On the Explainability of Automatic Predictions of Mental Disorders from Social Media Data

Ana Sabina Uban, Berta Chulvi and Paolo Rosso

Pattern Recognition and Human Language Technology (PRHLT),
Universitat Politècnica de València, València, Spain
ana.uban+acad@gmail.com, berta.chulvi@upv.es, prosso@dsic.upv.es

Abstract. Mental disorders are an important public health issue, and computational methods have the potential to aid with detection of risky behaviors online, through extracting information from social media in order to retrieve users at risk of developing mental disorders. At the same time, state-of-the-art machine learning models are based on neural networks, which are notoriously difficult to interpret. Exploring the explainability of neural network models for mental disorder detection can make their decisions more reliable and easier to trust, and can help identify specific patterns in the data which are indicative of mental disorders. We aim to provide interpretations for the manifestations of mental disorder symptoms in language, as well as explain the decisions of deep learning models from multiple perspectives, going beyond classical techniques such as attention analysis, and including activation patterns in hidden layers, and error analysis focused on particular features such as the emotions and topics found in texts, from a technical as well as psycho-linguistic perspective, for different social media datasets (sourced from Reddit and Twitter), annotated for four mental disorders: depression, anorexia, PTSD and self-harm tendencies.

1 Introduction and Previous Work

Mental disorders are a serious public health issue, and many mental disorders are underdiagnosed and undertreated. The early detection of signs of mental disorders is important, since, undetected, mental disorders can develop into more serious consequences, constituting a major predictive factor of suicide [33]. Computational methods have a great potential to assist with early detection of mental disorders of social media users, based on their online activity.

There is an extensive body of research related to automatic mental disorder detection from social media data. The majority of research has focused on the study of depression [6, 7, 1, 35], but other mental illnesses have also been studied, including generalized anxiety disorder [27], schizophrenia [16], post-traumatic stress disorder [3, 4], risks of suicide [19], anorexia [13] and self-harm [13, 34]. The majority of studies provide either quantitative analyses, or predictors built using simple machine learning models, such as SVMs and logistic regression [6, 5], with few studies using more complex deep learning methods [23, 29, 31, 26, 30]. As features, most previous works use traditional bag of words n-grams [3], as well as some domain-specific representations, such as lexicons [28,

5], or Latent Semantic Analysis [22, 28]. There are few studies which compare multiple different aspects of the language, such as topics and emotions [26, 27, 30].

Quantitative analyses in existing research on mental disorders have found that people suffering from depression manifest changes in their language, such as greater negative emotion and high self-attentional focus [5, 29], or an increased prevalence of certain topics, such as medications or bodily issues such as lack of sleep, expressing hopelessness or sadness [24, 28]. Nevertheless, correlation studies are limited in discovering more complex connections between features of the text and mental health disorder risks. Moreover, research on mental health disorders from a computational perspective has been generally disconnected from mental health research in psychology, with few computational studies providing interpretations from a psychological perspective [15].

In practice, models based on neural networks are vastly successful for most NLP applications. Nevertheless, neural networks are notoriously difficult to interpret. Recently, there is increasing interest in the field of explainability methods in machine learning including in NLP [8], which aim for providing interpretations of the decisions of neural networks. If any system for mental disorder detection is to be developed into a tool to assist social media users, it is essential that its decision-making process is understandable in the name of transparency. Especially in the medical domain, using black-box systems can be dangerous for patients and is not a realistic solution [36, 10]. Moreover, recently, the need of explanatory systems is required by regulations like the General Data Protection Regulation (GDPR) adopted by the European Union. 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 [2], 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 [32, 25]. In our study, we aim to go beyond explainability techniques based on the analysis of attention weights.

We intend to explore the explainability of mental disorder prediction models from different perspectives. We center our analysis around neural network models trained to identify signs of mental disorders from social media data for the four different mental health disorders, using various features to extract information reflecting different levels of the language, and through performing various complementary analyses of the behavior of the model and features used. In this way, we aim to discover the most relevant features that indicate mental disorder symptoms based on text data, analyze the way they manifest in text, as well as provide interpretations of our quantitative findings from a social psychology perspective.

