Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3202–3219
Marseille, 20-25 June 2022

Multi-Aspect Transfer Learning for Detecting Low Resource Mental Disorders on Social Media

Ana Sabina Uban[1,3], Berta Chulvi[1,2], Paolo Rosso[1]
[1]Universitat Politècnica de València, València, Spain,
[2]Universitat de València, Valencia, Spain,
[3]University of Bucharest, Bucharest, Romania
auban@fmi.unibuc.ro, berta.chulvi@upv.es, prosso@dsic.upv.es

Abstract

Mental disorders are a serious and increasingly relevant public health issue. NLP methods have the potential to assist with automatic mental health disorder detection, but building annotated datasets for this task can be challenging; moreover, annotated data is very scarce for disorders other than depression. Understanding the commonalities between certain disorders is also important for clinicians who face the problem of shifting standards of diagnosis. We propose that transfer learning with linguistic features can be useful for approaching both the technical problem of improving mental disorder detection in the context of data scarcity, and the clinical problem of understanding the overlapping symptoms between certain disorders. In this paper, we target four disorders: depression, PTSD, anorexia and self-harm. We explore multi-aspect transfer learning for detecting mental disorders from social media texts, using deep learning models with multi-aspect representations of language (including multiple types of interpretable linguistic features). We explore different transfer learning strategies for cross-disorder and cross-platform transfer, and show that transfer learning can be effective for improving prediction performance for disorders where little annotated data is available. We offer insights into which linguistic features are the most useful vehicles for transferring knowledge, through ablation experiments, as well as error analysis.

Keywords: Mental disorders, depression, transfer learning, explainability, social media, low resource

1. Introduction

Mental health disorders are an important and pervasive public health issue. Depression in particular affects approximately 300 million people worldwide (World Health Organization, 2012), and the problem has recently been exhacerbated by the COVID-19 pandemic (Lima et al., 2020; Shah et al., 2020; Shigemura et al., 2020; Torales et al., 2020; Xiang et al., 2020). Mental health disorders are closely connected to suicide, with depression present in 35% of suicides (Bridge et al., 2006). Self-harm is a strong predictor of suicide, present in the recent histories of around 40% of suicides (Cavanagh et al., 2003). Moreover, mental disorders are massively underdiagnosed and undertreated (Sheehan, 2004; Allan et al., 2014), with more than half of the people suffering from depression not receiving any treatment. People affected by mental disorders are often reluctant to approach a specialized clinician to seek help with treating the disorder. However, more and more frequently people turn to social media to discuss their issues and to seek emotional support. This opens up an important opportunity for automatic processing of social media data in order to identify changes in mental health status that may otherwise go undetected before they develop more serious health consequences.

The way mental disorders manifest and can be recognized is primarily through everyday communication. It has been shown that individuals suffering from mental disorders manifest changes in their language, either explicit, at the level of topics discussed, or implicit, such as through expressing greater negative emotion and high self-attentional focus (De Choudhury et al., 2014; Guntuku et al., 2017; Trotzek et al., 2018). Text data collected from social media can thus be a valuable source for analyzing signs of mental disorders. However, manually annotating datasets or cross-referencing medical records to obtain diagnosis labels can be challenging and may pose privacy concerns. Thus, several studies on mental health disorders provided datasets annotated semi-automatically, based on self-stated diagnoses. Most of these are focused on depression, while other disorders such as anorexia or PTSD have received much less attention, with very few annotated data available.

From a clinical perspective, the diagnosis of certain disorders (such as depression or anorexia) can also be a complicated issue, with the standards for diagnosis constantly evolving (Surís et al., 2016), and with significant overlap in symptomatology across some disorders (American Psychiatric Association, 2013). There is clinical evidence on the co-morbidity between certain disorders (Plana-Ripoll et al., 2019): being diagnosed with one mental disorder, such as depression, increases the risk of subsequently being diagnosed with another, such as anxiety, with up to 40% risk for some disorders (Kaufman and Charney, 2000). This suggests that different mental disorders may not only manifest similarly (through behavior and language), but also frequently occur in the same individuals, supporting the idea of analyzing these disorders jointly. Most computational studies in mental health treat each disorder in isolation, which misses the opportunity to model coinciding influence factors. Tasks in NLP with underlying commonalities have been shown to benefit from transfer learning and multi-task learning (MTL), as the learning implicitly leverages interactions between them (Caruana, 1997; Sutton et al., 2007; Collobert et al., 2011; Søgaard and Goldberg, 2016). Moreover, computational methods could help with a deeper understanding of the connections between these disorders and their particular manifestations and thus provide valuable insights to the clinicians developing diagnosis standards. Considering both the technical and practical potential, as well as the impact on clinical psychology, the research questions we propose to answer are as follows:

(RQ1) Can transfer learning be leveraged in order to improve the detection performance of automatic deep learning models for disorders where datasets are scarce, and can deep learning models for mental disorder detection be used successfully across different social media platforms?

(RQ2) What can we learn about the similarity between the different disorders through studying the effectiveness of transfer learning across different disorders, and how could this assist research on development of diagnostic criteria for certain disorders?

(RQ3) How can we use interpretable multi-aspect deep learning models to reveal qualitative conclusions about the specific linguistic dimensions which are more similar across different disorders?

In this study, we bring several contributions to research into the automatic detection of mental disorders, at different levels. Firstly, we use various datasets of social media posts from users suffering from different disorders: depression, PTSD, anorexia and self-harm, using data collected from Reddit and Twitter. We experiment with deep learning models for automatically predicting these disorders, including hierarchical attention networks (HAN) and transformers, and are the first to combine a deep learning architecture with a multi-aspect representation at the feature level in the context of transfer learning, through features that reflect various complementary levels of the language, including content, style and emotion. We explore the use of these models for transfer learning for mental disorder detection, and systematically compare different transfer learning strategies, showing how they can be used to leverage existing solutions tailored to a specific disorder for detecting other disorders from similar social media texts (cross-disorder), as well as for texts on different platforms (cross-platform). We complement these results with experiments for automatically distinguishing among different disorders. We additionally include ablation experiments, as well as error analysis, by leveraging our multi-aspect representations in order to compare which features are most useful for transferring knowledge between the different models, as a way to gain insight into which aspects of the language are most similar across disorders.

2. Previous Work

There is an extensive body of research related to automatic risk detection for mental disorders from social media data (Calvo et al., 2017; Guntuku et al., 2017), focusing especially on the study of depression (De Choudhury et al., 2013; Eichstaedt et al., 2018; Abd Yusof et al., 2017; Yazdavar et al., 2017), but other mental illnesses have also been studied, including generalized anxiety disorder (Shen and Rudzicz, 2017), schizophrenia (Mitchell et al., 2015), PTSD (Coppersmith et al., 2014; Coppersmith et al., 2015), risks of suicide (O'dea et al., 2015; Ramírez-Cifuentes et al., 2020; Sawhney et al., 2021), anorexia (Losada et al., 2019; Ramírez-Cifuentes et al., 2018; Ramıírez-Cifuentes et al., 2020) and self-harm (Losada et al., 2019; Yang et al., 2016).

Annotating datasets for mental disorder detection with reliable labels involves cross-referencing social media users and patients' medical records (De Choudhury et al., 2014), which can be a complicated process, and sometimes unfeasible due to privacy concerns. For this reason, most studies rely on self-stated diagnoses in order to semi-automatically annotate social media users who suffer from a mental disorder (Losada et al., 2018; Losada et al., 2019; Shen and Rudzicz, 2017; Coppersmith et al., 2015). For the study of depression, different social media platforms have been considered, such as Twitter (Chen et al., 2018; Shen and Rudzicz, 2017), Facebook (De Choudhury et al., 2014), or Reddit (Losada et al., 2018), and several datasets have been made available to the community. For other mental health disorders, there are fewer computational studies, and a lack of public annotated datasets, with at most one or two available datasets for each disorder, which generally cover only a few hundred users (e.g. for anorexia (Losada et al., 2019; Cohan et al., 2018), for PTSD (Coppersmith et al., 2015), self-harm (Losada et al., 2020), or suicide ideation (Ramírez-Cifuentes et al., 2020; Sawhney et al., 2021)).

