Automatic Detection and Classification of Mental Illnesses from General Social Media Texts

Anca Dinu[1] and Andreea-Codrina Moldovan[2]

[1]Faculty of Foreign Languages and Literatures, Digital Humanities Research Centre, University of Bucharest
[2]Faculty of Foreign Languages and Literatures, University of Bucharest
ancaddinu@gmail.com, moldovanandreeacodrina@gmail.com

Abstract

Mental health is getting more and more attention recently, depression being a very common illness nowadays, but also other disorders like anxiety, obsessive-compulsive disorders, feeding disorders, autism, or attention-deficit/hyperactivity disorders. The huge amount of data from social media and the recent advances of deep learning models provide valuable means to automatically detecting mental disorders from plain text. In this article, we experiment with state-of-the-art methods on the SMHD mental health conditions dataset from Reddit (Cohan et al., 2018). Our contribution is threefold: using a dataset consisting of more illnesses than most studies, focusing on general text rather than mental health support groups and classification by posts rather than individuals or groups. For the automatic classification of the diseases, we employ three deep learning models: BERT, RoBERTa and XLNET. We double the baseline established by Cohan et al. (2018), on just a sample of their dataset. We improve the results obtained by Jiang et al. (2020) on post-level classification. The accuracy obtained by the eating disorder classifier is the highest due to the pregnant presence of discussions related to calories, diets, recipes etc., whereas depression had the lowest F1 score, probably because depression is more difficult to identify in linguistic acts.

1 Introduction

An analysis performed by Chisholm et al. (2016) estimates that approximately 10% of the world's population is living with a mental illness. The Global Burden of Disease Study (2017) states that depression is a very common illness and there are more than 264 million people affected by it. At its worst, the illness can lead to suicide and it is the second highest cause of death for people between 15 and 29. Between 76% and 85% of the potentially diagnosed people, do not benefit from any treatment for their illness due to living in impoverished areas and not having access to mental care. It is difficult to discuss about digital solutions in the context of isolated areas with low data availability and limited access to professional help. Social stigma is another obstacle present regardless of age, gender or race, which makes early intervention difficult. Persons facing difficulties often avoid discussing their issues from various reasons. However, researchers working with Machine Learning algorithms can draw plenty of expertise from the unstructured data roaming the World Wide Web. The advent of social media platforms brings up an influx of large quantities of various types of unstructured textual data. The continuous advancements made in the field of Machine Learning enable the possibility to analyse such volumes of data efficiently. Experiments in this interdisciplinary domain managed to bring up useful input for mental health practitioners, socio-linguists, computer scientist and other researchers in the field. Pennebaker et al. (2015) perform one of the most influential quantitative studies, which reveals the way patterns of parts of speech, as labelled by LIWC founders, correlate with types of personalities and types of mental illnesses. The classes and the psychological dimensions mapped together served as a start for many projects including the prediction of Dark Triad personality traits by Sumner et al. (2012) and the risk of self-harm by Soldaini et al. (2018). Research in the area is conducted mainly on texts from mental health support groups, on just a few illnesses and some groups of individuals.

Our main research questions for this article are if and to what extent it is possible to detect and classify mental illnesses from general texts, if there

Proceedings of Recent Advances in Natural Language Processing, pages 358–366
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_041

are any differences between the difficulty of automatic detection and classification of different illnesses and, finally, if such a classification may rely on posts only. To this end, we experiment with state-of-the-art methods on Reddit, to improve previous results and provide new insights on mental illness discovery from general text. Reddit is a social media network hosting numerous communities where users join in order to participate in various discussions. Each community or "board", the way it is called by Reddit users, has a subject on which people must post. We employed highly performant deep learning models such as the Transformers, introduced by Vaswani et al. in 2017, which also led to the creation of BERT by Devlin et al. in 2018, a pretrained model trained on expansive general datasets, in order to be later fine-tuned on more specific tasks. Vale et al., (2021) efficiently applied BERT for question answering. Topal et al., (2021) use it for text generation and Sun et al., (2020) use it for text classification. Thus, we identify a solid ground for efficient usage in mental illnesses detection and classification.

