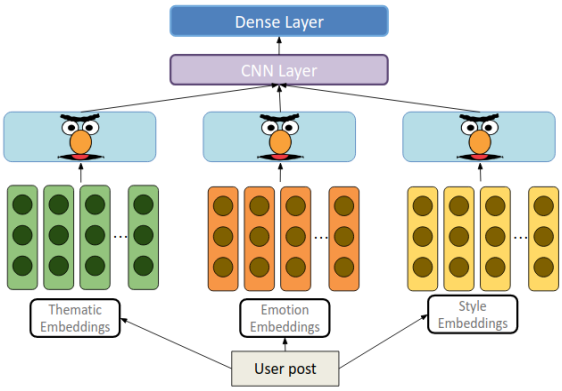

Figure 3: General diagram of the BERT-CNN model: Each channel separately enters to a BERT model, then join vectors feed a single CNN layer, and a classification layer.

For the third model, instead of concatenating the vectors and using a single CNN layer, we separate them into different convolutional layers and used the output for a dense layer. With this approach, the model obtains for each channel different feature maps of each region and concatenates them together to form a single feature vector. This can be interpreted as summarizing the local information to find patterns, and then combining the information. The hypothesis that the local information per channel is important and should be extracted before it is combined, we call this model BERT-3CNN. Figure 4 presents the general diagram for this model.

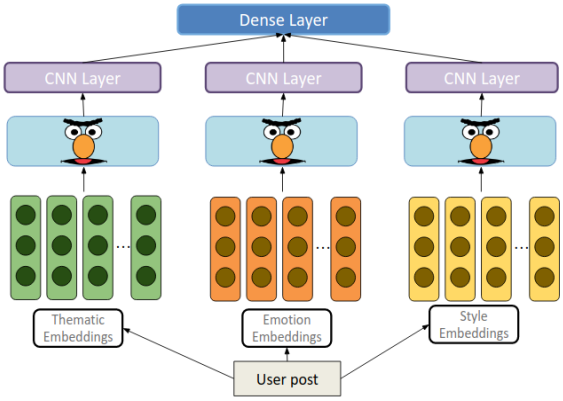

Figure 4: General diagram of the BERT-3CNN model: Each channel separately enters a BERT model and a CNN layer, then their outputs are concatenated and fed to a single classifier.

One of the challenges of this work is the problem of fusing information. A simple solution is to concatenate the representations of each channel into one vector or perform an operation like adding or taking the product. However, the use of these operations assumes that all channels have the same relevance, which is usually not the case. In recent work (Arevalo et al., 2019), 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 into the next one. 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.

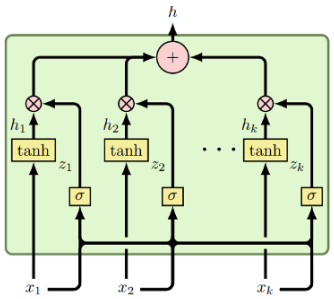

Figure 5: Overview of GMU module (Arevalo et al., 2019). Where x_i represents the ith input modality. The final fused representation of all modalities is represented by h at the top.

Motivated by the outstanding results of the GMU module in different multimodal tasks, our fourth model takes advantage of it. That is, after the feature extraction, we implement a Gated Multimodal Unit (GMU) module to learn the relations between each channel feature vector. Then, apply a dense layer to classify the final vector with the information of the three channels, we call this model BERT-GMU. Figure 6 describes the process for this model.

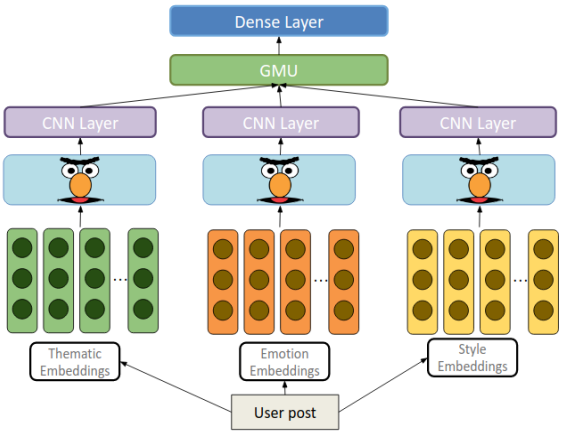

Figure 6: General diagram of the BERT-GMU model: Each channel separately enters a BERT model and a CNN layer, then their outputs are combined by a GMU module, and fed to a single classifier.