Historically, the majority of computational studies on mental health have provided either quantitative analyses, or predictors built using classical machine learning models (De Choudhury et al., 2013; De Choudhury et al., 2014). Fewer studies have made use of deep learning methods such as different types of convolutional (CNN) or recurrent neural networks (RNN) (Sadeque et al., 2017; Shen et al., 2017; Wang et al., 2018; Trotzek et al., 2018; Orabi et al., 2018), or pretrained transformers (Matero et al., 2019; Zirikly et al., 2019). Recently, (Rao et al., 2020) use hierarchical networks for depression detection, and (Amini and Kosseim, 2020) and (Mohammadi et al., 2019) use HANs for anorexia detection (the latter obtaining best results at the eRisk 2019 shared task (Losada et al., 2019)) - all of which rely on word n-gram features. In general, many previous works have used traditional bag of word n-grams (Coppersmith et al., 2014), while some have also applied more domain-specific representations, such as hand-crafted lexicons (Trotzek et al., 2017), LIWC (Pennebaker et al., 2001) features (De Choudhury et al., 2014), Latent Semantic Analysis (Resnik et al., 2013; Trotzek et al., 2017) or other linguistic features such as parts of speech (Bucur et al., 2021). There are few studies which jointly include in their models several different aspects of the language for assessing risk of mental disorders (Shen et al., 2017; Shen and Rudzicz, 2017; Leiva and Freire, 2017), and in the case of deep learning models, only word embedding features are typically used. We find few other computational studies which focus on an in-depth analysis of emotions to model mental disorders in language (Uban et al., 2021a; Aragón et al., 2019), as well as some interdisciplinary studies showing quantitative analyses of the relationship between negative emotions and mental disorders (O'Dea et al., 2021; Seabrook et al., 2018; Cheng et al., 2020).

Few studies consider several disorders jointly (Yates et al., 2017; Coppersmith et al., 2015; Saha et al., 2019; Gkotsis et al., 2017), and very few have studied the effectiveness of transfer learning for cross-disorder prediction of mental disorder risk. In (Saha et al., 2019), the authors study the effects of psychiatric medications across different disorders, using SVMs and statistical analyses. Gkotsis et al. (2017) use a CNN to predict different disorders and classify among disorders, showing in particular high confusion between depression and other disorders. A few participants in the eRisk shared tasks on unsupervised risk prediction for depression (Losada et al., 2018) and self-harm (Losada et al., 2019) have leveraged transfer learning through data augmentation, including data labelled for different disorders (Abed-Esfahani et al., 2019), with moderate success. On cross-domain transfer learning, one study (Shen et al., 2018) looks at detecting depression risk across different social media platforms: Reddit and the Chinese Weibo platform. Harrigian et al. (2020) investigate the cross-domain performance of depression detection systems, concluding that depression detection models do not generalize well across platforms. With respect to cross-disorder transfer, one study (Benton et al., 2017) shows the effectivenes of multi-task learning for predicting various mental disorders (including PTSD and depression) based on a multi-layer perceptron (MLP) with word embedding features. While complex deep learning models such as HANs have been used for individual disorder detection (Rao et al., 2020; Mohammadi et al., 2019; Amini and Kosseim, 2020; Zirikly et al., 2019), they have only used word embedding features. More importantly, no previous studies on transfer learning for this task use complex architectures or features, or analyze in detail the types of features that are successful vehicles for knowledge transfer. Especially in the medical domain, using black-box systems can be dangerous for patients and is not a realistic solution (Zucco et al., 2018; Holzinger et al., 2017). Moreover, recently, the need of explanatory systems is required by regulations like the General Data Protection Regulation (GDPR) adopted by the European Union. If any system for mental disorder detection is to be integrated into a tool to assist social media users, it is essential that its decision-making process is understandable in the name of transparency. Additionally, the behavior of powerful classifiers modelling complex patterns in the data has the potential to help uncover manifestations of the disease that are potentially difficult to observe with the naked eye, and thus assist clinicians in the diagnosis process. In the field of mental disorder detection, there are not many studies attempting to explain the behavior of models. We note one such example (Amini and Kosseim, 2020), where the authors analyze attention weights of a neural network trained for automatic anorexia detection. Nevertheless, recent studies have shown the limitations of using attention analysis for interpretability (Wiegreffe and Pinter, 2019; Serrano and Smith, 2019). Burdisso et al. (2019b) introduce a novel text classification model which inherently allows for visual explanations for its predictions, and apply it to the task of depression detection and depression level prediction (Burdisso et al., 2019a; Burdisso et al., 2021), as well as for anorexia and self-harm detection, as part of the eRisk shared tasks (Burdisso et al., 2019c). Uban et al. (2021b) study the automatic detection of depression, anorexia and self-harm from Reddit data using a similar hierarchical architecture with linguistic features and explore the model's explainability through different techniques including attention analysis, different types of error analysis and hidden layer analysis.

In our study, we choose to experiment with models which are inherently interpretable through the multi-aspect representations using different linguistic features, and analyze them in the context of transfer learning. We include ablation studies and error analysis for looking in more depth at what the models are learning, in order to better understand the effectiveness of transfer learning, and the similarities between the different disorders at the linguistic level.

3. Datasets

In order to obtain a wider picture on how mental disorders manifest in social media, we include in our analysis datasets from different sources, containing social media data labelled for several disorders and manifestations thereof: depression, anorexia, self-harm, and PTSD, and gathered from two different social media platforms: Reddit and Twitter. The diversity of the datasets used, both at the level of the targeted disorder, labelling methodology, and text genre allows us to explore the differences in how disorders can be detected automatically for the different types of social media data, and obtain more robust conclusions. While Twitter data are available in large volumes, tweets are limited in length (any tweet can have at most 280 characters) and can restrict the potential for contextual processing. By contrast, Reddit is a platform organized into ”sub-reddits”, or forums for people to discuss common interests, where there is no restriction on text length, and the text associated to a ”post” can either represent a short title or a longer comment. In the following paragraphs we describe the datasets used.

eRisk Reddit datasets on depression, anorexia and self-harm. The eRisk CLEF lab[1] is focused on the early prediction of mental disorder risk from social media data. Shared tasks were organized around specific disorders, including depression (Losada et al., 2018), anorexia and self-harm tendencies (Losada et al., 2019; Losada et al., 2020). Reddit users suffering from a mental disorder are annotated by automatically detecting self-stated diagnoses, followed by a manual curation step. Healthy users are selected from participants in the same sub-reddits (having similar interests), thus making sure the gap between healthy and diagnosed users is not trivially detectable. A long history of posts are collected for the users included in the dataset, up to years prior to the diagnosis. We use all three datasets, corresponding to different disorders: depression, anorexia and self-harm.

CLPsych Twitter dataset on depression and PTSD. CLPsych (Computational Linguistics and Clinical Psychology) is a workshop and shared task organized each year around a different topic concerning computational approaches for mental health. In 2015 (Coppersmith et al., 2015), the shared task challenged participants to detect Twitter users suffering from depression and PTSD. Labelling of the data was done semi-automatically, through an initial selection based on self-stated diagnoses, followed by curation by humans to remove any jokes or disingenuous statements. For each user, their most recent public tweets were included in the dataset. However, other mentions of the disease were not removed from the user's tweets.

Twitter dataset on depression. We include a second Twitter dataset labelled for depression. This dataset was introduced in (Shen et al., 2017), following a similar methodology as the previous ones: depressed Twitter users are selected based on self-stated diagnoses following the pattern ”(I'm/I was/ I am/ I've been) diagnosed depression”. Tweets published within a month of the diagnosis statement were included - this short time frame is an exception compared to the other datasets. Non-depressed users were selected among Twitter users never having posted any tweet containing the character string “depress”. In all datasets, the texts containing the mention of a diagnosis were excluded.

Dataset	**Users**	**Positive%**	**Posts**	**Words**
eRisk depr. (reddit)	1304	16.4%	811,586	25M
eRisk anorexia (reddit)	1287	10.4%	823,754	˜23M
eRisk self-harm (reddit)	763	19%	274,534	˜6M
CLPsych depr. (Twitter)	822	64.1%	1,919,353	˜26M
CLPsych PTSD (Twitter)	1078	72.6%	2,541,214	˜19M
Twitter depr (Shen et al., 2017)	519	50.2%	52,080	˜500K

Table 1: Datasets statistics.

Table 1 contains statistics for all datasets. Some differences between the datasets are observable already from the table: eRisk datasets are more imbalanced, with far more negative than positive cases, while CLPsych and (Shen et al., 2017) datasets are relatively balanced in positive and negative cases. While the total number of words is comparable among datasets, there are more posts per user for the Twitter datasets, while the length of individual posts is shorter. In terms of vocabulary, when looking at the most frequent 20,000 words in each dataset, the overlap between vocabularies for any two of the datasets considered ranges between 55% and 85%, with an average of 68%.

