Identifying Fine-grained Depression Signs in Social Media Posts

Augusto Rozendo Mendes, Helena de Medeiros Caseli
Federal University of São Carlos (UFSCar)
Rodovia Washington Luis, km 235 - São Carlos - SP - Brazil
augustorm@estudante.ufscar.br, helenacaseli@ufscar.br

Abstract

Natural Language Processing has already proven to be an effective tool for helping in the identification of mental health disorders in text. However, most studies limit themselves to a binary classification setup or base their label set on pre-established resources. By doing so, they don't explicitly model many common ways users can express their depression online, limiting our understanding of what kind of depression signs such models can accurately classify. This study evaluates how machine learning techniques deal with the classification of a fine-grained set of 21 depression signs in social media posts from Brazilian undergraduate students. We found out that model performance is not necessarily driven by a depression sign's frequency on social media posts, since evaluated machine learning techniques struggle to classify the majority of signs of depression typically present in posts. Thus, model performance seems to be more related to the inherent difficulty of identifying a given sign than with its frequency.

Keywords: mental health, social media, NLP, portuguese, depression

1. Introduction

Depression is a condition that affects a large portion of the global population, being the second largest contributor to decrease in the global healthy life expectancy[1]. However, according to the World Health Organization 2021 only a quarter of individuals afflicted with mental disorders receive proper care.

Depression is characterized by a clinically significant form of psychological suffering that leads to many impairments in someone's functionality, a reduction in quality of life, and in severe cases, it can lead to death due to the risk of suicide. Resources that aid in the detection of signs of depression can inform both individual treatment and public policy decision making. The most common tools for assessing signs of depression are psychometric scales and questionnaires, which can be administered by a healthcare professional or self-administered. These tools, although useful and validated through decades of practical application, have limitations in terms of reach, as they require either that a healthcare professional considers their application appropriate or that depressed people seek help despite the potential social stigma and impairment of functionality associated with the disorder.

These limitations motivate the creation of complementary depression detection techniques with broader reach, in order to enable the allocation of mental health resources to individuals and populations that would otherwise have no/poor access to them. These techniques include the use of content published in social media platforms, given their widespread adoption by the global population, as well as the general ease of access to such data. Among these efforts, the use of Natural Language Processing (NLP) stands out for automatic depression detection, as texts produced by depressed individuals can contain clear signs of depression (such as explicit mentions of symptoms and self-declarations of diagnosis/ongoing treatment) as well as more subtle clues like language style.

Typically, efforts to apply NLP for the task of depression detection deal with the problem as a binary classification task (Mowery et al., 2016; Coppersmith et al., 2014; De Choudhury et al., 2013b; Shoxiong et al., 2018; Liu et al., 2022). Possible reasons for choosing a binary configuration derive from the fact that it enables the use of automatic labeling strategies, and consequently the gathering of bigger corpora, which allows a wider array of viable techniques (such as deep learning) and typically results in models with good generalization capabilities.

A less explored alternative approach aims to classify a set of finer grained signs of depression, which are presumably more informative. Modeling the classification task in this way is also in accordance with medical practice, by assessing a set of characteristics associated with a mental health disorder according to a given psychopathological theory. These studies usually base their label schema on the PHQ-9 (Kroenke et al., 2001), a questionnaire that assesses 9 symptoms of depression codified by the DSM-5 (American Psychiatric Association, 2022). Using pre-established and clinically validated criteria ensures that resulting models will yield useful information about a user's mental state (if sufficiently accurate), however it must be acknowledged that tools like PHQ-9 were designed for a clinical scenario and might not be directly applicable to text content (for example, symptoms related to agitation need to be assessed through

[1] https://ghdx.healthdata.org/gbd-2019

LREC-COLING 2024, pages 8594–8604
20-25 May, 2024.

visual observation). It is also possible that online texts contain other signs of depression that are not usually covered by those tools, like external factors that can exacerbate or alleviate the depression.

Aiming at dealing with these limitations, this study focuses on a finer-grained set of signs of depression in social media posts, in an attempt to shed light on what signs are commonly found in social media posts, and which signs can be predicted with sufficient accuracy. This study takes into account 18 symptoms and 3 additional signs defined by a specialist committee that includes professionals in the fields of psychiatry, psychology, occupational therapy and NLP. A relatively small (in comparison to other studies) corpus of 780 Facebook posts published by Brazilian undergraduate students was annotated with the purpose of answering the following research questions:

RQ1 Which of the 21 signs of depression are more frequent in online text?