2 Classification Experiments

2.1 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

Dataset	Users	Positive%	Posts	Words
eRisk depression	1304	16.4%	811,586	25M
eRisk anorexia	1287	10.4%	823,754	~23M
eRisk self-harm	763	19%	274,534	~6M
CLPsych depression	822	64.1%	1,919,353	~26M
CLPsych PTSD	1078	72.6%	2,541,214	~19M
Twitter depression [26]	519	50.2%	52,080	~500K

Table 1: Datasets statistics.

labelled for several disorders and manifestations thereof: depression, anorexia, self-harm, and PTSD, and gathered from two different social media platforms: Reddit and Twitter.

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, focused on disorders such as depression [12], anorexia and self-harm tendencies [13, 14]. Data is collected from Reddit posts and comments selected from specific relevant sub-reddits. Users suffering from a mental disorder are annotated by automatically detecting self-stated diagnoses. 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.

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 [4], 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 human curation. For each user, their most recent public tweets were included in the dataset.

Twitter dataset on depression. To complement the CLPsych dataset, we include a second Twitter dataset labelled for depression. This dataset was collected and introduced in [26], following a similar methodology, based on self-stated diagnoses. Tweets published within a month of the diagnosis statement were included for each positive user. This short time frame is an exception compared to the other datasets considered. Non-depressed users were selected among Twitter users never having posted any tweet containing the character string "depress". In all datasets, the posts containing the mention of a diagnosis were excluded. Table 1 contains statistics describing all datasets considered.

2.2 Experimental Setup

We center our analyses on training deep learning models to predict mental disorders in social media data, which we will try to analyze in the following sections in order to explain their behavior.

First, we train and test our model for classifying between healthy users and those suffering from a disorder, for each of the datasets and disorders independently. Secondly, we perform similar experiments for cross-disorder classification: we try to automatically

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

distinguish between users suffering from different disorders, in an attempt to understand not only on linguistic patterns used by people diagnosed with an disorder, but also compare how these patterns differ (or coincide) across different disorders.

For the task of identifying users on social media suffering from a mental disorder, we model the problem as a binary classification task, training a deep learning model separately for each of the disorders and datasets considered. In the case of cross-disorder classification, we consider separately the two data sources: Reddit and Twitter, and perform experiments to distinguish between disorders present in each of the datasets: depression vs PTSD for the CLPsych (Twitter) datasets, and depression vs anorexia vs self-harm for the eRisk (Reddit) datasets. In this setup, we ignore the healthy users, and only focus on identifying the particular disorder that users are suffering from. We consider these as multi-label classification tasks (using a sigmoid activation for the final layer of our deep learning model for both tasks, instead of softmax), taking into account the fact that some users might be suffering from multiple disorders, given the known incidence of co-morbidity of mental disorders [11].

2.3 Model and Features

We choose a hierarchical attention network (HAN) as our model: a deep neural network with a hierarchical structure, including multiple features encoded with LSTM layers and two levels of attention. The HAN is made up of two components: a *post-level encoder*, which produces a representation of a post, and a *user-level encoder*, 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 using pre-trained GloVe embeddings and passed to the LSTM are then 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. Posts are truncated or padded to sequences of 256 words. The post-level encoder LSTM has 128 units, and the user-level LSTM has 32 units. The dense layers for encoding the lexicon features and the stopwords feature have 20 units each. We use the train/test split provided by the shared task organizers, done at the user level, making sure users occurring in one subset don't occur in the other. Since individual posts are too short to be accurately classified, we construct our datapoints by concatenating groups of 50 posts, sorted chronologically. We publish all the code used for experiments reported in this paper in a public repository, which includes more details on the network's architecture[2].

We represent social media texts using features that capture different levels of the language (semantic, stylistic, emotions etc.) and train the model to predict mental disorder risk for each user.

Content features. We include a general representation of text content by transforming each text into word sequences.