4. Methodology

In the following section we describe the features and models used for our classification experiments.

4.1. Features

As previous studies have shown, mental disorders manifest in language at different levels: topics discussed, emotions conveyed, as well as the author's style. We aim to build multi-dimensional representations of the texts in our datasets to account for the different levels of the language where markers of mental illnesses can manifest, through our selection of various features.

Content features. We include a general representation of text content by transforming each text into word sequences. The obtained sequences will constitute the main input of the recurrent and convolutional layers of our neural networks. Preprocessing includes removing punctuation, numbers, and username mentions from tweets (function words and hashtags are kept). The most frequent 20,000 words in all datasets were selected to form a common vocabulary.

As input to the models, words are encoded as 300-dimensional embeddings, initialized with pre-trained GloVe embeddings.

Style features. We aim at representing the stylistic level of texts through including function word and syntactical features. The usage pattern of function words is known to be reflective of an author's style, at a subconscious level (Argamon and Levitan, 2005). The increased use of first person pronouns, has been shown to correlate with mental disorder risk (Trotzek et al., 2017). Additionally, we expect the differences between texts posted on the two platforms considered (Twitter and Reddit), which are known to have different standards for published posts (the main difference being the 280-character limit on Twitter), might be well captured by stylistic features. We extract from each text a numerical vector representing function words frequencies as bag-of-words. Separately, we include several syntactical features extracted using the LIWC lexicon (including pronoun usage), as described below.

LIWC features. The LIWC lexicon (Pennebaker et al., 2001) has been widely used in computational linguistics as well as some clinical studies for analyzing how suffering from mental disorders manifests in an

[1] https://early.irlab.org/

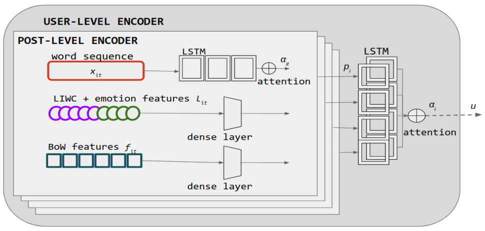

Figure 1: Deep architecture based on HAN.

author's writings. LIWC is a lexicon mapping words of the English vocabulary to lexico-syntactic features of different kinds, with high quality associations curated by human experts and capturing different levels of language: including style (through syntactic categories), emotions (through affect categories) and topics (through content-oriented categories, referring to cognitive processes, or to topics such as money, health or religion). We include in our analysis all 64 categories in LIWC 2015 (Pennebaker et al., 2015).

Emotions and sentiment. We dedicate a few features to representing emotional content in our texts, since the emotional state of a user is known to be highly correlated with her mental health. Aside from some LIWC categories designed to capture emotional content (e.g. *negative emotions*, *positive emotions*, *sadness*, *anxiety*), we additionally include a second lexicon: the NRC emotion lexicon (Mohammad and Turney, 2013), which is dedicated exclusively to emotion representation, containing 10 categories based on Plutchik's emotions (Plutchik, 1991): *anger*, *anticipation*, *disgust*, *fear*, *joy*, *negative*, *positive*, *sadness*, *surprise*, *trust*. For both lexicons, we build features as numerical vectors by computing for each category the ratio of words in a text that are related to the category.

4.2. Experimental Setup

All tasks approached are modelled as supervised classification tasks, with labels at the user level. We follow the typical machine learning workflow, and split each of our datasets into training, validation and test subsets. For datasets released as part of a shared task (eRisk and CLPsych data), we maintain the original train/test split provided by the shared task organizers; for the rest, we use a random 70/30 split. Train and test sets are disjoint at the user level. Social media posts are not considered individually as datapoints, since (according to preliminary experiments) they are too short to be sufficiently predictive. Instead, we generate our datapoints by grouping sequences of 50 chronologically consecutive posts into larger chunks.

4.3. Architecture

We propose using a sophisticated deep architecture based on HANs with multiple linguistic features. Previous work showed improved performance for individual disorder detection compared to other baselines reported including a sequential BiLSTM, pretrained transformers, and logistic regression (Uban et al., 2021a; Uban et al., 2021b). Additionally, using this multi-aspect model allows us to gain a better understanding of the importance of individual features and linguistic levels on the classification performance, and on the particular differences between the disorders, which we will explore in more detail in Section 7. Configuration details of all models are included in the Appendix, we make the code publicly available[2].

Hierarchical attention networks for text classification were introduced in (Yang et al., 2016). We assume that social media data are well suited to a hierarchical representation; in our case the hierarchy consists of user post histories, which are composed of social media posts, composed of word sequences. The hierarchical network is made of two components: a *post-level encoder* , which produces a representation of a post, and a *user-level encoder* (modelled as an LSTM), which generates a representation of a user's post history. The post-level encoder and the user-level encoder are modelled as LSTMs. The word sequences encoded as embeddings are passed to the post-level LSTM and the output is concatenated with the other features to form the hierarchical post encoding. The obtained representation is passed to the user-encoder LSTM, which is connected to the output layer. A depiction of the hierarchical architecture is shown in Figure 1.

5. Classification among Disorders

As a way to provide an initial understanding of the similarity between the targeted disorders from the perspective of their manifestations in language, we perform experiments to automatically classify among the different disorders. This is bound to be a more difficult problem than the distinction between healthy users and those suffering from the disorder - the similarity between the linguistic patterns across disorders is in itself the premise for the effective use of transfer learning.

We use the same HAN model, in a multi-class, multi-label classification task (using a sigmoid activation for the final layer), taking into account the fact that in principle some users might be suffering from multiple disorders, even though our datasets don't contain examples with multiple labels, or any information about the overlap between users with different disorders. We exclude the healthy users, and separately classify among disorders from the two data sources: depression, anorexia, and self-harm for Reddit, and depression and PTSD for Twitter. We use weighted sampling in order to balance the classes. For the Reddit datasets, we obtain an accuracy of 0.44, and 0.44 macro F1-score. The confusion matrix, illustrated in Table 2 shows that classification is most accurate for depression, while self-harm is most difficult to distinguish from the other disorders. This coincides with the binary classification results, where self-harm was the most difficult to classify. In the case of Reddit users suffering from depression, while they are not very easy to distinguish from healthy users (compared to people suffering

[2] https://github.com/ananana/mental-disorders

	Prediction		
	D	S-H	A
Label			
D	139	2	113
S-H	60	67	144
A	201	16	218

	Prediction	
	D	PTSD
Label		
D	126	24
PTSD	65	95

Table 2: Disorder classification confusion matrices for depression (D) vs. self-harm (S-H) vs. anorexia (A) (reddit); and depression (D) vs. PTSD (Twitter).

from anorexia for example), they are easier to distinguish from users suffering from other disorders. For the Twitter datasets, we obtain an F1-score of 0.72 for distinguishing between depression and PTSD, and an AUC of 0.75. These results suggest that PTSD might be easier to distinguish from depression than anorexia or self-harm are. This is not surprising, since, unlike the other three disorders, PTSD implies a triggering event, and has a different effect on the affected person's self-concept, which could also appear at the linguistic level. On the other hand, depression and anorexia, for example, have overlapping symptoms and often cooccur in the same patients (American Psychiatric Association, 2013).

6. Transfer Learning Experiments

We further apply transfer learning in order to understand compatibility between disorders (similarity between the manifestation of the different disorders in texts on social media) and between datasets from different social media platforms, as well as to attempt enhancing performance on individual disorder prediction. We use our multi-aspect HAN model for the following experiments, and compare the transfer learning results with baseline results obtained on individual disorder classification using the same model architecture.

6.1. Transfer Learning Strategies

Strategy 0. Zero-shot. As a baseline, we assess the cross-disorder performance on models before any transfer learning, by simply training a model on one dataset and testing it on another. The results of these experiments also give us an initial understanding of how compatible two tasks are, as measured by how well a model trained on data for one task can predict labels of a different class without any specific training.

Strategy 1. Transfer layer. As the first transfer learning strategy, we append an additional dense layer connected to the final layer of the model, which will serve as a *transfer layer*. All weights except for the transfer layer weights are frozen (not trainable) in this second training phase, and initialized with the weights of the pre-trained model on the original task.