RQ2 Which of these signs are most easily identified using machine learning techniques?

We have observed that signs related to a user's emotional state and to external factors that worse/alleviate one's condition are frequently expressed in social media, along with text indicative of some behavioral signs of depression. However, only a few of the depression signs can be accurately predicted by machine learning techniques, suggesting that more complex and specialized model training setups are needed in order to cover the varied expressions of depression in social media posts.

2. Related work

One of the first works to investigate the use of social media text for the automatic prediction of depression was conducted by De Choudhury et al. 2013b. For the task of user-level depression classification, they used features including: (i) emotional content, from Linguistic Inquiry and Word Count categories (Tausczik and Pennebaker, 2010); (ii) activation and arousal extracted from ANEW (Bradley and Lang, 1999); and (iii) occurrences of depression related terms, behavioral features (frequency and time of posts, retweets and replies), language style (syntactic features) and social graph stats (density of networks, reciprocity of interaction, etc.). This feature engineering approach has informed several subsequent works despite recent developments in deep learning, since mental health professionals tend to prefer explainable models (Ji et al., 2022).

A smaller number of studies investigate the use of NLP in the domain of mental health focused on Portuguese language. Santos et al. 2023 developed a large corpus for user-level depression and anxiety classification in Brazilian Twitter users. Depressed users were identified by the presence of self-reports of formal diagnosis or ongoing treatment in their post history, and a control group of gender, post frequency and posting period matched users in a scale of 7:1[2] was selected. The study evaluated logistic regression, CNNs, LSTM networks and a fine-tuned BERT model with a Bi-LSTM classification head, with the latter achieving best performance for both depression and anxiety classification.

Mann et al. 2020 developed multimodal classifiers for user-level depression detection in Brazilian undergraduate students based on textual and visual information from their Instagram posts. Labeling was performed based on the participants' answers to a Beck Depression Inventory questionnaire. Both deep and traditional models were evaluated, with deep models utilizing TF-IDF, BoW, pretrained FastText or ELMo embeddings as text features and ResNet/ResNext embeddings for image features, fusing them in a fully connected multimodal layer. Their traditional models utilized LIWC category counts for text and hue, brightness and saturation values for images. Traditional models demonstrated competitive performance with deep models, but were not benefited by the multimodal setup, unlike the deep learning strategies.

Most work on depression symptom/sign classification are based on the PHQ-9 categories, with deviations either slightly expanding the label set or grouping the symptoms into simplified taxonomies. These studies also often evaluate techniques designed to mitigate data scarcity problems, which arise as a consequence of a more complicated task that necessitates manual annotation, preferably by someone with mental health expertise, resulting is small datasets. Yazdavar et al. 2017 put forward a semi-supervised approach based on a lexicon of PHQ-9 related terms (as well as a 10th category for common medications) created by a mental health expert. The technique (ssToT) selects a subset of lexicon terms frequently used by a user in order to guide Latent Dirichlet Allocation topic extraction. Lee et al. 2021 also start with a collection of seed terms and iteratively search for similar sentences in an unlabeled corpus in order to generate datasets for PHQ-9 symptom detection.

Casani et al. 2021 research multiclass depression symptom classification in Portuguese Twitter posts based on a simplified taxonomy consisting of psychological, physiological and behavioral symptoms, as well as a neutral category. The corpus was manually labeled by a mental health professional, resulting in 2008 annotated posts. The resulting models demonstrated good performance (above

[2]That is, 7 control users for each depressed/anxious user.

0.9 AUC) with a simple set of features (TF-IDF and BoW) and techniques (SVM, MLP, Naive Bayes).

3. Data collection and annotation

As part of a broader project aimed at devising depression detection and intervention solutions for Brazilian undergraduate students, the Amive[3] project, posts were collected from public Facebook pages where users vent their feelings. These pages offer a semi-anonymous way for students to express their feelings, by anonymously submitting text through a form that is then approved and posted by a centralized account, obfuscating the original author. This pairing of a certain degree of anonymity with a platform that allows long-form text posts encourages more varied expressions of signs of depression than other common data sources, such as tweets, that promote a direct communication style (Shoxiong et al., 2018). Posts were filtered through keyword search ("suicide", "depression", "cut myself", "will to live", "kill myself" and "want to die") and date of publication (from January 1st 2012 to December 31st 2021). All collected posts contained text in Brazilian Portuguese[4].