2 Related Work

NLP researchers have shown an increased interest in the area at the intersection of Machine Learning and Psychiatry in the last years. Social media is an indispensable resource for research. Yet, the particularities of the online setting rise a range of challenges. As there are not any standards established for using social data, practitioners from many fields pointed to the dangers of using such data without a clear framework. Olteanu et al., (2019) address the issue of "biases, methodological pitfalls, and ethical boundaries" - discussing the problems often left unaddressed by researchers working with this kind of data. Selbst et al. (2019) analyse not only the ethical dilemma revolving around this type of studies, but also their feasibility and the integration of the social component into the compound of a socio-technical system.

When it comes to detecting mental illnesses from social media data, we have many examples at hand, which often look at data coming from those Reddit communities, which are support groups for people struggling with an illness or another. Most articles look at a single illness in comparison to a control group: Vedula et al. (2017) and Tsugawa (2019) – depression, and Birnbaum et al. (2020) – schizophrenia. Our goal is to detect a wide range of mental illnesses using deep learning techniques, which seem like the best candidates for this task. Jiang et al. (2020) employ deep learning methods similar to ours, but we concentrate on obtaining better results by training the models on individual posts rather than posts grouped by users, which might not work as expected. For example, if a user produced few contributions or has a fresh account, they would probably have few posts available. On the other hand, some types of user are the observing type and rarely contribute to discussions. One aspect worth mentioning is the nature of the data used in many classification tasks. Texts containing explicit content and linguistic cues pertaining to the properties of a certain illness are often used. Kim et al. (2020) and Thorstad et al. (2019) perform automatic text classification by their author's mental illnesses, with good results, on texts that specifically discussed these conditions on dedicated forums. Nevertheless, these classifications are of little help in finding risk population, when looking at general text, which does not include mental illness topics. Among the few researchers who report using datasets containing general discussions coming from people who self-reported their diagnosis in one of the support communities are Jiang et al. (2020) and Cohan et al. (2018).

The results are favorable and leave room for improvement. We believe it is important to experiment further for a better understanding of the ways in which mental illnesses can be detected in earlier stages and how even general discussions contain traces of how mental illnesses manifest themselves in language. In addition, this is a direction worthy of exploration because the persons asking for guidance represent a very small and idiosyncratic part of the population battling with mental illnesses, thus early mental illness detection from general text might be of a real help.

3 Data

We used the SMHD dataset introduced by Cohan et al. (2018). This dataset contains non-explicit texts: A Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions. They test some classification algorithms, but no deep learning. Also, employed LIWC categories for classification. These categories include standard linguistic dimensions – pro-nouns, articles, present tense, future tense; psychological processes – positive emotions,

negative emotions, anger, anxiety; personal concerns – work, achievements. The SMHD dataset contains texts extracted from Reddit's general discussion communities grouped on users and illnesses. Individuals diagnosed with a mental illness were detected by searching for self-reports in the dedicated support groups. The dataset features multiple illnesses, which are present in the psychiatric taxonomy DSM-5 (American Psychiatric Association, 2013). As stated by the authors of the dataset, "Six conditions are top-level DSM-5 disorders: schizophrenia spectrum disorders (schizophrenia), bipolar disorders (bipolar), depressive disorders (depression), anxiety disorders (anxiety), obsessive-compulsive disorders (ocd) and feeding and eating disorders (eating). The three other conditions are one rank lower: post-traumatic stress disorder (ptsd) is classified under trauma- and stress-related disorders, and autism spectrum disorders (autism) and attention-deficit/hyperactivity disorder (adhd) under neurodevelopmental disorders". The opposing group of users is the control one, whose members are selected based on having no posts in the support groups and at least 50 posts on Reddit. The complete dataset contains 20,406 diagnosed users and 335,952 control users. The texts do not contain any terms related to mental health, neither the diagnosed groups, nor the control ones. Our experiments will use just a selection of each group of illnesses to speed up the computation process. The models are not user centered and will learn from each individual post. Selecting data based on a fixed number of users was not suitable for our tasks due to the imbalance at the user level when it comes to the number of comments and posts available. Therefore, we selected randomly 50,000 posts for each group of users.

The numbers shown in tables 1 and 2 might reflect certain particularities about an illness, how the diagnosed users communicate in the online environment. This variation depends also on how the users engage, whether they create posts or comment on somebody else's and on the format adopted by each community – if pictures are posted often, then the comments are on the shorter side, if story telling is the center of the community, people engage with the purpose of telling their opinion or a similar story, hence the lengthier texts.