5 *Experimental settings*

5.1 Data sets

We performed experiments over data sets from the eRisk 2019 and 2020 evaluation tasks (Losada, Crestani, and Parapar, 2019; Losada, Crestani, and Parapar, 2020). These data sets consist of the detection of depression, anorexia, and self-harm, and contain the post history of several users from the Reddit platform. For each mental disorder, we have two types of users: 1) the control group, people collected who do not suffer from any mental disorder; and 2) positive users, a group composed of people affected by either depression, anorexia, or self-harm.

In the tasks of anorexia and self-harm, the positive class is composed of users who explicitly mentioned that they were diagnosed by a medical specialist or that they had committed self-harm. On the other hand, the control class for both tasks is composed of random users from the Reddit platform. However, to add realism to the data sets and make the

detection of positive users challenging, the control group also contains users who often interact in the threads of anorexia and self-harm.

For depression, the organizers of the shared task asked the participants to predict, for each user, the possible answer to each input of the BDI questionnaire (Beck et al., 1961), which contains 21 questions that allow assessing the level of severity of the depression. In contrast to them, in this work 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 fewer in their final result.

Table 1 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 (Losada, Crestani, and Parapar, 2018), this data set was constructed similarly to anorexia and self-harm data sets.

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	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	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	24.8	18.8	21.4	11.9
avg. days	495	500	270	426

Table 1: Data sets used for experimentation, where P indicates the positive users and C is used for control users.

5.2 Preprocessing

The texts were normalized by lowercasing all words and removing special characters like URLs, emoticons, and #; the stopwords were kept. Our decision to keep stopwords was completely experimental, we performed experiments removing them before masking the texts, but consistently we got slightly lower performances.

5.3 Classification

The main goal is to classify users into one of the two classes (Depressed / Control, Anorexia / Control, or Self-harm / Control). 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 depression, anorexia, or self-harm through all the user posts.

5.4 Baselines

The results are compared to the traditional Bag-of-Words representation combined with a SVM classifier. This representation was created using word unigrams and n-grams; these are common baseline approaches for text classification. For both approaches, we selected the same number of features using tf-idf representation and chi^2 distribution X_k^2. We also add some baselines based on deep learning approaches, using a CNN and a Bi-LSTM. The neural networks used 100 neurons, an adam optimizer, and word2vec and Glove embeddings with a dimension of 300. For the CNN we use 100 random filters of sizes 1, 2, and 3 (parameters recommended in literature). We also add 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, precision, and recall over the positive class (Losada, Crestani, and Parapar, 2018).

6 Evaluation and Analysis

For the evaluation, besides the baselines, we also performed experiments using independently the proposed thematic, emotion and style representations. Table 2 presents the results in terms of F_1 score, precision, and recall 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 proposal using all information channels

Method	Anor			Dep			SH		
Baselines									
	F1	P	R	F1	P	R	F1	P	R
BoW-unigrams	0.67	0.85	0.55	0.58	0.56	0.60	0.50	0.95	0.34
BoW-Ngrams	0.66	0.83	0.55	0.57	0.55	0.59	0.50	0.92	0.33
Bag of char 3grams	0.67	0.85	0.55	0.58	0.56	0.60	0.52	0.97	0.36
RNN-word2vec	0.65	0.95	0.49	0.57	0.62	0.53	0.55	0.60	0.51
CNN-word2vec	0.66	0.94	0.52	0.60	0.57	0.62	0.56	0.54	0.59
RNN-GloVe	0.65	0.92	0.51	0.58	0.59	0.57	0.57	0.62	0.53
CNN-GloVe	0.67	0.93	0.52	0.61	0.56	0.68	0.57	0.62	0.53
RNN-Attention	0.66	0.94	0.52	0.50	0.67	0.40	0.58	0.76	0.47
Best eRisk participants									
1st	0.71	0.64	0.79	-	-	-	**0.75**	0.82	0.69
2nd	0.68	0.77	0.60	-	-	-	0.62	0.62	0.62
3rd	0.68	0.67	0.68	-	-	-	0.62	0.59	0.65
Our methods: Single-channel									
Thematic	0.77	0.70	0.85	0.62	0.55	0.72	0.60	0.44	0.94
Emotion	0.70	0.85	0.60	0.61	0.62	0.61	0.63	0.68	0.59
Style	0.69	0.86	0.57	0.62	0.64	0.61	0.64	0.70	0.59
Our methods: Multi-channel Bi-transformers									
MMBT	0.76	0.72	0.84	0.65	0.51	0.91	0.65	0.75	0.58
BERT-CNN	**0.82**	0.81	0.82	**0.70**	0.54	0.96	0.70	0.73	0.68
BERT-3CNN	0.80	0.81	0.79	0.68	0.53	0.95	0.70	0.69	0.71
BERT-GMU	0.81	0.80	0.81	**0.70**	0.55	0.95	0.73	0.73	0.74