The collected posts were relatively lengthy in comparison to Twitter posts, which is the most common data source found in recent literature (Liu et al., 2022), and can vary greatly in size, with an average of 178 words and a standard deviation of 160 words[5]. The shortest collected post has only 4 words, while the longest has 949.

In total, 780 posts were selected for manual annotation. Given the sensitive nature of their contents, a manual process of further anonymization was conducted, replacing any possibly identifying information – including references to places, institutions, events, dates and courses – by generic tags (e.g. <city>). Links and user mentions were also anonymized in this same fashion, but in these cases, this process serves a dual function as a preprocessing step. While links, usernames and other such artifacts are typically removed from *corpora*, there is evidence that interaction between users can be a relevant factor for depression detection (De Choudhury et al., 2013b), in which case links to other posts and user mentions can carry useful information, and were therefore standardized into a generic representation as part of this anonymization effort.

The depression signs were defined by a specialist committee composed of psychiatry, psychology, occupational therapy and NLP experts. Signs were defined from a collective analysis of samples of posts as well as the professional experience of the mental health experts on the young and undergraduate population. The final depression signs set includes 18 depression symptoms: (1) Sleep disorder, (2) Alteration in efficiency/functionality, (3) Helplessness/Social harm/Loneliness, (4) Worry/Fear/Anxiety, (5) Despair, (6) Feeling of worthlessness/Low self-esteem, (7) Irritation/Aggressiveness, (8) Physical symptom, (9) Feeling of guilt, (10) Difficulty in decision making, (11) Tiredness/Discouragement/Fatigue, (12) Attention/memory deficit, (13) Feeling of emptiness, (14) Alteration in weight/eating habits, (15) Loss/Diminishment of pleasure/libido, (16) Sadness/Depressed mood, (17) Suicide/Self-extermination and (18) Agitation/Restlessness. In addition to these symptoms, 3 signs pertaining to external factors that are potentially informative in a mental health context were added: (1) Risk factors, which are factors that can worsen the depression state of a person and are related to a particular environment or to a general view of the world (e.g. bullying, racism, illness, interpersonal conflict); (2) Protective factors, which are factors that can help the depressed person face theirf issues (e.g. signs of a healthy support network, engaging in activities that promote well-being or that a user finds pleasurable); and (3) Death/suicide of third party. Thus, this work investigates a total of 21 signs.

The data was labeled by 4 annotators with familiarity in the domain of mental health (psychology, psychiatry and occupational therapy students). They were allowed to label any span of text representative of a sign, regardless of sentence boundaries or overlap with other annotations. This free form labeling strategy was employed in order to provide more autonomy to annotators in comparison with stricter approaches such as limiting labeled spans to sentence boundaries and forbidding span overlapping, which could hinder annotator's ability to provide full context for the attribution of a given sign in each span. A special tag was created to mark those posts from a person with a potential depressive profile (<PDP>). By doing so, we also delivered a binary depression classification dataset with little additional effort on the part of annotators, since they would have already carried out the assignment of signs of depression in a given post, and could easily label a post as PDP or not based on both the intensity and the amount of depressive signs.

Each annotator was assigned a distinct subset of posts with no overlap between them, and a subset of 100 posts was annotated by all in order to measure inter annotator agreement. This subset had moderate to high inter-rater agreement (75%) on a per-post basis, but only moderate Krippen-

[3] https://www.amive.ufscar.br/

[4] Data was collected from Crowdtangle, a public insights tool owned and operated by Facebook.

[5] as estimated by the NLTK(https://www.nltk.org/) word tokenizer

dorf's nominal alpha agreement (42.2%), which means that while annotators generally agreed on what signs were present in a post, they tended to annotate different (though often overlapping) spans of text as indicative of these signs. This subset of posts was also validated by psychiatry and occupational therapy professionals through a curation process, in order to create a set of posts with a more reliable gold-standard, reserved as a test set for classification models.