The authors of the dataset conducted a linguistic analysis based on LIWC categories. Several differences were observed between the diagnosed groups and the control users. Pennebaker et al. (2015) and Ireland and Mehl (2014) underline that pronounced usage of first-person singular with most conditions is consistent with the theory that illness drives one towards self-focus. An interesting finding underlining the bias of the dataset towards the predominantly male demographic is the female references that point to discussions about relationships and love related issues with the bipolar, depression and anxiety groups.

Illness	No. of users	No. of comments
Depression	500	71017
ADHD	500	73201

Table 1: The number of comments produced by the two groups are similar.

Illness	No. of users	No. of comments
PTSD	300	40885
Eating	300	10526

Table 2: The number of comments produced by the two groups of users is not balanced.

Illness	Total No. of tokens per group	Mean no. of tokens per group
DEPR	3,246,814	38.11
ANX	3,304,634	24.16
BIP	3,266,525	38.54
EAT	2,206,672	42.92
ADHD	3,241,564	40.47
PTSD	3,558,287	46.42
SCHIZO	4,611,530	37.10
OCD	3,068,948	42.01
AUT	3,348,654	39.24

Table 3: The number of tokens per group of illnesses and average number of tokens per person in a given group.

Reddit does not impose a very strict post limit hence we have diverse lengths. However, the deep learning models we used impose a limit for training. The SMHD has already undergone preprocessing, but we needed more cleaning. We remove any posts shorter than 4 tokens. Very short texts are often noise like thankful comments or very short approval phrases, which would confuse the model and do not contain significant meaning

pertaining to a class or another. Special characters and symbols are removed, and contractions are handled using a dictionary automatically translating them into the expanded forms. Table 3 shows the average number of tokens produced by users diagnosed with a certain illness and the average number of tokens a post from a user has.
The next section will look at another data related problem, namely the ethics and biases of working with social media data.

4 Ethics and Biases

Reddit represents a social media application whose users are part of communities and engage in discussions. Each social media network represents a cluster of people who are defined by certain characteristics. The Hootsuite yearly report (2021) shows that more than 60% of the Reddit users are males aged 18 to 34. Accordingly, studies show that there is a tendency in males to display less emotionally charged input due to the social stigma in the offline world. Nadeau et al. (2016) find that men often avoid seeking professional help or talking about their problems. Concealing their emotional state in real life is a strategy in order to avoid prejudices and is not something specific for the female population. Ireland and Mehl's (2014) research conducted in the psychology area proves that manifestations of negative emotions are muted across many settings and situations. Alternatively, Schoenebeck (2013) and Shelton et al. (2015) demonstrate that people tend to discuss personal things in anonymous spaces and share unpopular opinions. In this situation, Reddit represents a good source of data for a population, which is underrepresented in clinical studies. Anderson (2015) prove that some platforms might be more attractive for a demographic than others. Behavioral biases imply that users of a platform display a particular behavior, observable in how they interact with each other or what type of content they create. One such bias is the way in which users seek and share information. De Choudhury et al. (2014) discovered that users diagnosed with one illness behave differently in this aspect from the others. Nevertheless, we cannot claim that this is representative for all the individuals diagnosed with a mental illness. There are certain biases plaguing the studies based on social media, which should be at least mentioned for awareness. Here, we consider the population bias a positive fact, which enables studies targeting the young adult and adult males. However, this bias does not affect much our dataset, because the data collected comes from neutral communities where a variety of topics is discussed.