Strategy 2. Fine-tuning. As a second strategy, we employ fine-tuning of pre-trained models, which consists of training a model initialized with weights of a different model of the same architecture that was previously trained for a different task. Training continues then on the new target task, with a lower learning rate, for all parameters.

Strategy 3. Multi-task learning. Finally, we experiment with applying multi-task learning, which consists of training a single model jointly on different types of labels, corresponding to different tasks or types of data. In this setup, all layers are shared (and trainable), except for a final dense layer which is task-specific.

Cross-disorder transfer. Our goal is to study if and how knowledge can be transferred between disorders. Transferring knowledge from disorders for which more data is available (such as for depression), to other disorders where data is more scarce (such as anorexia, PTSD or self-harm) could help increase the performance for the latter. We experiment with transferring knowledge from our depression datasets to the smaller datasets corresponding to the other disorders. We apply the transfer layer strategy (*Strategy 1*) and the fine-tuning strategy (*Strategy 2*) between datasets different disorders but from the same platform (so as to isolate the cross-disorder problem from the cross-platform one).

Cross-platform transfer. It would additionally be useful to understand whether models can be used cross-platform: from one type of data to another. Nowadays multiple social media platforms are being used, and new ones appear every day, with each having different standards and typical ways of engaging with other users. In our case, we have datasets for the same disorder (depression) collected from different platforms (Reddit and Twitter) and using different methodologies. We study how knowledge can be transferred across different data genres (social media platforms) using the first two strategies to transfer knowledge from a source dataset (the eRisk dataset) to other datasets on the same task; separately we use multi-task learning to jointly train a model on all three depression datasets, and assess its performance on each. Table 3 summarizes results for cross-disorder and cross-platform transfer learning experiments using the transfer layer and the fine-tuning strategy. In the cased of zero-shot learning, (*Strategy 0*) the best AUC score is obtained for self-harm detection, in line with the pattern of confusion between self-harm and depression seen in the 3-way classification experiments, and with the relation between depression and suicidal tendencies.The high jump in performance for all experiments between results for *Strategy 0* and *Strategy 1* shows that transfer of knowledge between the disorders we consider is achievable to a high degree by training very few parameters, reaching performances close to those obtained by training the full models for the target task, and in the case of the cross-disorder transfer between depression and self-harm, it even exceeds it (0.87 AUC compared to 0.83 AUC). Self-harm is also the smallest of the Reddit datasets, and the disorder which was the most difficult to distinguish in cross-disorder classification - premises which support the utility of transfer learning in this case. Results of fine-tuning experiments (*Strategy 2*) show that the baseline performance of models trained for a disorder can be enhanced through trans-

	CROSS-DISORDER						**CROSS-PLATFORM**			
Source	**eRisk depression**				**CLPsych depression**		**eRisk depression**			
Target	**eRisk Anorexia**		**eRisk Self-harm**		**CLPsych PTSD**		**Shen et al. depression**		**CLPsych depression**	
	F1	**AUC**	**F1**	**AUC**	**F1**	**AUC**	**F1**	**AUC**	**F1**	**AUC**
Strategy 0	.17	.62	.13	.69	.31	.60	.69	.59	.38	.57
Strategy 1	.64	.90	.54	**.87**	.43	.73	.65	.74	.61	.72
Strategy 2	.63	**.93**	.67	**.87**	.58	**.78**	.86	**.94**	.60	**.74**
Baseline HAN	.46	.91	.51	.83	.57	.70	.77	.81	.53	.73

Table 3: Cross-disorder and cross-platform transfer learning results, compared to individual disorder prediction.

fer learning on related disorders: for all disorders and datasets, performance is improved compared to the results for the same model in a single-task setting, at least for one metric. The largest improvement is obtained for depression on the Twitter dataset (Shen et al., 2017), which might benefit from transfer learning since it is the smallest dataset.

Source	**All depression**					
Target	**eRisk**		**Shen et al.**		**CLPsych**	
	F1	**AUC**	**F1**	**AUC**	**F1**	**AUC**
Strategy 3	.39	.81	.74	**.83**	.56	**.82**
Single-task	.44	**.86**	.77	.81	.53	.73

Table 4: Cross-platform multi-task learning results.

In Table 4 we show the performance of the model after being trained in a multi-task setting on the three depression datasets. Multi-task learning seems to benefit especially the Twitter datasets, while for the Reddit dataset performance is not improved. The superior results of cross-disorder compared to cross-platform experiments confirm previous findings on the difficulty of cross-domain transfer for depression detection (Harrigian et al., 2020). In general, across strategies, cross-platform transfer from Reddit to Twitter data has the largest impact on performance, which might be due to the longer texts typically found on Reddit.

For depression and PTSD detection, we compare our results to a previous study using MTL (Benton et al., 2017) for mental disorder detection. The authors trained a MLP model and evaluated it on the same CLPsych datasets. We show superior improvements after employing MTL, with the best models reaching or exceeding 0.78 AUC for both disorders (compared to 0.78 for PTSD and around 0.75 for depression in the previous study). The best performance across tasks reported in this study is 0.84 AUC for suicide ideation detection, whereas several of our models exceed this score. For the other disorders, there are no previous studies, to our knowledge, that use cross-disorder transfer. Anorexia detection was approached within the 2019 eRisk shared task (Losada et al., 2019), with the best team obtaining a 0.70 F1-score (Mohammadi et al., 2019). We obtain similar results with the RoBERTa transformer. Most previous literature on self-harm detection consists of the solutions submitted in the eRisk 2019 shared task (Losada et al., 2019), where self-harm detection was framed as an unsupervised task, and the best team obtained 0.52 F1 scores (using a comparable strategy based on training a model on an external dataset), whereas we obtain results up to 0.67 F1 with our best model.

Source	**eRisk**				**CLPsych**	
Target	**Anorexia**		**Self-harm**		**PTSD**	
	F1	**AUC**	**F1**	**AUC**	**F1**	**AUC**
All-word seq	.49	.88	.24	.77	.57	.74
All-function words	.51	.90	.61	.83	.57	.77
All-lexicon feat	.50	.91	.42	.81	.54	.75
All features	.63	**.93**	.67	**.87**	.58	**.78**

Table 5: Ablation for cross-disorder transfer, using different depression datasets as source data.

7. Interpretability

7.1. Ablation Experiments

In the previous sections we have shown that transfer learning can be effectively used to improve classification of mental disorders, suggesting as well that the way mental disorders manifest in language show similar patterns across disorders. Our choice of features extracted to represent different levels of the language allows us to isolate the features used in order to attempt to explain which features are most useful as vehicles for knowledge transfer between disorders.

In order to answer this question, we perform a series of ablation experiments, where we ignore each type of feature one by one (along with the hidden layers encoding it), as a way to measure its impact on the effectiveness of the knowledge transfer. Based on the HAN model, we remove each of the features one by one: first the word sequences (along with the post-level LSTM layers), then the bag-of-function-words feature, and finally the lexicon features. We then train the partial model on the source task, then fine-tune it on the target task (*Strategy 2*). Results are shown in Table 5. We notice that using all features is generally the optimal strategy, proving that including multi-aspect features is useful for successful transfer learning. Among them, removing word sequences significantly reduces the performance of the transferred model across disorders, suggesting they are the most effective at capturing common patterns. Lexicon features (reflecting emotions and other psycho-linguistic categories) are more useful than function words in the case of self-

Experiment	Psycho-linguistic categories (LIWC features)	Emotions (NRC features)
Depression (eRisk) baseline	verbs, tentative, *I* (1st pers pron), adverbs, past tense, pronouns, present tense, conjunctions	fear, anger, negative emotion, sadness
Self-harm baseline	health, insight, cognitive processes, pronouns function words, adverbs	sadness, negative emotion
Anorexia baseline	future tense, positive emotion, affective, function words, adverbs, present tense, pronouns	anger, fear, negative emotion
PTSD baseline	they (3rd pers pron), health, insight, she/he	fear, joy, positive emotion, negative emotion, sadness
Depr→self-harm transfer	*you* (2nd pers pron), function words, impersonal pronouns, verbs	positive emotion
Depr→anorexia transfer	future tense, affective, function words, adverbs, present tense, *I* (1st pers pron), verbs, social	fear, negative emotion
Depr→PTSD transfer	exclusive, sad, conjunctions, adverbs, friend, biology	anger, positive emotion, sadness

Table 6: Features with highest differences between correctly classified and misclassified texts.

harm and PTSD, suggesting these affective and psychological features encode some commonalities between these disorders better than the stylistic level. As a reference baseline, we also include ablation experiments performed on the original trained models prior to transfer learning. Results are listed in the Appendix, showing similar patterns of feature importance, with word sequences and lexicons features proving most useful. One interesting effect is the importance of function words for detecting anorexia, self-harm and PTSD: in the baseline results, the ablation experiments show it is not a useful feature for predicting these disorders, while in the transfer learning results, the behavior changes - removing the function words feature harms the prediction performance. This suggests that pre-training the models on the depression detection task helps learn patterns in function word distribution which are useful for predicting the target disorders.