Table 1 shows the number of instances for each sign in Train and Test sets. In order to get a better sense of how much of the text did not contain any depression sign, a neutral class was established consisting of non-labeled spans from non-PDP posts. Non-labeled spans from PDP posts were not included as part of the neutral class given the moderate Krippendorf's nominal alpha agreement, since we expected that neutral spans extracted from such posts would be noisy and may contain some indications of signs of depression.

Sign	Train	Test
Agitation/Restlessness	5	0
Attention/memory deficit	9	4
Alteration in weight/eating habits	10	5
Loss/Diminishment of pleasure/libido	15	2
Physical symptom	16	8
Difficulty in decision making	22	4
Sleep disorder	17	10
Feeling of emptiness	28	6
Death/suicide of third party	37	10
Feeling of guilt	34	14
Irritation/Aggressiveness	50	38
Tiredness/Discouragement/Fatigue	81	25
Despair	85	27
Alteration in efficiency/functionality	85	28
Worry/Fear/Anxiety	111	29
Protective factor	115	52
Feeling of worthlessness/Low self-esteem	126	42
Suicide/Self-extermination	218	26
Helplessness/Social harm/Loneliness	220	63
Risk factor	177	109
Sadness/Depressed mood	278	63
Neutral	593	89
Overall	2152	523

Table 1: Number of positive instances for each sign

The collected data has a long-tail label distribution, typical of fine-grained label sets, with some signs being much more frequent than the rest ("Sadness/Depressed mood" and "Helplessness/Social harm/Loneliness", for example) and some categories having only a scarce number of positive instances ("Sleep disorder" and "Agitation/Restlessness", for instance). This property is also illustrated by Figure 1, which additionally demonstrates that label distributions between train and test sets are mostly similar, save for an increase in the number of neutral instances in the test set (a consequence of the smaller number of PDP posts) and a higher frequency of "Risk Factor", the only signs that reject the null hypothesis of the Kolmogorov-Smirnov test (p-values of 0.0001 and $2.67e^{-6}$ respectively). Even though a lot of signs are infrequent, only one is scarce

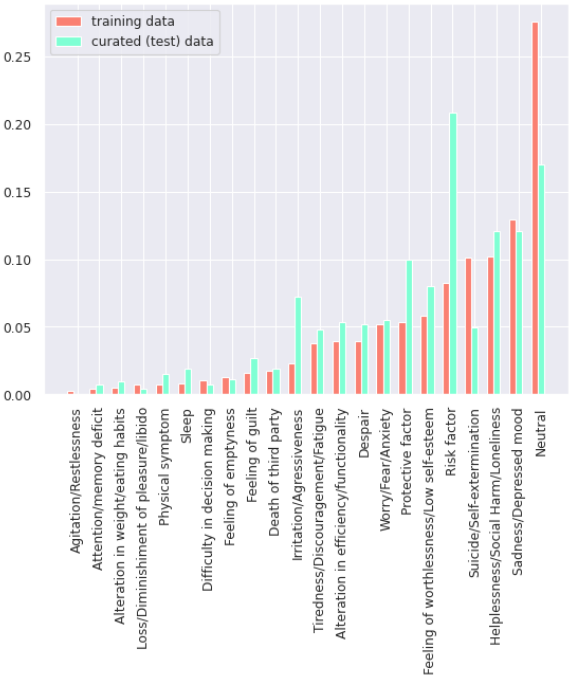

Figure 1: Label distribution for train and test sets

enough as to be ineligible for model training: "Agitation/Restlessness" does not occur in the test set.

By adopting a simplified taxonomy that groups signs into behavioral, emotional, somatic, and external, a pattern begins to emerge: somatic signs are very rare in texts, emotional and external signs are often present, and behavioral sign frequency varies on a case by case basis. In broad terms, students preferred to externalize their inner state of mind and discuss everyday events that impacted them. Note that infrequent signs are not necessarily difficult to automatically classify, and are potentially useful, but this scarcity is a relevant factor for both data collection and model performance, and indicative of which kind of sign might be better served by alternative methods of detection (such as bio-metric sensor data for somatic signs). This taxonomy was inspired by the one proposed by (Casani et al., 2021), validated by one member of the specialist committee, and detailed in Table 2.