5 Discriminative Features

We run a Naïve Bayes Classifier in order to find out the most informative features from each category in our dataset. We used the classifier implemented in the scikit-learn library by Pedregosa, et al. (2011) to get a top of n most informative words by scores. Our experiment includes the 9 illnesses as labels, and the control group. The top *n* words can be seen in table 4. We notice that across the dataset, the top 8 words are mostly associated with the illness's groups. The words belong to the category of meaning words as established by the LIWC taxonomy. Even if the texts come from general discussion boards with less restrictive discussion topics, we can see that the terms are related to the diagnosis. There are words pertaining to the group of emotion: "awkwardness", "spiritual", "psychotic", "guilt"; love and sexuality: "attraction", "sexuality", "cis", "trans", "hormones"; terms related to illness and medication: "illness", "doses", "medication", "therapy", "pharmacist", "relapse" and acronyms: "NC" (no contact), "AA" (Alcoholics Anonymous), HSV (Herpes Simplex Virus), STD (Sexually Transmitted Diseases), TRP (in this case it refers to a Reddit community called r/The Red Pill – a controversial Men's Rights Activists (MRA) space which has now been removed). The idiosyncrasy of the network is seen in the occurrence of terms related to its own internal structure. Wide ranges of topics are discussed and that might differ in a clinician's office, in face-to-face situations and where one's identity is known. These distinctive features will help our classifiers to better distinguish the control users from the diagnosed ones because in some cases like in the depressed users and the ones suffering from anxiety the language is less distinctive.

6 Classification Methods

Identifying significant differences between our groups was the main drive for training classifiers. We trained 3 different models based on the Transformers architecture to see how each performs binary classification between a diagnosed group and a control one. We obtained state-of-the-

art results for text classification using BERT, RoBERTa and XLNET, as it follows.

BERT is a language representation model designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both the left and right context in all layers introduced by Devlin et al 2018). BERT's key technical innovation is applying the bidirectional training of Transformer, which is an advanced Long Short Term Memory Network (LSTM) to language modelling. The model uses the Transformer architecture to capture long distance dependencies within sentences. The pre-trained model contains a general knowledge of language and by giving it task-specific data we can obtain promising results. BERT uses a special tokenizer, which has a specific way of dealing with words outside its vocabulary. It accepts input in a required formatting: we have to add tokens marking the start and end of a sentence, pad and truncate sentences in order to have a uniform length. An attention mask is added in order to differentiate between the padding and the first type of tokens. The [CLS] token marks the beginning of a sentence and must be present for any sentence-level classification task. The [SEP] token separates one sentence from the next so the model can learn entailment. [CLS] and [SEP] tokens were used when the developers pretrained BERT, and we must preserve the same scheme for the model to work properly. The maximum sentence length supported by BERT is 512 tokens. Our experiment uses the Hugging Face Pytorch implementation

PTSD	**Chi ill**	**Chi Cont**	**ADHD**	**Chi ill**	**Chi Cont**	**Eating**	**Chi ill**	**Chi Cont**
NC	80.9	1.0	*kratom*	56.4	1.0	*carbs*	71.0	1.0
TRP	58.1	1.0	*NC*	41.3	1.0	*deficit*	53.0	1.0
abusive	53.8	1.0	*backyard*	22.6	1.0	*calorie*	47.5	1.0
meetings	51.5	1.0	*attraction*	21.8	1.0	*pound*	42.9	1.0
therapy	36.0	1.0	*pharmacist*	20.9	1.0	*obese*	42.0	1.0
pregnancy	26.1	1.0	*allergy*	20.0	1.0	*oily*	42.0	1.0
boundaries	35.2	1.0	*childfree*	26.0	1.0	*guilt*	28.2	1.0
hobbies	31.4	1.0	*hormonal*	20.3	1.0	*rice*	24.4	1.0

Depr	**Chi ill**	**Chi Cont**	**Anx**	**Chi ill**	**Chi Cont**	**Bip**	**Chi ill**	**Chi Cont**
caffeine	93.3	1.0	*HSV*	72.4	1.0	*Meds*	59.2	1.0
meds	38.0	1.0	*bernie*	61.1	1.0	*IGN*	1.0	58.4
attraction	34.1	1.0	*Eyeshadow*	40.2	1.0	*Kratom*	45.7	1.0
flirting	33.4	1.0	*Cannabis*	37.4	1.0	*Sweetheart*	41.8	1.0
hormones	31.1	1.0	*boyfriends*	27.0	1.0	*Bondage*	38.0	1.0
symptoms	26.1	1.0	*kratom*	33.6	1.0	*Vaping*	29.3	1.0
Mii	1.0	27.0	*STD*	26.0	1.0	*Literature*	27.4	1.0
cigars	1.0	29.3	*sexuality*	25.3	1.0	*hormones*	23.6	1.0