Style features. The usage pattern of function words is known to be reflective of an author's style, at an unconscious level [18]. As stylistic features, we extract from each text a numerical vector representing function words frequencies as bag-of-words, which are passed through an additional dense layer of 20 units. We complement function word

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

	SELF-HARM		ANOREXIA		DEPRESSION						PTSD	
	eRisk		eRisk		eRisk		Shen et al.		CLPsych		CLPsych	
Model	**F1**	**AUC**	**F1**	**AUC**	**F1**	**AUC**	**F1**	**AUC**	**F1**	**AUC**	**F1**	**AUC**
HAN	.51	**.83**	.46	**.91**	.44	**.86**	.77	.81	.53	**.73**	.57	**.70**
LogReg	.45	.75	.49	.90	.36	.76	.71	.81	.55	.72	.49	.69

Table 2: F1 and AUC scores for all datasets and models trained on individual tasks.

distribution features with other syntactical features extracted from the LIWC lexicon, as described below.

LIWC features. The LIWC lexicon [20] has been widely used in computational linguistics as well as some clinical studies for analysing how suffering from mental disorders manifests in an author's writings. LIWC is a lexicon mapping words of the English vocabulary to 64 lexico-syntactic features of different kinds, with high quality associations curated by human experts, capturing different levels of language: including style (through syntactic categories), emotions (through affect categories) and topics (such as money, health or religion).

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 the sentiment and emotion categories in the LIWC lexicon, we include a second lexicon: the NRC emotion lexicon [17], which is dedicated exclusively to emotion representation, with categories corresponding to a wider and a more fine-grained selection of emotions, containing the 8 Plutchik's emotions [21], as well as *positive*/*negative* sentiment categories: *anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness*, *surprise*, *trust*. We represent LIWC and NRC features by computing for each category the proportion of words in the input text which are associated with that category.

Our choice of model is motivated both by its hierarchical attention mechanism, and by the multiple features used, which allow for interpretability from different perspectives.

2.4 Classification Results

Results for individual disorder detection are shown in Table 2. As performance metrics we compute the F1-score of the positive class and the area under the ROC curve (AUC), which is more robust in the case of data imbalance. We show results for our model, in comparison with a baseline logistic regression model with bag-of-word features.

In the case of cross-disorder classification, we obtain an F1-score of 0.72 for the depression class in depression vs PTSD classification, and an AUC score of 0.75. For the eRisk datasets, we obtain an accuracy of 0.44 for discriminating between depression, anorexia and self-harm, and a macro-F1 of 0.44. The results suggest the task of cross-disorder classification is significantly more difficult than distinguishing healthy users from ones suffering from a disorder, especially in the case of depression/anorexia/self-harm classification.

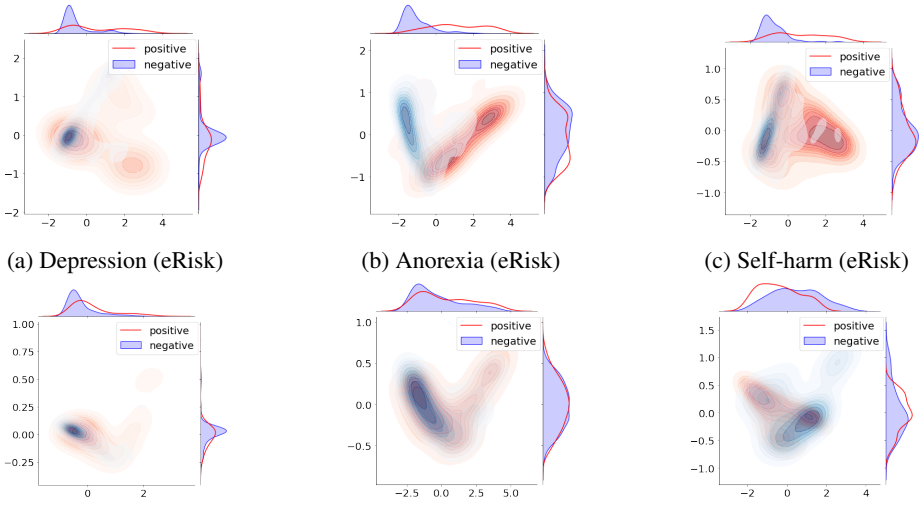
3 Explaining Predictions

In this section we present different analyses meant to uncover insights into how the model arrives at its predictions, first looking at the abstract internal representations of

the data in the layers of the network, and secondly providing several feature-focused analyses of misclassifications, using the lexicon-based features (emotions and LIWC categories) in order to identify particular interpretable patterns among users which the model cannot classify correctly.