4. Experimental setup

In order to answer our second research question (about which signs can be adequately classified by machine learning techniques), we selected a collection of 9 feature sets and 2 pre-trained language models. The feature sets were used for the purposes of training machine learning algorithms, and can be further subdivided into 5 engineered features (features derived from handcrafted resources or simple heuristics) and 4 different em-

Sign	Type
Less than 100 train instances	
Agitation/Restlessness	Behavioral
Attention/memory deficit	Behavioral
Alteration in weight/eating habits	Somatic
Loss/Diminishment of pleasure/libido	Behavioral
Physical symptom	Somatic
Difficulty in decision making	Behavioral
Sleep disorder	Somatic
Feeling of emptiness	Emotional
Death/suicide of third party	External
Feeling of guilt	Emotional
Irritation/Aggressiveness	Emotional
Between 100 and 200 train instances	
Tiredness/Discouragement/Fatigue	Behavioral
Despair	Emotional
Alteration in efficiency/functionality	Behavioral
Worry/Fear/Anxiety	Emotional
Protective factor	External
Feeling of worthlessness/Low self-esteem	Emotional
More than 200 train instances	
Suicide/Self-extermination	Behavioral
Helplessness/Social harm/Loneliness	Emotional
Risk factor	External
Sadness/Depressed mood	Emotional

Table 2: Mapping of depression related signs to a simplified taxonomy

bedding strategy categories. These feature sets were mostly selected based on reported performance in depression detection tasks found in the literature (De Choudhury et al., 2013a; Paraboni et al., 2020; Liu et al., 2022; Shoxiong et al., 2018; Casani et al., 2021), with the exception of NILCMetrix and AffectiveSpace embeddings.

4.1. Feature sets

We experimented with algorithms that have proven successful in works that evaluated binary depression classification or symptom classification in the literature: SVM, logistic regression, naive Bayes, RandomForest and XGBoost. Nine different types of features were extracted for the purposes of model training:

Engineered features

1. **LIWC**: A set of 72 categories from the Brazilian Portuguese version of the Linguistic Inquiry and Word Count (Filho et al., 2013). These include syntactic and emotional (such as polarity and sentiment) categories, as well as some tags potentially capable of capturing information related to external signs of depression (e.g. family, friends, health, money, work, etc.).

2. **AnewBR**: Average of dominance (or intensity) and arousal (or pleasantness) values of words present in each post as per ANEWbr, a Brazilian Portuguese version of ANEW (Kristensen et al., 2011). These values carry information regarding the sentiment expressed by a given text span.

3. **PHQ-9 categories**: A set of terms associated with a given symptom covered by PHQ-9, as per a lexicon of terms. This resource was originally constructed by (Yazdavar et al., 2017) and automatically translated to Portuguese by (Mendes et al.). For the purposes of this study, this lexicon was validated by a mental health professional and a medicine student.[6]

4. **POS+morph**: A set of course-grained POS and morphological categories following the Universal Dependencies (Nivre et al., 2016) UPOS and FEATS guidelines, as implemented by the spaCy library[7].This feature set is intended to capture language style, as various studies point to differences in style between depressed and non-depressed individuals, such as the more frequent use of first person by those afflicted by the condition.

5. **NILCMetrix**: A collection of 72 metrics extracted by the NILCMetrix library (Leal et al., 2023), that include descriptive, cohesion, readability, morphosyntatic and psycholinguistic metrics.This feature set was included since it could capture information related to cognitive impairment associated with depression.

Embedding features

6. **Static embeddings**: Average of the embeddings of every token in each post, extracted from Word2Vec, FastText, GloVe and LexVec models trained from social media posts in Portuguese and in the mental health domain (Paraboni et al., 2020). A pre-trained version of Word2Vec[8] was also used in order to evaluate if there's benefit in utilizing in-domain data as opposed to general resources trained on larger corpora.

7. **TF-IDF**: Term frequency-inverted document frequency values. The term frequencies were fitted on our training set.

8. **Contextualized embeddings**: Internal representations of the special token [CLS] produced by a pre-trained model, BERTimbau (Souza et al., 2020).

[6]The 10th category (common medications) was discarded as there was no mention of specific drugs in any collected post.