Schiz	**Chi ill**	**Chi Cont**	**OCD**	**Chi ill**	**Chi Cont**	**Autism**	**Chi ill**	**Chi Cont**
spiritual	84.6	1.0	*trans*	133.4	1.0	*feminine*	75.0	1.0
psychotic	76.1	1.0	*therapy*	81.7	1.0	*sensory*	64.5	1.0
meditation	49.9	1.0	*feminine*	51.3	1.0	*trans*	57.5	1.0
consciousness	44.8	1.0	*NC*	44.5	1.0	*NC*	50.9	1.0
cannabis	39.1	1.0	*cis*	31.4	1.0	*relapse*	29.9	1.0
doses	37.0	1.0	*TRP*	29.4	1.0	*TRP*	27.8	1.0
literature	37.7	1.0	*scarring*	29.4	1.0	*awkwardness*	24.6	1.0
AA	37.1	1.0	*literature*	28.1	1.0	*illness*	23.6	1.0

Table 4: Chi-square scores per each illness and control group.

BertForSequenceClassification by Wolf et al. (2019). In order to setup this model, we experimented using different hyperparameters, loss functions, batch sizes and number of epochs. The authors of BERT recommend using it with the following specifications:

- batch sizes: 8, 16, 32, 64, 128;
- learning rates: 3e-4, 1e-4, 5e-5, 3e-5.

Hyperparameters	BERT	RoBERTa	XLNet
Sequence Length	256	256	126
Batch Size	3	8	8
Weight Decay	0.0	0.0-0.1	0.0
AdamW ε	1e-5	1e-5	2e-5
AdamW β	0.9, 0.999	0.9, 0.999	0.9, 0.999
No. of training posts	100,000	100,000	100,000
No. of testing posts	20,000	20,000	20,000

Table 5: Hyperparameters for each model.

Our machine needed smaller batch sizes to be able to train the model, so we used 3. We established a learning rate (Adam ε) of 1e-5 for the AdamW loss function implemented by Loschilov and Hutter (2017). We trained the model for 3 epochs only, because we noticed overfitting starting with the 4th epoch.

The second method we used is XLNet, which is another method for pre-training language representations introduced by Yang et al. (2019). XLNet was meant to overcome the limitations imposed by BERT with its autoregressive model and does so by outperforming it on 20 tasks as shown by Yang et al. (2019). For this method, we have a different formatted input and there is no limit for the length of the input texts. However, the input arrays need to be of the same size. This is addressed by padding the inputs that do not meet the size of the longest sequence. Padding means simply adding 0s until the length is met. For this classifier we had to limit the length of sequences to 126 due to computational resources. The optimum batch size was 8. The loss function we used was AdamW with the same hyperparameters as for BERT. We trained this model for 4 epochs. With a training set of approximately 100000 texts, we get a number of 50000 training steps.

The last model we used, RoBERTa implemented by Liu et al. (2019), is Facebook AI's training method and it promises to improve on BERT. The researchers involved in implementing RoBERTa prove that BERT was undertrained and there is still a long way to go in terms of design choices and the way in which the improvements are reported. We did not use the full size of our dataset due to its large size and subsequent long training times.

Finetuning RoBERTa implies loading the weights of the pretrained model, in our case, the RobertaForSequenceClassification model. We use a sequence length of 256 and a batch size of 8. The loss function used here is AdamW with Adam ε of 2e-5.

7 Results

We obtained the results using 50,000 posts for each group alike. The compound of 100,000 posts for each binary classifier was split in 80,000 for training and 20,000 for testing. We trained our models with different hyperparameters until we reached the optimum ones detailed in table 6. We manage to overrun the baseline established by Cohan et al. (2018) using Transformers-based models on just a sample of their dataset. Their best results lie at approximately 50% accuracy with 57% being the best result obtained using Supervised FastText on Bipolar Disorder, while ours lie at approximately 75%, with 81% the best result. We also improve the results obtained by Jiang et al. (2020) on post-level classification as seen in table 7 by a considerable margin. We compare the results obtained using BERT and calculate the difference between our results and the ones from Jiang et al. Higher results were obtained with XLNet and RoBERTa in some cases. The BERT model achieves the highest accuracy for an illness: schizophrenia, OCD, eating disorder, autism and anxiety. We notice that discriminative features play an important role in building a performant model. The accuracy obtained by the eating disorder classifier is the highest due to the pregnant presence of discussions related to calories, diets, recipes etc. (as seen in table 6), whereas for depression we obtained the lowest F1 score, probably because depression is not always identifiable in linguistic acts, cf. Ireland and Mehl (2014). It is often a matter of contextual factors that might drive a user to discuss their emotional state or any other