7.2. Error Analysis

We attempt to understand what causes misclassifications by comparing correctly versus incorrectly classified examples in terms of the metrics computed using lexicons, which are the most easily interpretable features in our multi-aspect models. The features for which the misclassified examples and correctly classifed examples differ most significantly ($p \leq 0.05$) on average are shown in Table 6 for each experiment (including baselines and cross-disorder transfer learning experiments using strategy 2). We notice a high incidence of grammatical categories across all experiments, suggesting that inclusion of a feature to model part-of-speech distribution could improve performance. Our results also help us identify some aspects of the language used by people suffering from different disorders which become more accurately modelled by the trained neural networks after transfer learning, such as the expression of some emotions: for example sadness in the case of self-harm, anger for anorexia, and fear and joy for PTSD. These results could suggest that the identified emotions show similar patterns between depression and the different target disorders, which the models learn to better identify after transfer learning.

8. Conclusions

In this study we have explored the issue of automatic prediction of mental disorders from social media data using transfer learning, from both cross-disorder and cross-platform perspectives. We are the first to experiment with deep learning architectures including different types of linguistic features for transfer learning, and we compare different transfer strategies, demonstrating compatibility between the manifestation of the different disorders at the level of language, and showing that transfer learning could be leveraged as a solution for building models in the case of disorders where annotated data is scarce (RQ1). The results of the different transfer learning experiments and of the cross-disorder classification experiments show that we can find some common features between any of the studied disorders, but some disorders (such as self-harm and depression) show more similarity than others (such as PTSD and depression), confirming existing evidence of co-morbidity between certain disorders (RQ2). At the level of cross-platform transfer, we show that it can provide performance improvements in the case of transferring knowledge from data sources where texts are longer (e.g. Reddit) to datasets with shorter writings (such as Twitter), whereas the reverse is not as effective (RQ1). We have included features to capture different linguistic levels, which can be useful for a better understanding of how mental disorders manifest in language, and performed ablation experiments to understand which features are most useful for transfer learning (RQ3). We provided some interpretability with regards to feature importance through ablation experiments, and through feature-based error analysis.

In the future, it could be promising to combine the different linguistic features with transformer-based representations, for example by integrating pre-trained sentence embeddings as the HAN's post-level sentence encoder. Moreover, multi-modal solutions have rarely been used in computational studies related to mental health (Ramírez-Cifuentes et al., 2020; Liu et al., 2016; Guntuku et al., 2019), and to our knowledge they have not been explored in the context of transfer learning. Finally, extending the analysis to additional disorders with known co-morbidities would help with better understanding the overlap between certain disorders' definitions and manifestations.

Ethical Considerations

All the data we use is anonymized, sourced from external research groups according to each dataset's respective data usage policy. We have obtained our ethical board's approval for carrying out our project.

Acknowledgements

The authors thank the EU-FEDER Comunitat Valenciana 2014-2020 grant IDIFEDER/2018/025. The work of Paolo Rosso was in the framework of the research project PROMETEO/2019/121 (DeepPattern) by the Generalitat Valenciana. The work of Ana Uban was partially funded by a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI—UEFISCDI, project number 108, PNCDI III.

Bibliographical References

Abd Yusof, N. F., Lin, C., and Guerin, F. (2017). Analysing the causes of depressed mood from depression vulnerable individuals. In *Proceedings of the International Workshop on Digital Disease Detection using Social Media 2017 (DDDSM-2017)*, pages 9–17.

Abed-Esfahani, P., Howard, D., Maslej, M., Patel, S., Mann, V., Goegan, S., and French, L. (2019). Transfer learning for depression: Early detection and severity prediction from social media postings. In *L. Cappellato, N. Ferro, D. Losada and H. Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings.CEUR-WS.org*, volume 2380.

Allan, C. E., Valkanova, V., and Ebmeier, K. P. (2014). Depression in older people is underdiagnosed. *The Practitioner*, 258(1771):19—22, 2—3, May.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5*. Autor, Washington, DC, 5th ed. edition.

Amini, H. and Kosseim, L. (2020). Towards explainability in using deep learning for the detection of anorexia in social media. In *International Conference on Applications of Natural Language to Information Systems*, pages 225–235. Springer.

Aragón, M. E., Monroy, A. P. L., González-Gurrola, L. C., and Montes, 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, volume 1 (long and short papers)*, pages 1481–1486.

Argamon, S. and Levitan, S. (2005). Measuring the usefulness of function words for authorship attribution. In *Proceedings of the 2005 ACH/ALLC Conference*, pages 4–7.

Benton, A., Mitchell, M., and Hovy, D. (2017). Multitask learning for mental health using social media text. *arXiv preprint arXiv:1712.03538*.

Bridge, J. A., Goldstein, T. R., and Brent, D. A. (2006). Adolescent suicide and suicidal behavior. *Journal of child psychology and psychiatry*, 47(3-4):372–394.

Bucur, A.-M., Podină, I. R., and Dinu, L. P. (2021). A psychologically informed part-of-speech analysis of depression in social media. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 199–207.

Burdisso, S., Errecalde, M. L., and Montes y Gómez, M. (2019a). Towards measuring the severity of depression in social media via text classification. In *XXV Congreso Argentino de Ciencias de la Computación (CACIC)(Universidad Nacional de Río Cuarto, Córdoba, 14 al 18 de octubre de 2019)*.

Burdisso, S. G., Errecalde, M., and Montes-y Gómez, M. (2019b). A text classification framework for simple and effective early depression detection over social media streams. *Expert Systems with Applications*, 133:182–197.

Burdisso, S. G., Errecalde, M., and Montes-y Gómez, M. (2019c). Unsl at erisk 2019: a unified approach for anorexia, self-harm and depression detection in social media. In *CLEF (Working Notes)*.

Burdisso, S. G., Errecalde, M. L., and Montes y Gómez, M. (2021). Using text classification to estimate the depression level of reddit users. *Journal of Computer Science & Technology*, 21.

Calvo, R. A., Milne, D. N., Hussain, M. S., and Christensen, H. (2017). Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, 23(5):649–685.

Caruana, R. (1997). Multitask learning. *Machine learning*, 28(1):41–75.

Cavanagh, J. T., Carson, A. J., Sharpe, M., and Lawrie, S. M. (2003). Psychological autopsy studies of suicide: a systematic review. *Psychological medicine*, 33(3):395.

Chen, X., Sykora, M. D., Jackson, T. W., and Elayan, S. (2018). What about mood swings: Identifying depression on twitter with temporal measures of emotions. In *Companion Proceedings of the The Web Conference 2018*, pages 1653–1660.

Cheng, X., Wang, X., Ouyang, T., and Feng, Z. (2020). Advances in emotion recognition: Link to depressive disorder. In *Neurological and Mental Disorders*. IntechOpen.

Cohan, A., Desmet, B., Yates, A., Soldaini, L., MacAvaney, S., and Goharian, N. (2018). Smhd: a large-scale resource for exploring online language usage for multiple mental health conditions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1485–1497.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(ARTICLE):2493–2537.

Coppersmith, G., Dredze, M., and Harman, C. (2014). Quantifying mental health signals in twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 51–60.

Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., and Mitchell, M. (2015). Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39.

De Choudhury, M., Gamon, M., Counts, S., and Horvitz, E. (2013). Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*.

De Choudhury, M., Counts, S., Horvitz, E. J., and Hoff, A. (2014). Characterizing and predicting postpartum depression from shared facebook data. In

Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing, pages 626–638.

Eichstaedt, J. C., Smith, R. J., Merchant, R. M., Ungar, L. H., Crutchley, P., Preoţiuc-Pietro, D., Asch, D. A., and Schwartz, H. A. (2018). Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208.