[7]https://spacy.io/

[8]Available at http://www.nilc.icmc.usp.br/embeddings

9. **AffectiveSpace embeddings**: Average of Portuguese AffectiveSpace (Cambria et al., 2015) embeddings for every token in each post. AffectiveSpace is a graph embedding trained on the SenticNet knowledge graph (Cambria et al., 2022), which aims to model common sense knowledge as it pertains to emotions.

4.2. Pre-trained language models

Besides feature extraction reserved for traditional machine learning algorithms, the following pre-trained language models were selected for fine-tuning:

- **General domain models**: BERTimbau and mBERT (Devlin et al., 2019), both capable of processing texts in Portuguese.
- **Domain-specific models**: mentalBERT and mentalRoBERTa (Ji et al., 2022), both pre-trained on social media content extracted from mental health communities. These specialized models were trained for English, and require translation of our corpus before fine-tuning[9].

5. Experiments

Our experiments aimed to identify which signs could be discerned automatically from text using strategies that have previously proven useful in the literature, in order to better understand which type of content these models can identify, and which they tend to miss. Separate classification models were trained for each sign in a one-vs-all setup .

In order to better understand what type of feature is useful, traditional models were trained on the whole collection of engineered feature sets, akin to feature engineering efforts (LIWC, NilcMetrix, arousal and dominance, POS+morph), but were trained on each embedding feature set separately (as each embedding strategy has the same aim of capturing general word meaning and context). NilcMetrix is an outlier in the engineered set of features since it is composed of comparatively complex metrics, which were in general designed for whole documents instead of our shorter annotated spans that usually encompass one or a few sentences, and these features are aggregations (ratios, averages, etc.) instead of simple counts. Because of these differences, and considering NilcMetrix represents roughly a third of all engineered feature data, it was ablated in order to measure its impact on performance.

All traditional ML algorithms hyper-parameters were fitted through a 60-iteration random search with 5-fold cross validation (stratified, given the imbalanced nature of the dataset). Deep models were only adjusted in terms of the pre-trained model used and loss function: either cross entropy or one of its variations, focal loss (Lin et al., 2017). Focal loss was included in order to potentially help the model cope with label imbalance, especially considering the large number of neutral instances present in the dataset, while remaining hyperparmeters were fixed[10]. We acknowledge that sampling, regularization, few-shot or semi-supervised techniques can be effective given the scarce and imbalanced data, however these fall outside the scope of this set of experiments, which only takes into account standard practices.

6. Results and Analysis

Table 5 shows the best results for each sign on the curated test set, consisting of a single experimental run after hyper-parameter fitting for each sign. We report average precision scores (which is the average precision along a precision-recall curve as the classification threshold varies from 0 to 1 recall) instead of AUC, as it is more robust to label imbalance (Saito and Rehmsmeier, 2015). Some signs can be classified with relatively good performance (average precision score above 70%), but the evaluated models struggle with most categories. The number of instances is a relevant factor for model performance, which tends to increase with more positive instances (as shown in Figure 2), but is not by itself a good predictor of performance, as signs that occur infrequently can still be predicted more accurately than others with an order of magnitude more data. For example, “Sleep disorder” reaches 90% average precision score with only 17 train instances of the positive class while the best model for “Risk Factor” reaches only 46% with 177 instances.

An analysis of model performance according to our simplified taxonomy also shows that somatic features in general are more easily discernible than the rest, despite their infrequency, reaching 76.14% macro-average precision score (as shown in Table 3).

A possible explanation for these results is that model performance is primarily dependent on the inherent complexity of each task, and as such even infrequent signs can be predicted if they are sufficiently simple. For example, while “Sleep disorder” concerns a relatively simple phenomenon (sleeping a lot/little, insomnia, irregular sleep schedules) and can thus be expressed in text in a simple and direct

[9] A m2m_100_418M model was used for neural machine translation.

[10] Learning rate of $10e^{-5}$, AdamW optimizer, early stopping after 4 epochs and constant schedule with linear warmup for the first 10% of training steps, 10% dropout before classification head.