	Metric	**DEPR**	CONT	**SCHIZ**	CONTR	**OCD**	CONT	**EAT**	CONT	**BPD**	CONT	**ADHD**	CONT	**PTSD**	CONT	**AUT**	CONT	**ANX**	CONT
B	P	0.73	0.63	0.73	0.72	0.74	0.76	0.82	0.80	0.75	0.72	0.73	0.66	0.75	0.76	0.72	0.69	0.73	0.72
	R	0.70	0.66	0.71	0.74	0.72	0.77	0.83	0.78	0.74	0.73	0.71	0.68	0.76	0.75	0.71	0.70	0.73	0.72
	F1	0.68		**0.73**		**0.75**		**0.81**		0.73		0.70		0.75		**0.71**		**0.73**	
XL	P	0.74	0.57	0.68	0.69	0.73	0.70	0.73	0.87	0.76	0.69	0.71	0.67	0.76	0.76	0.70	0.69	0.74	0.71
	R	0.67	0.81	0.57	0.78	0.71	0.72	0.87	0.70	0.69	0.76	0.69	0.70	0.74	0.77	0.72	0.67	0.71	0.74
	F1	**0.70**		0.68		0.72		0.79		0.72		0.69		**0.76**		0.70		**0.73**	
R	P	0.66	0.71	0.69	0.71	0.69	0.74	0.77	0.79	0.78	0.70	0.74	0.70	0.73	0.78	0.67	0.69	0.75	0.69
	R	0.71	0.59	0.73	0.66	0.77	0.65	0.78	0.78	0.74	0.73	0.70	0.68	0.76	0.75	0.74	0.70	0.70	0.77
	F1	0.68		0.70		0.72		0.78		**0.75**		**0.71**		0.75		0.70		**0.73**	

Table 6: Results for BERT (B), XLNET (XL) and RoBERTa (R) classifiers. Best results are in bold for each illness.

topics that might point towards a diagnosis or another.

8 Conclusions

Our automatic classification experiments used the Transformers-based models BERT, XLNet and RoBERTa on the SMHD dataset for the classification of 9 mental health illnesses. We manage to overrun Jiang et al. (2020) by approximately 0.10-0.15 on the single-post classification task and prove that individual posts yield satisfactory accuracy. We overrun Cohan et al. (2017) by 0.20-0.30 who did not employ any deep learning methods.

We used a Naïve Bayes Classifier to discover the most important features for each group of users. Our results add to the group of articles showing good prospects for this field. An encouraging finding is the sufficiency of focusing on general text rather than mental health support groups and classification by posts rather than individuals or groups. Another takeaway is the sufficiency of post-level classification and avenue to improve this approach in future work by paying attention to contextual cues such as time, events, entailment of posts or any other possible triggers that might help the earlier detection of a mental illness. Further experimentation with different setups and data that are more diverse is also required. This would benefit our research and increase the possibility of future integration of automated tools, which could assist clinicians in the earlier detection of mental health issues.

Illness	**Jiang et al.(2020)**	**This Work BERT**	**This work Overall**	**Cohan et al.**	**Difference**
	F1	**F1**	**F1**	**F1**	
DEPR	0.59	**0.68**	**0.70(XL)**	0.53	**+0.09**
SCHIZ	0.61	**0.73**	**0.73(B)**	0.45	**+0.12**
OCD	0.62	**0.75**	**0.75(B)**	0.44	**+0.13**
EAT	0.73	**0.81**	**0.81(B)**	0.44	**+0.08**
BPD	0.61	**0.73**	**0.75(R)**	0.57	**+0.12**
ADHD	-	**0.71**	**0.71(R)**	0.47	**N/A**
PTSD	0.57	**0.76**	**0.76(XL)**	0.57	**+0.19**
AUT	-	**0.71**	**0.71(B)**	0.49	**N/A**
ANX	0.68	**0.73**	**0.73(ALL)**	0.54	**+0.05**

Table 7: Comparison of BERT classification results – Jiang et al. (2020), Cohan et al. (2017) and our model. We report the results obtained by our binary classifiers in comparison with the binary classifiers trained by them. Cohan et al. did not employ any deep learning methods at the time, so we picked the highest F1 scores obtained with SVMs, Logistic Regression, FastText and CNNs.