3.1 User Embeddings

We start by analyzing the internal representations of the network. We can regard the final layer of the trained network as the most compressed representation of the input examples, which is, in terms of our trained model, the optimal representation for distinguishing between healthy users and those suffering from a disorder. Thus, the final layer (the output of the 32-dimensional user-level LSTM) can be interpreted as a 32-dimensional embedding for the input points, corresponding to the users to be classified.



(a) Depression (eRisk) (b) Anorexia (eRisk) (c) Self-harm (eRisk)

(d) PTSD (CLPsych) (e) Depress. (CLPsych) (f) Depress. (Shen et al.)

Fig. 1: User embeddings for classification of users with a disorder vs healthy ones.

We analyze the output of the *user embedding* layer by reducing it to 2 dimensions using principal component analysis (PCA) and visualizing it in 2D space with a kernel density estimate (KDE) plot to show the distribution of scores across the 2 dimensions, separately for each dataset and disorder (Figure 1). We make sure to train the PCA model on a balanced set of positive and negative users, then we extract 2D representations for all users in the test set. By looking at these representations, we can gain insight into the separability of the classes, from the perspective of the trained model, and better understand where it encounters difficulties in separating between the datapoints belonging to different classes. Separately, we perform the same experiments for cross-disorder classification, as shown in Figure 2.

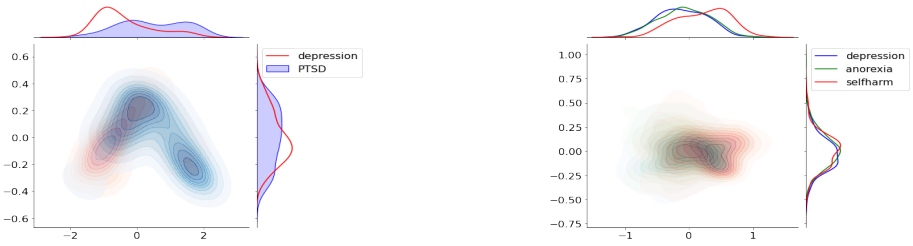
We notice that, in accordance with the classification performance reported previously, the highest separation in user embedding space seems to be achieved for anorexia and for

Label	Prediction Depression	Self-harm	Anorexia
Depression	139	2	113
Self-harm	60	67	144
Anorexia	201	16	218

Label	Prediction Depression	PTSD
Depression	126	24
PTSD	65	95

Table 3: Cross-disorder classification confusion matrices.

depression on the Twitter (Shen et al.) dataset, while depressed users in the other datasets (eRisk and CLPsych) show higher overlap with healthy ones, as do users suffering from self-harm. Moreover, we notice an interesting pattern of multiple clusters of positive users, while healthy users' representations seem to be more compact.



(a) Depression/PTSD (b) Depr./anorexia/self-harm

Fig. 2: User embeddings for cross-disorder classification.

In the case of cross-disorder classification, user embeddings seem highly overlapping, especially in the case of the 3-way classification of disorders in the eRisk datasets, suggesting that the model has difficulties in producing separate representations for these disorders, leading to a high misclassification rate.

In the following subsection we take a deeper dive into misclassified examples for each of the analyzed disorders and datasets. Focusing on misclassifications could also help to further explain the patterns noticed through user embedding analysis - particularly the clusters of false positives in the user embedding spaces for several disorders.

3.2 Error Analysis

We provide some insight into misclassified examples for cross-disorder classification through confusion matrices, as seen in Table 3. We notice a high rate of confusion for the 3-way classification between depression, anorexia, and self-harm, and particularly that users suffering from other disorders tend to be classified as depressed. The difficulty to distinguish between these disorders might be due to their common linguistic patterns, but also to possible cases of co-morbidities.