Table 2: F1, precision and recall results over the positive class in three eRisk's tasks.

combined by the multimodal transformer as well as by different architectures based on multiple BERTs.

From this evaluation, we observe that most of our proposals outperform the baseline results. Firstly, the single-channel representations obtain an improvement in comparison with baselines, in particular those based on style and emotion information. Unexpectedly, for the baselines, the performance of deep learning models applied over word-based representations is closer to traditional approaches using a Bag-of-Words. We think this could be due to the small size of the data set and the intersection of thematic content. Something interesting to notice is that CNN networks obtain better performance than RNN networks. The latter could be because CNN networks search for the presence of specific local information important for the detection of these disorders.

For our representations based on the fusion of information, their performance is higher in comparison with the other models, suggesting the relevance of combining the information from different channels. Something interesting to notice is that all models using multiple BERTs outperformed the multimodal BERT model in F1, indicating that, for this particular task, and with this way of representing the channels, it is better to represent each channel independently and combine them later. In general, the model that use the GMU module showed the best average performance, which suggest that weighting the information helps to create a better representation of the posts and the users.

From these experiments, we highlight the following observations:

1. Most single-channel representations obtain better results than the baselines.

2. The architectures using multiple BERTs obtained better performance than the multimodal BERT.

3. These results confirms our intuition that learning to combine different types of information is very relevant to capture signs of mental disorders in users.

4. In general, our models obtain an harmonic result between precision and recall deriving in a better F1 score.

6.1 Comparison against the eRisk participants

To expand our analysis and add context to the results, we also include a comparison against the original participants of the eRisk tasks. These evaluation forum considers a total of 54 models for the anorexia detection task and 57 for the self-harm detection task in eRisk-19 and 20 editions (Losada, Crestani, and Parapar, 2019; Losada, Crestani, and Parapar, 2020). It is important to mention that the participants focused on obtaining early and accurate predictions on the users, while our approach focuses exclusively on determining accurate classifications.

We observe that our models achieve competitive results in both tasks; they surpassed the first place results in Anorexia, and showed a slightly lower performance than the first place in Self-harm. For the depression task, organizers changed the evaluation strategy. While our approach focuses on binary classification, the eRisk task considered the assessment of the level of depression severity for each user. For this reason, we cannot directly compare our results against the participants.

6.2 Contribution of each information channel

To understand how each information channel contributes to the final decision we will utilize the GMU units in the fourth model (BERT-GMU) and analyze the weighting of the gates, where the gates in the unit let the model regulate the flow of information into the next one. The main idea in a GMU is that the unit learns to weigh 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.

For this analysis, we obtained the gates' z_i values of the GMU module corresponding to the test set posts. Figure 7 presents the results for the three tasks, where each value already takes into account the average of all posts. For anorexia, we can appreciate that the thematic information contributes the most to the final decision, followed by emotion and style information. For depression, we can observe that the thematic information is also the most important and the value for the style information is higher than the emotional value. Finally, for the self-harm task, the thematic information obtains the lowest value and the style information the highest. In general, it can be noticed how the activation for each channel are different depending on the mental disorder. Something interesting is that the thematic channel presents the highest variation, with the lowest value in self-harm and the highest value in anorexia. We think that this variation indicates that the posts of users who suffer from anorexia are probably more homogeneous than those who suffer self-harm.

For a further analysis of the GMU, Table 3 presents the posts of the depression data set with the highest z_i value for each channel. We can notice that the posts are related to personal opinions, different topics, and in general express negative emotions even when they are not directly related to mental disorders. Take for example the emotion channel where the post is related to regrets in life and feelings, or the style channel where the post talks about a mental illness, but also contains words such as "don't" and "nothing".