Acknowledgements

This research is supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI

UEFISCDI, project number 108, COTOHILI, within PNCDI III.

References

American Psychiatric Association and American Psychiatric Association, editors. *Diagnostic and statistical manual of mental disorders: DSM-5*. 5th ed, American Psychiatric Association, 2013.

Anderson, M. *Men Catch Up With Women on Overall Social Media Use*. Technical report, Pew Research Center, 2015.

Birnbaum, M. L., S. K. Ernala, A. F. Rizvi, E. Arenare, A. R. Van Meter, M. De Choudhury & J. M. Kane. "Detecting Relapse in Youth with Psychotic Disorders Utilizing Patient-Generated and Patient-Contributed Digital Data from Facebook". *Npj Schizophrenia*, vol. 5, no. 1, 2019, p. 17, doi:10.1038/s41537-019-0085-9.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., Vanderplas, J., Joly, A., Holt, B., & Varoquaux, G.. "API design for machine learning software: experiences from the scikit-learn project". *arXiv:1309.0238 [cs]*, 2013. *arXiv.org*, http://arxiv.org/abs/1309.0238.

Chisholm, D., Sweeny, K., Sheehan, P., Rasmussen, B., Smit, F., Cuijpers, P., & Saxena, S. "Scaling-up Treatment of Depression and Anxiety: A Global Return on Investment Analysis". *The Lancet Psychiatry*, vol. 3, no. 5, May 2016, pp. 415–24, doi:10.1016/S2215-0366(16)30024-4.

Cohan, A., Desmet, B., Yates, A., Soldaini, L., MacAvaney, S., Goharian, N. "SMHD: a Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions". *Proceedings of the 27th International Conference on Computational Linguistics*, Association for Computational Linguistics, 2018, pp. 1485–97. *ACLWeb*, https://www.aclweb.org/anthology/C18-1126.

Choudhury, M.D., Morris, M., & White, R.W. "Seeking and Sharing Health Information Online: Comparing Search Engines and Social Media". Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2014, pp. 1365–76, doi:10.1145/2556288.2557214.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". *arXiv:1810.04805 [cs]*, May 2019. *arXiv.org*, http://arxiv.org/abs/1810.04805.

GBD 2017 Child and Adolescent Health Collaborators et al. "Diseases, Injuries, and Risk Factors in Child and Adolescent Health, 1990 to 2017: Findings From the Global Burden of Diseases, Injuries, and Risk Factors 2017 Study." *JAMA pediatrics* vol. 173,6 (2019): e190337. doi:10.1001/jamapediatrics.2019.0337

Holtgraves, Thomas M., et al. "Natural Language Use as a Marker of Personality". *The Oxford Handbook of Language and Social Psychology*, Thomas M. Holtgraves (ed.), Oxford University Press, 2014, doi:10.1093/oxfordhb/9780199838639.013.034.

Hootsuite (2021), "Digital 2021 USA report," https://www.hootsuite.com/pages/digital-trends-2021

Jiang, Z., Levitan, S., Zomick J., Hirschberg, J., "Detection of Mental Health from Reddit via Deep Contextualized Representations". *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, Association for Computational Linguistics, 2020, pp. 147–56, doi:10.18653/v1/2020.louhi-1.16.

Kim, Jina, Lee, J., Park, E., "A Deep Learning Model for Detecting Mental Illness from User Content on Social Media". Scientific Reports, vol. 10, no. 1, 2020, p. 11846, doi:10.1038/s41598-020-68764-y.

Loshchilov, I., & Hutter, F. (2019). Decoupled Weight Decay Regularization. ICLR.. "Decoupled Weight Decay Regularization". arXiv:1711.05101 [cs, math], 2019. arXiv.org, http://arxiv.org/abs/1711.05101.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". *arXiv:1907.11692 [cs]*, July 2019. *arXiv.org*, http://arxiv.org/abs/1907.11692.