Gkotsis, G., Oellrich, A., Velupillai, S., Liakata, M., Hubbard, T. J., Dobson, R. J., and Dutta, R. (2017). Characterisation of mental health conditions in social media using informed deep learning. *Scientific reports*, 7(1):1–11.

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

Guntuku, S. C., Preotiuc-Pietro, D., Eichstaedt, J. C., and Ungar, L. H. (2019). What twitter profile and posted images reveal about depression and anxiety. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 236–246.

Harrigian, K., Aguirre, C., and Dredze, M. (2020). Do models of mental health based on social media data generalize? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 3774–3788.

Holzinger, A., Biemann, C., Pattichis, C. S., and Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain? *arXiv preprint arXiv:1712.09923*.

Kaufman, J. and Charney, D. (2000). Comorbidity of mood and anxiety disorders. *Depression and anxiety*, 12(S1):69–76.

Leiva, V. and Freire, A. (2017). Towards suicide prevention: early detection of depression on social media. In *International Conference on Internet Science*, pages 428–436. Springer.

Lima, C. K. T., de Medeiros Carvalho, P. M., Lima, I. d. A. S., de Oliveira Nunes, J. V. A., Saraiva, J. S., de Souza, R. I., da Silva, C. G. L., and Neto, M. L. R. (2020). The emotional impact of coronavirus 2019-ncov (new coronavirus disease). *Psychiatry research*, page 112915.

Liu, L., Preotiuc-Pietro, D., Samani, Z. R., Moghaddam, M. E., and Ungar, L. (2016). Analyzing personality through social media profile picture choice. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 10.

Losada, D. E., Crestani, F., and Parapar, J. (2018). Overview of erisk: early risk prediction on the internet. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 343–361. Springer.

Losada, D. E., Crestani, F., and Parapar, J. (2019). Overview of erisk 2019 early risk prediction on the internet. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 340–357. Springer.

Losada, D., Crestani, F., and Parapar, J. (2020). Overview of erisk 2020: Early risk prediction on the internet. In *L. Cappellato, C. Eickhoff, N. Ferro, and A. Névéol (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings.CEUR-WS.org*, volume 2696.

Matero, M., Idnani, A., Son, Y., Giorgi, S., Vu, H., Zamani, M., Limbachiya, P., Guntuku, S. C., and Schwartz, H. A. (2019). Suicide risk assessment with multi-level dual-context language and bert. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 39–44.

Mitchell, M., Hollingshead, K., and Coppersmith, G. (2015). Quantifying the language of schizophrenia in social media. In *Proceedings of the 2nd workshop on Computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 11–20.

Mohammad, S. M. and Turney, P. D. (2013). Nrc emotion lexicon. *National Research Council, Canada*, 2.

Mohammadi, E., Amini, H., and Kosseim, L. (2019). Quick and (maybe not so) easy detection of anorexia in social media posts. In *L. Cappellato, N. Ferro, D. Losada and H. Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings.CEUR-WS.org*, volume 2380.

O'dea, B., Wan, S., Batterham, P. J., Calear, A. L., Paris, C., and Christensen, H. (2015). Detecting suicidality on twitter. *Internet Interventions*, 2(2):183–188.

Orabi, A. H., Buddhitha, P., Orabi, M. H., and Inkpen, D. (2018). Deep learning for depression detection of twitter users. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 88–97.

O'Dea, B., Boonstra, T. W., Larsen, M. E., Nguyen, T., Venkatesh, S., and Christensen, H. (2021). The relationship between linguistic expression in blog content and symptoms of depression, anxiety, and suicidal thoughts: A longitudinal study. *Plos one*, 16(5):e0251787.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). The development and psychometric properties of liwc2015. Technical report.

Plana-Ripoll, O., Pedersen, C. B., Holtz, Y., Benros, M. E., Dalsgaard, S., De Jonge, P., Fan, C. C., Degenhardt, L., Ganna, A., Greve, A. N., et al. (2019). Exploring comorbidity within mental disorders among a danish national population. *JAMA psychiatry*, 76(3):259–270.

Plutchik, R. (1991). *The emotions*. University Press of America.

Ramı́rez-Cifuentes, D., Largeron, C., Tissier, J., Freire, A., and Baeza Yates, R. (2020). Enhanced word embeddings for anorexia nervosa detection on social media. In *Berthold M, Feelders A, Krempl G, editors. Advances in Intelligent Data Analysis XVIII. 18th International Symposium on Intelligent Data Analysis, IDA 2020 Proceedings; 2020 Apr 27-29; Konstanz, Germany. Cham: Springer; 2020. p. 404-17.(LNCS; no. 12080)*. Springer.

Ramírez-Cifuentes, D., Mayans, M., and Freire, A. (2018). Early risk detection of anorexia on social media. In *International Conference on Internet Science*, pages 3–14. Springer.

Ramírez-Cifuentes, D., Freire, A., Baeza-Yates, R., Puntí, J., Medina-Bravo, P., Velazquez, D. A., Gonfaus, J. M., and Gonzàlez, J. (2020). Detection of suicidal ideation on social media: multimodal, relational, and behavioral analysis. *Journal of medical internet research*, 22(7):e17758.

Rao, G., Zhang, Y., Zhang, L., Cong, Q., and Feng, Z. (2020). Mgl-cnn: A hierarchical posts representations model for identifying depressed individuals in online forums. *IEEE Access*, 8:32395–32403.

Resnik, P., Garron, A., and Resnik, R. (2013). Using topic modeling to improve prediction of neuroticism and depression in college students. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1348–1353.

Sadeque, F., Xu, D., and Bethard, S. (2017). Uarizona at the clef erisk 2017 pilot task: linear and recurrent models for early depression detection. In *CEUR workshop proceedings*, volume 1866. NIH Public Access.

Saha, K., Sugar, B., Torous, J., Abrahao, B., Kıcıman, E., and De Choudhury, M. (2019). A social media study on the effects of psychiatric medication use. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 440–451.

Sawhney, R., Joshi, H., Shah, R., and Flek, L. (2021). Suicide ideation detection via social and temporal user representations using hyperbolic learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2176–2190.

Seabrook, E. M., Kern, M. L., Fulcher, B. D., and Rickard, N. S. (2018). Predicting depression from language-based emotion dynamics: longitudinal analysis of facebook and twitter status updates. *Journal of medical Internet research*, 20(5):e9267.

Serrano, S. and Smith, N. A. (2019). Is attention interpretable? In *ACL*, pages 2931–2951.

Shah, K., Kamrai, D., Mekala, H., Mann, B., Desai, K., and Patel, R. S. (2020). Focus on mental health during the coronavirus (covid-19) pandemic: applying learnings from the past outbreaks. *Cureus*, 12(3).

Sheehan, D. V. (2004). Depression: underdiagnosed, undertreated, underappreciated. *Managed care (Langhorne, Pa.)*, 13(6 Suppl Depression):6—8, June.

Shen, J. H. and Rudzicz, F. (2017). Detecting anxiety through reddit. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality*, pages 58–65.

Shen, G., Jia, J., Nie, L., Feng, F., Zhang, C., Hu, T., Chua, T.-S., and Zhu, W. (2017). Depression detection via harvesting social media: A multimodal dictionary learning solution. In *IJCAI*, pages 3838–3844.

Shen, T., Jia, J., Shen, G., Feng, F., He, X., Luan, H., Tang, J., Tiropanis, T., Chua, T. S., and Hall, W. (2018). Cross-domain depression detection via harvesting social media. International Joint Conferences on Artificial Intelligence.

Shigemura, J., Ursano, R. J., Morganstein, J. C., Kurosawa, M., and Benedek, D. M. (2020). Public responses to the novel 2019 coronavirus (2019-ncov) in japan: Mental health consequences and target populations. *Psychiatry and clinical neurosciences*, 74(4):281.

Søgaard, A. and Goldberg, Y. (2016). Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 231–235.

Surís, A., Holliday, R., and North, C. S. (2016). The evolution of the classification of psychiatric disorders. *Behavioral Sciences*, 6(1):5.

Sutton, C., McCallum, A., and Rohanimanesh, K. (2007). Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. *Journal of Machine Learning Research*, 8(Mar):693–723.

Torales, J., O'Higgins, M., Castaldelli-Maia, J. M., and Ventriglio, A. (2020). The outbreak of covid-19 coronavirus and its impact on global mental health. *International Journal of Social Psychiatry*, page 0020764020915212.