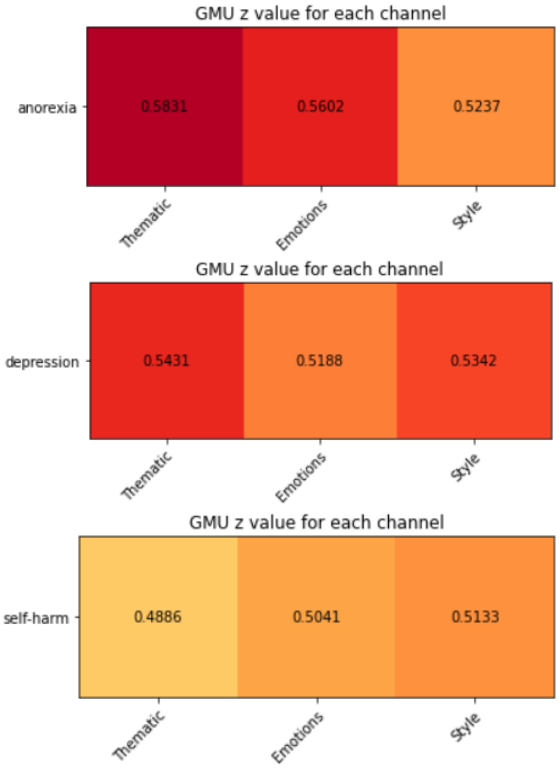

Figure 7: Average z_i value for the three mental disorders over the test set instances.

7 Conclusion and future work

In this work, we explored the detection of users that suffer anorexia, depression, or self-

Channel	Post
thematic	*"...I have no idea what either of them were trying to communicate tbh i was* ***having a really good day*** *and then you had to bring up hughes..."*
emotion	*"...take the chance and have no* ***regrets*** *in life, its always better to know if the other person* ***feels*** *something so that you are not wasting your time..."*
style	*"...these days parents* ***don't*** *know a whole lot about* ***mental illness*** *they were told it was* ***nothing*** *so that's all..."*

Table 3: Posts with highest z_i value for each channel over the depression task.

harm. For this task, we used the users' thematic interests, emotions and writing style. We tested different strategies to combine these information channels inspired by the usage of transformers to learn contextual knowledge. Our results suggest that enriching emotional and style data using a transformer improves the detection of users with mental disorders. Moreover, a striking result is the superiority of using a combination of multiple BERTs over the recent multimodal BERT transformer; this finding by its own opens an opportunity to explore models inspired by transformers to create new representations and continue improving the performance in the detection of people with mental disorders. The results outperform traditional and state-of-the-art baselines and are competitive with the performance of top eRisk participants. We believe that it is important to mention that these models, although they obtain better results, are extremely resource-consuming (processor, memory, energy, etc.) in comparison with simple models. For future work, we want to explore more sophisticated combination techniques that could improve the results and understanding of mental disorders detection, for example, multi-modal transformers. We note that most of the analysis of mental disorders has been made for the English language, then, one of our interests lies in the expansion of this study to Spanish language.

References

Aragon, M. v., A. Lopez-Monroy, L. Gonzalez-Gurrola, and M. Montes-y Gomez. 2019. Detecting depression in social media using fine-grained emotions. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).*

Arevalo, J., T. Solorio, M. Montes-y Gómez, and F. González. 2019. Gated multimodal networks. *Neural Computing and Applications.*

Baer, J. 2021. Multimodal machine learning: A survey and taxonomy. *https://www.convinceandconvert.com/social-media-research/social-media-usage-statistics/.*

Baltrusaitis, T., C. Ahuja, and L. Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2):423–443.*

Beck, A., C. Ward, M. Mendelson, J. Mock, and J. Erbaugh. 1961. An inventory for measuring depression. *JAMA Psychiatry 4(6), 561–571.*

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov. 2016. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics.*

Coopersmith, G., M. Dredze, and C. Harman. 2014. Quantifying mental health signals in twitter. *Workshop on Computational Linguistics and Clinical Psychology.*

Coppersmith, G., M. Dredze, C. Harman, and K. Hollingshead. 2015. From adhd to sad: analyzing the language of mental health on twitter through self-reported diagnoses. *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology*, pages 1–10.