In the case of models for detecting individual disorders, errors of classification can have serious negative impacts on the users' well-being, if such as system would be deployed into a tool for assisting social media users. False negative predictions in particular can lead to missing cases of people with high risk of suffering from mental health disorders, and, left undetected, the disorders might further develop. We attempt

Experiment	False negatives	False positives
Depression (eRisk)	clemson, game, lemieux, team, uio play, song, pka, you, season	I, my, her, she, me, is was, the, are, trump, of
Depression (CLPsych)	earning, mpoints, video, rewards, patientchat, thank, besties, you, gameinsight, ipadgames	dundee, I, my, me, lol, the, vitamin, win, of, fuck, mobile, syria, love
Depression (Shen et al.)	I, rt, to, you, the, and, my, is, of, me	rt, prayer, bestmusicvideo, iheartawards, zain, pillowtalk, location, hiphopnews, ghetsis, via
Self-harm	I, the, que, is, me, de, a, despacito, feel, myself	the, I, a, to, and, it, you, of, is, that, in, for
Anorexia	I, the, my, her, she, r, me, eating, I'm, u, senate	I, the, am, you, of, their, transfer, college, him, from, in, girls
PTSD	mpoints, earning, reward, thank following, you, plz, ff, ptsd, cptsd	I, besties, gameinsights, ipadgames, thatsheartgiveaway, vietnam, coins, collected

Table 4: Top words (χ^2 test) that discriminate between incorrect and correct predictions.

to understand what causes misclassifications by comparing correctly versus incorrectly classified examples in terms of different features, including words, NRC emotions and LIWC categories. We thus compare the different types of misclassified and correctly classified examples, across the four groups: true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN).

In Table 4 we show the vocabulary words which are most distinctive for misclassifications for each disorder, separately for FP and FN cases. We select these words by applying the chi^2 test to extract the most discriminative features between FN and FP cases on one hand, and FP and TP cases on the other hand, and report the words with the highest scores. In some cases, these keywords can shed some light on what characterizes the sub-clusters of FN users identified with the user embedding representations. For depression, the FN group (both for eRisk and CLPsych) appear to be distinguished by discussing topics related to games. In the case of anorexia, we notice words related to college and social life in the FP group. Another interesting finding is the occurrence of "Vietnam" for FP in PTSD: the model learns to excessively associate PTSD sufferers with the topic of Vietnam, possibly showing a topic bias in the dataset.

In order to understand the effect of lexicon features on the model's prediction, we measure for each of the lexicon categories their comparative prevalence in misclassified and correctly classified examples, separately for healthy users and users suffering from a disorder. We identify four categories of features, based on their prevalence FP, FN, TP and TN examples comparatively:

Feature bias type 1 (**FN<TP; FP>TN**): features which occur to a lower degree in misclassified positive examples than in correctly classified positive examples; while for negative examples they occur more in incorrectly classified ones than in correctly classified ones. The model likely relies too much on the connection between their high prevalence and high risk scores.

Feature bias type 2 (FN<TP; FP<TN): generally under-represented features in misclassified examples - if they are not well represented, the model tends to make mistakes.

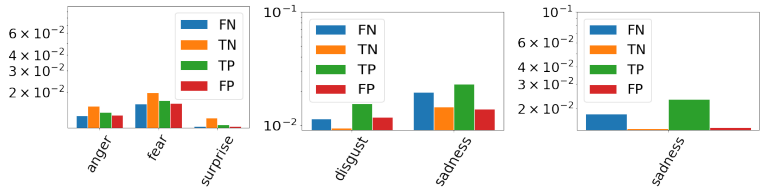
Feature bias type 3 (FN>TP; FP>TN): features which are generally over-represented in misclassified examples - when they are highly prevalent, the model is less accurate.

Feature bias type 4 (FN>TP; FP<TN): features which are over-represented in FN cases and under-represented in FP cases. The model likely relies too much on their low prevalence to emit high risk scores.