Trotzek, M., Koitka, S., and Friedrich, C. M. (2017). Linguistic metadata augmented classifiers at the clef 2017 task for early detection of depression. In *L. Cappellato, N. Ferro, L. Goeuriot and T. Mandl (eds.) CLEF 2017 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings.CEUR-WS.org*, volume 1866.

Trotzek, M., Koitka, S., and Friedrich, C. M. (2018). Word embeddings and linguistic metadata at the clef 2018 tasks for early detection of depression and anorexia. In *L. Cappellato, N. Ferro, J. Nie and L. Soulier (eds.) CLEF 2018 Labs* *and Workshops, Notebook Papers. CEUR Workshop Proceedings.CEUR-WS.org*, volume 2125.

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

Uban, A. S., Chulvi, B., and Rosso, P. (2021b). On the explainability of automatic predictions of mental disorders from social media data. In *International Conference on Applications of Natural Language to Information Systems*, pages 301–314. Springer.

Wang, Y.-T., Huang, H.-H., and Chen, H.-H. (2018). A neural network approach to early risk detection of depression and anorexia on social media text. In *L. Cappellato, N. Ferro, J. Nie and L. Soulier (eds.) CLEF 2018 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings.CEUR-WS.org*, volume 2125.

Wiegreffe, S. and Pinter, Y. (2019). Attention is not not explanation. In *EMNLP-IJCNLP*, pages 11–20.

World Health Organization, W. (2012). Depression: A global crisis. world mental health day, october 10 2012. *World Federation for Mental Health, Occoquan, Va, USA*.

Xiang, Y.-T., Yang, Y., Li, W., Zhang, L., Zhang, Q., Cheung, T., and Ng, C. H. (2020). Timely mental health care for the 2019 novel coronavirus outbreak is urgently needed. *The Lancet Psychiatry*, 7(3):228–229.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Yates, A., Cohan, A., and Goharian, N. (2017). Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978.

Yazdavar, A. H., Al-Olimat, H. S., Ebrahimi, M., Bajaj, G., Banerjee, T., Thirunarayan, K., Pathak, J., and Sheth, A. (2017). Semi-supervised approach to monitoring clinical depressive symptoms in social media. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 1191–1198.

Zirikly, A., Resnik, P., Uzuner, O., and Hollingshead, K. (2019). Clpsych 2019 shared task: Predicting the degree of suicide risk in reddit posts. In *Proceedings of the sixth workshop on computational linguistics and clinical psychology*, pages 24–33.

Zucco, C., Liang, H., Di Fatta, G., and Cannataro, M. (2018). Explainable sentiment analysis with applications in medicine. In *IEEE BIBM*, pages 1740–1747. IEEE.

Appendix

A. Hyperparameters

- LSTM units (post encoder) = 128
- dense BoW units = 20
- dense lexicon units = 20
- CNN filters (post encoder) = 100
- CNN kernel size (post encoder) = 5
- LSTM units (user encoder) = 32
- dropout = 0.0
- l_2 = 0.00001
- optimizer = Adam
- learning rate = 0.0001
- early stopping patience = 20
- epochs = 20
- maximum sequence length = 256
- posts per chunk = 50

B. Additional Metrics for Experiment Results

Source	**All depression**					
Target	**eRisk**		**(Shen et al.)**		**CLPsych**	
	P	**R**	**P**	**R**	**P**	**R**
Strategy 3	.46	.37	.60	.84	.71	.49

Table 9: Precision and recall for cross-genre multi-task learning results for depression detection.

	SELF-HARM		**ANOREXIA**		**DEPRESSION**						**PTSD**	
	eRisk		**eRisk**		**eRisk**		**(Shen et al)**		**CLPsych**		**CLPsych**	
Model	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
HAN	.67	.44	.44	.57	.71	.57	.65	.95	.48	.63	.51	.67

Table 7: Precision and recall scores for all datasets and models trained on individual tasks.

	CROSS-TASK						**CROSS-GENRE**			
Source	**eRisk depression**				**CLPsych depression**		**eRisk depression**			
Target	**eRisk Anorexia**		**eRisk Self-harm**		**CLPsych PTSD**		**(Shen et al.) depression**		**CLPsych depression**	
	P	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
Strategy 0	.09	.92	.35	.11	.37	.28	.54	.97	.35	.44
Strategy 1	.71	.64	.41	.85	.49	.40	.60	.71	.48	.41
Strategy 2	.43	.71	.71	.57	.48	.63	.87	.86	.50	.82

Table 8: Precision and recall scores for cross-task and cross-genre transfer learning.

Source	**eRisk depression**				**CLPsych depression**	
Target	**eRisk Anorexia**		**eRisk Self-harm**		**CLPsych PTSD**	
	P	**R**	**P**	**R**	**P**	**R**
All-word seq	.45	.61	.55	.16	.44	.85
All-function words	.70	.56	.61	.65	.53	.63
All-lexicon feat	.44	.67	.75	.30	.49	.63
All features	.43	.71	.71	.57	.48	.63

Table 11: Precision and recall scores for ablation experiments, using HAN.

	Depression			**Anorexia**			**Self-harm**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Depression/anorexia/self-harm (reddit)	.23	.61	.33	.53	.50	.51	.82	.22	.34
Depression/PTSD (Twitter)	.65	.84	.72	-	-	-	-	-	-

Table 12: Precision, recall and F1 scores per class, for classification among disorders: multi-class classification for depression/anorexia/self-harm or reddit dataset, and binary classification on depression/PTSD on Twitter dataset (depression considered the positive class).

	SELF-HARM		**ANOREXIA**		**DEPRESSION**				**PTSD**	
	eRisk		**eRisk**		**eRisk**		**CLPsych**		**CLPsych**	
Model	**F1**	**AUC**	**F1**	**AUC**	**F1**	**AUC**	**F1**	**AUC**	**F1**	**AUC**
All-word seq	.34	.83	.47	.85	.34	.79	.51	.68	.50	.53
All-function words	.55	.84	.46	.93	.50	.81	.53	.69	.57	.71
All-lexicon feat	.59	.79	.57	.90	.43	.84	.50	.65	.60	.72
All features	.51	.83	.46	.91	.44	.86	.53	.73	.57	.70

Table 10: F1 and AUC scores for ablation experiments on individual disorder classification with HAN (no transfer learning).

C. Error Analysis - Additional Details for Distinctive Features of Misclassifications

Emotion/ category	Mean value correct classif	Mean value misclassif	p-value	Emotion/ category	Mean value correct classif	Mean value misclassif	p-value
space	.053	.048	5.38e-6	insight	.018	.021	2.46e-6
verb	.182	.235	4.50e-30	tentat	.022	.027	8.80e-15
body	.006	.008	.001	death	.003	.002	.0006
quant	.022	.027	9.40e-11	excl	.020	.025	2.42e-14
i	.030	.049	2.48e-34	achieve	.017	.015	9.37e-7
adverb	.038	.048	1.37e-17	preps	.0103	.097	1.54e-5
past	.029	.037	2.07e-17	ppron	.064	.092	9.38e-38
present	.145	.189	2e-27	conj	.040	.047	1.35e-10
they	.005	.006	.002	cogmech	.127	.145	1.1e-15
pronoun	.110	.151	3.75e-38	discrep	.012	.016	6.34e-14
assent	.006	.008	5.97e-6	shehe	.007	.010	1.64e-6
incl	.028	.030	.001	anger	.017	.015	.001
future	.007	.103	2.95e-31	feel	.004	.006	.0001
auxverb	.136	.178	1.12e-30	ipron	.046	.058	2.97e-21
motion	.015	.017	.006	swear	.003	.004	.0001
money	.008	.007	.002	affect	.063	.069	6.21e-5
sexual	.004	.006	.001	nonfl	.0012	.0018	1.90e-5
negate	.018	.023	8.02e-12	you	.016	.022	1.22e-10
work	.024	.018	6.37e-11	bio	.020	.022	.034
social	.074	.086	5.71e-12	posemo	.040	.046	0.0002
funct	.426	.493	2.26e-27	percept	.022	.024	.002
certain	.011	.013	9.88e-6				
anger	.015	.013	.0002	fear	.020	.017	6.34e-7
negative	.033	.029	.0001	positive	.049	.046	.003
sadness	.016	.015	.013	surprise	.012	.011	.001
trust	.031	.029	.0008				

Table 13: Difference between average values of emotions and LIWC categories between correctly classified and misclassified examples for significant features (p≤0.05) for depression detection (eRisk dataset, no trasfer learning).