Nadeau, M. M., Balsan, M. J., & Rochlen, A. B. "Men's Depression: Endorsed Experiences and Expressions." *Psychology of Men & Masculinity*, vol. 17, no. 4, October 2016, pp. 328–35, doi:10.1037/men0000027.

Olteanu, A., Castillo, C., Diaz, F., Kıcıman, E. "Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries". *Frontiers in Big Data*, vol. 2, July 2019, p. 13, doi:10.3389/fdata.2019.00013.

Pedregosa, F. et al. "Scikit-learn: Machine Learning in Python". *Journal of Machine Learning Research*, vol. 12, no. 85, 2011, pp. 2825–30.

Pennebaker, J.W., Boyd, R.L., Kayla, N.J., Blackburn, K., et al. *The Development and Psychometric Properties of LIWC2015*. 2015, doi:10.15781/T29G6Z.

Schoenebeck, S. "The Secret Life of Online Moms: Anonymity and Disinhibition on YouBeMom.Com". *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 7, no. 1, June 2013. *ojs.aaai.org*, https://ojs.aaai.org/index.php/ICWSM/article/view/14379.

Selbst, A.D., Boyd, D., Friedler, S., Venkatasubramanian, S., Vertesi, J., "Fairness and Abstraction in Sociotechnical Systems". *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ACM, 2019, pp. 59–68, doi:10.1145/3287560.3287598.

Shelton, Martin, et al. *Online Media Forums as Separate Social Lives: A Qualitative Study of Disclosure Within and Beyond Reddit*. March 2015. *www.ideals.illinois.edu*, https://www.ideals.illinois.edu/handle/2142/73676.

Soldaini, L., Walsh, T., Cohan, A., Han, J., Goharian, N. "Helping or Hurting? Predicting Changes in Users' Risk of Self-Harm Through Online Community Interactions". *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, Association for Computational Linguistics, 2018, pp. 194–203, doi:10.18653/v1/W18-0621.

Sumner, C., Byers, A., Boochever, R., Park G.J. "Predicting Dark Triad Personality Traits from Twitter Usage and a Linguistic Analysis of Tweets", *ResearchGate*, doi:10.1109/ICMLA.2012.218.

Sun C., Xipeng Q., Yige X., Huang X. "How to Fine-Tune BERT for Text Classification?" arXiv:1905.05583 [cs], February 2020. arXiv.org, http://arxiv.org/abs/1905.05583.

Thorstad, R., and Wolff, P. "Predicting Future Mental Illness from Social Media: A Big-Data Approach". Behavior Research Methods, vol. 51, nr. 4, 2019, pp. 1586–600, doi:10.3758/s13428-019-01235-z.

Topal, M. Onat, Bas A., van Heerden J. "Exploring Transformers in Natural Language Generation: GPT, BERT, and XLNet". arXiv:2102.08036 [cs], February 2021. arXiv.org, http://arxiv.org/abs/2102.08036.

Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y., & Ohsaki, H. Recognizing Depression from Twitter Activity. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. "Recognizing Depression from Twitter Activity". *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, ACM, 2015, pp. 3187–96, doi:10.1145/2702123.2702280.

Vale, L.D., & Maia, M. "Towards a question answering assistant for software development using a transformer-based language model". arXiv:2103.09423 [cs], 2021. arXiv.org, http://arxiv.org/abs/2103.09423.

Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. "Attention Is All You Need". *arXiv:1706.03762 [cs]*, 2017. *arXiv.org*, http://arxiv.org/abs/1706.03762.

Vedula, N., and Parthasarathy S. "Emotional and Linguistic Cues of Depression from Social Media". *Proceedings of the 2017 International Conference on Digital Health*, ACM, 2017, pp. 127–36, doi:10.1145/3079452.3079465.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., & Brew, J. "HuggingFace's Transformers: State-of-the-art Natural Language Processing". *arXiv:1910.03771 [cs]*, July 2020. *arXiv.org*, http://arxiv.org/abs/1910.03771.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q.V. "XLNet: Generalized Autoregressive Pretraining for Language Understanding". *arXiv:1906.08237 [cs]*, January 2020. *arXiv.org*, http://arxiv.org/abs/1906.08237.