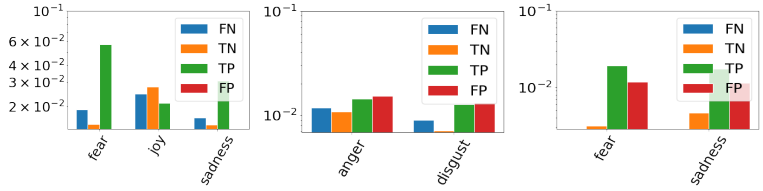
Experiment	Feat. bias 1	Feat. bias 2	Feat. bias 3	Feat. bias 4
Depression (eRisk)	ppron, quant, auxverb, verb, present, you, pronoun, excl, I, conj, adverb, future, cogmech, funct	-	ipron	-
Depression (CLPsych)	hear, conj, ipron, present, article, auxverb, certain, negative, verb,	-	-	-
Depression (Shen et al,)	-	-	-	ppron, adverb, bio, funct, verb past, funct, article, health, present
Self-harm	conj, excl, pronoun, future, cogmech, I, funct, ppron	-	-	incl
Anorexia	ppron, I, adverb, cogmech, auxverb, verb, pronoun, future, quant, excl, present, conj, funct	anxiety, health ingest, bio	-	-
PTSD	money, number, article work, achieve, preps	-	-	cogmech, fear, assent, pronoun, bio, I, leisure, swear, affect, feel

Table 5: LIWC features with highest differences for misclassified groups ($p < 10^{-6}$).

For each dataset, we identify misclassifications grouped into the two categories (FP and FN), and find those features for which there is a statistically significant difference of the average value between the misclassified group and the correctly classified group. We do this separately for emotions (see Figure 3) and for LIWC features (shown in Table 5, categories with p-values below 10^{-6}). We provide more interpretations for the patterns of misclassifications in relation to emotions and psycho-linguistic categories in the following sub-section, from a deeper psychological perspective.



(a) Depression (eRisk) (b) Anorexia (eRisk) (c) Self-harm (eRisk)



(d) PTSD (Twitter) (e) Depression (CLPsych) (f) Depression (Shen et al.)

Fig. 3: Mean values for emotions that are significantly different for misclassif. ($p<0.05$).

4 Cognitive Styles and Error Analysis: Some Interpretations

Cognitive style is a concept used in cognitive psychology to describe the way individuals think, perceive, and remember information [9]. Research in psychology suggests that some cognitive styles are more prevalent in some patients suffering from depression

and anorexia. In our error analysis we find that some errors have a relation with the under-representation of these cognitive styles. Some of the features that are relevant to explain the misclassifications of the model (see Table 5) are related to cognitive styles.

For instance, for depression we find that in the case of FN, features as *cogmech*, that refers to cognitive processes (causation, discrepancy, tentative, certainty, etc), occurs to a lower degree in FP examples. We can conclude that the model is confused when the depressed users do not express themselves in the typical pattern that refers some way of reasoning about causes, consequences, etc.

For anorexia, the under-representation of some features like *anxiety*, *health* or *ingest* leads to misclassifications. We can conclude that the model is relying on the use of these words to detect anorexia, but there are positive cases where we do not find the typical semantics of this disorder, and these will be more difficult to detect also for clinicians.

There is an interesting result related to the use of the *future* feature of LIWC. We found that in depression (eRisk corpus), anorexia and in self-harm, if this feature that speaks about *future* occurs to a lower degree, the model tends to make more mistakes in the classification of positive examples. We can infer that the model is able to detect the mental health disorders when people speak about what life is preparing for them, but has more difficulties when users that suffer from these mental health disorders don't speak about plans and focus more on the moment.