Emotion/ category	Mean value correct classif	Mean value misclassif	p-value	Emotion/ category	Mean value correct classif	Mean value misclassif	p-value
future	.080	.102	2.64e-10	article	.056	.052	.018
space	.050	.047	.019	affect	.063	.073	1.11e-5
inhib	.005	.004	.034	posemo	.041	.052	5.36e-8
verb	.196	.236	1.39e-9	past	.031	.034	.011
funct	.446	.492	8.20e-7	adverb	.041	.050	1.24e-7
insight	.019	.022	.001	negate	.020	.022	.018
leisure	.018	.015	.010	anger	.010	.007	.001
excl	.022	.026	.0002	present	.157	.193	3.79e-10
work	.020	.017	.012	percept	.024	.026	.028
auxverb	.147	.176	2.97e-8	i	.035	.055	1.47e-19
quant	.047	.027	.022	conj	.042	.049	8.19e-6
cogmech	.132	.145	.0001	incl	.028	.031	.001
assent	.006	.008	.0006	tentat	.024	.027	.001
ipron	.049	.057	.0002	feel	.005	.006	.001
death	.003	.001	.001	you	.018	.022	.004
pronoun	.121	.152	3.80e-11	anx	.002	.003	.048
discrep	.014	.015	.045	ppron	.071	.095	3.24e-13
fear	.018	.015	.004	negative	.030	.027	.010

Table 14: Difference between average values of emotions and LIWC categories between correctly classified and misclassified examples for significant features (p$\leq$0.05) for anorexia detection (no trasfer learning).

Emotion/ category	Mean value correct classif	Mean value misclassif	p-value	Emotion/ category	Mean value correct classif	Mean value misclassif	p-value
future	.098	.113	.0001	article	.057	.051	.030
health	.004	.008	6.73e-6	funct	.488	.540	1.51e-7
adverb	.051	.063	1.57e-5	insight	.020	.026	3.47e-5
excl	.026	.036	7.15e-8	present	.189	.210	.002
sexual	.006	.008	.019	percept	.024	.028	.036
auxverb	.171	.191	.001	i	.047	.066	8.82e-9
ppron	.092	.112	5.43e-7	money	.006	.004	.025
conj	.045	.061	1.19e-11	cogmech	.142	.168	1.66e-7
relig	.004	.002	.025	bio	.022	.029	.0006
incl	.026	.033	9.19e-7	tentat	.025	.032	7.77e-6
feel	.006	.008	.009	verb	.233	.258	.001
sad	.004	.006	.023	pronoun	.149	.175	1.06e-6
head	.005	.006	.017	we	.005	.003	.008
anx	.002	.003	.094	discrep	.016	.019	.016
negative	.030	.034	.030	sadness	.015	.019	2.37e-5

Table 15: Difference between average values of emotions and LIWC categories between correctly classified and misclassified examples for significant features (p$\leq$0.05) for self-harm detection (no transfer learning).

Emotion/ category	Mean value correct classif	Mean value misclassif	p-value	Emotion/ category	Mean value correct classif	Mean value misclassif	p-value
motion	.015	.016	.024	anx	.003	.002	.039
percept	.019	.021	.019	they	.004	.004	.001
see	.007	.008	.014	time	.044	.047	.013
article	.039	.037	.022	health	.006	.005	.004
negemo	.024	.021	.020	posemo	.040	.045	.0004
insight	.015	.013	.003	shehe	.007	.006	.004
death	.0019	.0015	.017				
anticipation	.024	.026	.005	fear	.017	.014	1.96e-5
joy	.023	.027	.0003	negative	.030	.027	.002
positive	.044	.049	.0009	sadness	.016	.014	.001

Table 16: Difference between average values of emotions and LIWC categories between correctly classified and misclassified examples for significant features (p$\leq$0.05) for PTSD detection (no transfer learning).

Emotion/ category	Mean value correct classif	Mean value misclassif	p-value	Emotion/ category	Mean value correct classif	Mean value misclassif	p-value
future	.080	.102	2.64e-10	space	.050	.047	.003
affect	.063	.071	5.77e-6	inhib	.005	.004	.004
posemo	.041	.050	2.98e-7	past	.030	.034	.0005
friend	.001	.002	.0004	funct	.443	.492	4.80e-10
nonfl	.0014	.0019	.020	adverb	.041	.049	3.89e-9
insight	.019	.022	4.79e-6	negate	.020	.023	.0003
excl	.022	.026	5.83e-7	present	.154	.192	4.18e-14
work	.020	.017	.002	preps	.100	.096	.026
auxverb	.145	.177	2.34e-12	i	.034	.052	1.24e-20
quant	.024	.027	.0006	conj	.042	.048	3.63e-6
cogmech	.131	.145	1.63e-6	incl	.038	.030	.012
assent	.006	.009	1.41e-6	tentat	.023	.027	1.82e-6
ipron	.049	.058	1.41e-8	feel	.005	.006	.005
verb	.193	.23	1.04e-13	death	.003	.002	.003
you	.018	.021	.001	pronoun	.119	.152	5.46e-16
social	.076	.085	7.18e-5	discrep	.013	.015	.001
ppron	.070	.093	3.25e-16				
anger	.014	.012	.013	fear	.018	.015	.0002
negative	.030	.027	.001				

Table 17: Difference between average values of emotions and LIWC categories between correctly classified and misclassified examples for significant features (p$\leq$0.05) for anorexia detection (with transfer strategy 2).

Emotion/ category	Mean value correct classif	Mean value misclassif	p-value	Emotion/ category	Mean value correct classif	Mean value misclassif	p-value
verb	.232	.257	.0004	future	.098	.110	.002
excl	.028	.031	.023	auxverb	.170	.191	.0003
ipron	.056	.063	.004	funct	.493	.516	.015
excl	.028	.031	.023	affect	.079	.087	.029
ppron	.094	.104	.010	sexual	.006	.008	.023
present	.188	.209	.0009	you	.021	.027	.0009
pronoun	.150	.168	.0004	posemo	.051	.059	.023

Table 18: Difference between average values of emotions and LIWC categories between correctly classified and misclassified examples for significant features (p$\leq$0.05) for self-harm detection (transfer learning strategy 2).

Emotion/ category	Mean value correct classif	Mean value misclassif	p-value	Emotion/ category	Mean value correct classif	Mean value misclassif	p-value
excl	.018	.021	3.08e-5	funct	.382	.410	.0006
work	.013	.009	.001	feel	.0051	.0059	.022
i	.049	.059	.0001	ppron	.088	.200	.0007
future	.080	.091	.0002	leisure	.016	.012	.002
death	.001	.002	.001	space	.042	.039	.010
anger	.012	.010	.006	tentat	.012	.018	.0002
swear	.006	.008	.013	ipron	.034	.034	.0001
sad	.003	.004	1.71e-5	auxverb	.135	.150	.0005
conj	.034	.040	9.49e-7	motion	.017	.014	.001
pronoun	.122	.138	.0001	negemo	.021	.026	.0001
adverb	.038	.042	.0007	bio	.025	.028	.017
body	.008	.010	.001	friend	.0020	.0026	.006
relativ	.109	.101	.002	we	.006	.005	.015
cogmech	.106	.112	.017	money	.005	.004	.013
anger	.010	.012	.006	positive	.048	.043	.005
sadness	.014	.015	.048				

Table 19: Difference between average values of emotions and LIWC categories between correctly classified and misclassified examples for significant features (p$\leq$0.05) for PTSD detection (with transfer strategy 2).

Task	**Distinctive terms**
Depression baseline	her, I, in, is, me, my, of, she, trump, was
Self-harm baseline	I, she, que, meme, is, film, game, bpd, video, suicide
Anorexia baseline	u, the, am, el, guerreriar, i, me, my, guerrero, reddit
PTSD baseline	besties, dundee, fife, gameinsight, ipadgames, **ptsd**, nipclub
Depression transfer to self-harm	que, de, despacito, o, you, para, se, that, cleetus, y
Depression transfer to anorexia	you, the, my, I, am, u, shinies, trump, senate, el, guerreriar
Depression transfer to PTSD	besties, brain, dundee, fife, neuro, salary, vietnam, thatsheartgiveaway

Table 20: Most distinctive terms in the vocabulary between correctly classified and misclassified examples, using chi^2 test on tf-idf scores.