De Choudhury, M., S. Counts, and E. Horvitz. 2013. Social media as a measurement tool of depression in populations. *In Proceedings of the 5th Annual ACM Web Science Conference.*

Devlin, J., M. Chang, K. Lee, and K. Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HTL.*

Dyson, M., L. Hartling, J. Shulhan, A. Chisholm, A. Milne, P. Sundar, S. Scott, and A. Newton. 2016. A systematic review of social media use to discuss and view deliberate self-harm acts. *PLOS ONE.*

Guntuku, S., D. Yaden, M. Kern, L. Ungar, and J. Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, pages 43–49.

Hilton, C. 2016. Unveiling self-harm behaviour: what can social media site twitter tell us about self-harm? a qualitative exploration. *Journal of clinical nursing.*

Htait, A., S. Fournier, and P. Bellot. 2017. Lsis at semeval-2017 task 4: Using adapted sentiment similarity seed words for english and arabic tweet polarity classification. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017).*

Ji, S., X. Li, Z. Huang, and E. Cambria. 2020. Suicidal ideation and mental disorder detection with attentive relation networks. *arXiv:2004.07601.*

Kang, K., C. Yoon, and E. Kim. 2016. Identifying depressive users in twitter using multimodal analysis. *In Big Data and Smart Computing (BigComp), 2016 International Conference on. IEEE, 231–238.*

Kessler, R., E. Bromet, P. Jonge, V. Shahly, and Marsha. 2017. The burden of depressive illness. *Public Health Perspectives on Depressive Disorders.*

Kiela, D., S. Bhooshan, H. Firooz, E. Perez, and D. Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. *Visually Grounded Interaction and Language (ViGIL), NeurIPS 2019 Workshop.*

Losada, D., F. Crestani, and J. Parapar. 2018. Overview of erisk 2018: Early risk prediction on the internet (extended lab overview). *Proceedings of the 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France.*

Losada, D., F. Crestani, and J. Parapar. 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).*

Losada, D. v., F. Crestani, and J. Parapar. 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.*

Mohammad, S. and P. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence.*

Pennington, J., R. Socher, and C. Manning. 2014. Glove: global vectors for word representation. *In Proceedings of the Conference on Empirical Methods on Natural Language Processing.*

Preotiuc-Pietro, D., J. Eichstaedt, G. Park, M. Sap, L. Smith, V. Tobolsky, H. Schwartz, and L. Ungar. 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.*

Qianli, M., S. Lifeng, C. Enhuan, T. Shuai, W. Jiabing, and C. Garrison. 2017. Walking walking walking: Action recognition from action echoes. *Twenty-Sixth International Joint Conference on Artificial Intelligence.*

Ragheb, W., J. Aze, S. Bringay, and M. Servajean. 2019. Attentive multi-stage learning for early risk detection of signs of anorexia and self-harm on social media. *Proceedings of the 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland.*

Renteria-Rodriguez, M. 2018. Salud mental en mexico. *NOTA-INCyTU NÚMERO 007.*

Tausczik, Y. and J. Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, pages 24–54.

Trifan, A. and J. Oliveira. 2019. Bioinfo@uavr at erisk 2019: delving into social media texts for the early detection of mental and food disorders. *Proceedings of the 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland.*

Trotzek, M., S. Koitka, and C. Friedrich. 2018. Word embeddings and linguistic metadata at the clef 2018 tasks for early detection of depression and anorexia. *Proceedings of the 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France.*

Uban, A., B. Chulvi, and P. Rosso. 2021. An emotion and cognitive based analysis of mental health disorders from social media data. *Future Generation Computer Systems.*

Van Rijen, P., D. Teodoro, N. Naderi, L. Mottin, J. Knafou, M. Jeffryes, and P. Ruch. 2019. A data-driven approach for measuring the severity of the signs of depression using reddit posts. *Proceedings of the 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland.*

Wang, T., M. Brede, A. Ianni, and E. Mentzakis. 2017. Detecting and characterizing eating-disorder communities on social media. *In Proceedings of the Tenth ACM International conference on web search and data mining.*

World Health Organization, W. 2019. Mental health: Fact sheet. *https://www.euro. who.int/en/health-topics/noncommunicable-diseases/mental-health.*