Considering the analysis of emotions (see Figure 3) we found also that the unclear expression of some emotions leads the model to make mistakes and that these emotions are just the ones that are relevant for each mental health disorder. For instance, we see that the model makes more mistakes when people that suffer from depression do not express anger and fear. In the case of anorexia, the FN examples are more frequent when people do not speak about disgust. This suggests that anorexia is a much more complex disorder that the one that express the development of strange eating habits. We also observe that in terms of emotions, the people with self-harm tendencies do not express their sadness emotion are more difficult to detect for the model and maybe also for clinicians. It suggests the need to explore other narratives that must being used for these people with self-harm tendencies that show a low expression of negative emotions.

5 Conclusions and Future Work

Explainability of machine learning models, especially in the domain of mental health, where automatic tools can have significant social impact, is an essential topic. In this study, we have presented several analyses for interpreting the decisions of models trained to profile users at risk of developing mental disorders from social media, going beyond more common techniques such as attention weight analysis, and including hidden layer analysis and error analysis at different levels of the language for better understanding how mental disorders manifest in social media data. In addition, we interpret our findings though the lens of psychology, identifying connections between specific topics (e.g. health, biology) or emotions (e.g. anger, fear) and certain disorders, which can lead the model to over-rely on these features.

Although we approach a novel topic in the computational research on mental disorders and present new findings, the methods used in this study could be developed

into deeper and more sophisticated analyses. As future work, we intend to continue the analysis of emotion markers through applying time series analysis methods, in order to automatically detect trends and seasonal patterns in the evolution of the usage of emotion-related vocabulary for users suffering of mental health disorders. Moreover, the results of the user embedding analysis encourage us to further study the distinct patterns of symptoms for certain disorders.

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.

References

1. Abd Yusof, N.F., Lin, C., Guerin, F.: Analysing the causes of depressed mood from depression vulnerable individuals. In: DDDSM-2017. pp. 9–17 (2017)
2. Amini, H., Kosseim, L.: Towards explainability in using deep learning for the detection of anorexia in social media. NLDB **12089**, 225
3. Coppersmith, G., Dredze, M., Harman, C.: Quantifying mental health signals in Twitter. In: CLPsych 2014. pp. 51–60 (2014)
4. Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., Mitchell, M.: CLPsych 2015 shared task: Depression and PTSD on twitter. In: CLPsych 2015. pp. 31–39 (2015)
5. De Choudhury, M., Counts, S., Horvitz, E.J., Hoff, A.: Characterizing and predicting postpartum depression from shared facebook data. In: ACM on Computer supported cooperative work & social computing. pp. 626–638 (2014)
6. De Choudhury, M., Gamon, M., Counts, S., Horvitz, E.: Predicting depression via social media. In: AAAI (2013)
7. Eichstaedt, J.C., Smith, R.J., Merchant, R.M., Ungar, L.H., Crutchley, P., Preoţiuc-Pietro, D., Asch, D.A., Schwartz, H.A.: Facebook language predicts depression in medical records. Proc. of the National Academy of Sciences **115**(44), 11203–11208 (2018)
8. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: An overview of interpretability of machine learning. In: IEEE DSAA. pp. 80–89. IEEE (2018)
9. Grigorenko, E.L., Sternberg, R.J.: Thinking styles. In: International handbook of personality and intelligence, pp. 205–229. Springer (1995)
10. Holzinger, A., Biemann, C., Pattichis, C.S., Kell, D.B.: What do we need to build explainable AI systems for the medical domain? arXiv preprint arXiv:1712.09923 (2017)
11. Kaufman, J., Charney, D.: Comorbidity of mood and anxiety disorders. Depression and anxiety **12**(S1), 69–76 (2000)
12. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk: early risk prediction on the internet. In: CLEF. pp. 343–361. Springer (2018)
13. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk 2019 early risk prediction on the internet. In: CLEF. pp. 340–357. Springer (2019)
14. Losada, D.E., Crestani, F., Parapar, J.: eRisk 2020: Self-harm and depression challenges. In: ECIR. pp. 557–563. Springer (2020)
15. Mehltretter, J., Rollins, C., Benrimoh, D., Fratila, R., Perlman, K., Israel, S., Miresco, M., Wakid, M., Turecki, G.: Analysis of features selected by a deep learning model for differential treatment selection in depression. Frontiers in Artificial Intelligence **2**, 31 (2020)

16. Mitchell, M., Hollingshead, K., Coppersmith, G.: Quantifying the language of schizophrenia in social media. In: CLPsych 2015. pp. 11–20 (2015)
17. Mohammad, S.M., Turney, P.D.: NRC emotion lexicon. National Research Council, Canada **2** (2013)
18. Mosteller, F., Wallace, D.L.: Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. Journal of the American Statistical Association **58**(302), 275–309 (1963)
19. O'dea, B., Wan, S., Batterham, P.J., Calear, A.L., Paris, C., Christensen, H.: Detecting suicidality on Twitter. Internet Interventions **2**(2), 183–188 (2015)
20. Pennebaker, J.W., Francis, M.E., Booth, R.J.: Linguistic inquiry and word count: LIWC 2001. Mahway: Lawrence Erlbaum Associates **71**(2001), 2001 (2001)
21. Plutchik, R.: Emotions: A general psychoevolutionary theory. Approaches to emotion **1984**, 197–219 (1984)
22. Resnik, P., Garron, A., Resnik, R.: Using topic modeling to improve prediction of neuroticism and depression in college students. In: EMNLP. pp. 1348–1353 (2013)
23. Sadeque, F., Xu, D., Bethard, S.: UArizona at the CLEF eRisk 2017 pilot task: linear and recurrent models for early depression detection. In: CLEF 2017 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings.CEUR-WS.org. vol. 1866. NIH Public Access (2017)
24. Schwartz, H.A., Eichstaedt, J., Kern, M., Park, G., Sap, M., Stillwell, D., Kosinski, M., Ungar, L.: Towards assessing changes in degree of depression through facebook. In: CLPsych. pp. 118–125 (2014)
25. Serrano, S., Smith, N.A.: Is attention interpretable? In: ACL. pp. 2931–2951 (2019)
26. Shen, G., Jia, J., Nie, L., Feng, F., Zhang, C., Hu, T., Chua, T.S., Zhu, W.: Depression detection via harvesting social media: A multimodal dictionary learning solution. In: IJCAI. pp. 3838–3844 (2017)
27. Shen, J.H., Rudzicz, F.: Detecting anxiety through reddit. In: CLPsych 2017. pp. 58–65 (2017)
28. Trotzek, M., Koitka, S., Friedrich, C.M.: Linguistic metadata augmented classifiers at the CLEF 2017 task for early detection of depression. In: CLEF 2017 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings.CEUR-WS.org. vol. 1866 (2017)
29. Trotzek, M., Koitka, S., Friedrich, C.M.: Word embeddings and linguistic metadata at the CLEF 2018 tasks for early detection of depression and anorexia. In: CLEF 2018 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings.CEUR-WS.org. vol. 2125 (2018)
30. Uban, A.S., Rosso, P.: Deep learning architectures and strategies for early detection of self-harm and depression level prediction **2696** (2020)
31. Wang, Y.T., Huang, H.H., Chen, H.H.: A neural network approach to early risk detection of depression and anorexia on social media text. In: CLEF 2018 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings.CEUR-WS.org. vol. 2125 (2018)
32. Wiegreffe, S., Pinter, Y.: Attention is not not explanation. In: EMNLP-IJCNLP. pp. 11–20 (2019)
33. World Health Organization, W.: Depression: A global crisis. world mental health day, october 10 2012. World Federation for Mental Health, Occoquan, Va, USA (2012)
34. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: NAACL-HLT 2016. pp. 1480–1489 (2016)
35. Yazdavar, A.H., Al-Olimat, H.S., Ebrahimi, M., Bajaj, G., Banerjee, T., Thirunarayan, K., Pathak, J., Sheth, A.: Semi-supervised approach to monitoring clinical depressive symptoms in social media. In: IEEE/ACM in Social Networks Analysis and Mining. pp. 1191–1198 (2017)
36. Zucco, C., Liang, H., Di Fatta, G., Cannataro, M.: Explainable sentiment analysis with applications in medicine. In: IEEE BIBM. pp. 1740–1747. IEEE (2018)