Sign	Average precision score
Somatic	76.14% ± 12.08%
Behavioral	52.01% ± 14.18%
External	51.73% ± 4.50%
Emotional	48.07% ± 20.90%

Table 3: Macro average precision scores for each type of sign

way (e.g. “I didn’t sleep well last night”), “Despair” can be expressed in so many different ways as illustrated in Table 4. Texts pertaining to the same sign can also illustrate this complexity: while “Protective Factor” can include any number of self-care practices, positive encounters with others, and any other experience that gives a depressed person more resilience, it can also include sentences that confound models given the negative connotation of the language used (e.g. “let’s all talk a little more, cry a little more together[...]”, in which crying represents being more vulnerable and open to others), or referring to the protective factor indirectly (e.g. “I couldn’t have done it by myself”, implying a healthy support network).

With each day that passes my world gets darker, and I can’t see the future!
Why do I talk about myself in the past? I’m still here, aren’t I? Wasn’t university supposed to be a den of knowledge? Why do I feel its a tomb swallowing me whole?
If I had a mission here, I’m failing it
Life is horrible by itself
Nothing bears fruit for me
Knowing I’ll need months of therapy to get over this is horrible, I’m already torn apart now, imagine in the future...

Table 4: Examples for the “Despair” sign instances. Note the use of figurative language and the varied ways users can broach the subject.

Some of the literature regarding depression symptom classification suggests that people usually apply figurative language to express feelings that are hard to communicate (Yadav et al., 2020), which can further hamper the ability of models to accurately classify signs. Our results also demonstrate that despite general lack of data (fewer than 300 positive instances seen during training for every sign) and minimal adjustments in the fine-tuning setup that account for this scenario, deep models achieved better performance than traditional ones in 11 out of 20 cases, even in some of the least frequent categories (“Physical symptom”, “Loss/Diminishment of pleasure/libido”).

With regard to traditional models, engineered features achieved best performance in 9 of 20 signs when compared with their embedding counterparts,

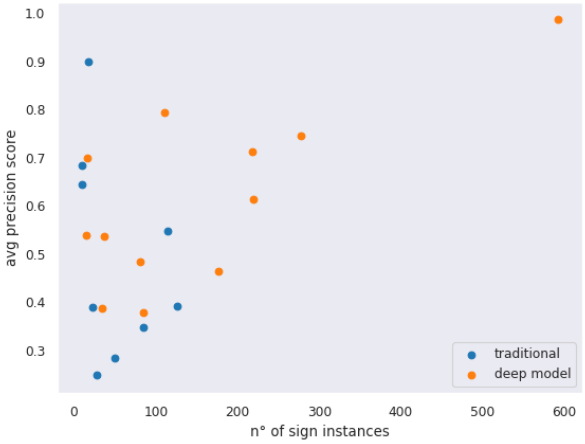

Figure 2: Visualization of model performance per number of instances.

and 2 out of 20 when compared to all models. While these engineered features by themselves are generally not sufficiently informative for the purpose of sign classification, they carry information that might not be easily derived from dense self-supervised text representations, particularly in a data-starved context. The use of NilcMetrix resulted in a deterioration in performance for all signs of depression.

Training models with in-domain data also proved to be relevant. Among models that utilize embeddings as features, the in-domain Word2Vec embeddings surpass its pre-trained counterpart, contextual embeddings, and AffectiveSpace embeddings for every sign. The fact that mentalBERT and mentalRoBERTa models achieved better performance than their native model counterparts for 8 of 20 signs despite noise introduced by automatic translation corroborates this observation, and points to the potential benefit of a mental health domain specific language model for Portuguese[11].

Finally, accounting for label imbalances also proved to be relevant, with most of the best models found through hyperparameter search utilizing some form of class weight balancing strategy. A set of subsequent experiments utilizing common sampling (random sampling, near miss sampling, and sampling of centroid clusters) and regularization techniques (SMOTE and back translation) was conducted. These techniques led to a drop in performance in all cases.

6.1. Answering the research questions

Regarding the research questions we posed in the beginning of this paper:

[11] By the time of this paper’s writing, we were not able to find such model publicly available.

Sign	Model	Average precision
Feeling of emptiness	XGBoost (word2vec)	0.248
Irritation/Aggressiveness	SVC (word2vec)	0.284
Alteration in efficiency/functionality	SVC (word2vec)	0.348
Despair	BERTimbau	0.379
Feeling of guilt	BERTimbau	0.387
Difficulty in decision making	RandomForest (engineered)	0.391
Feeling of worthlessness/Low self-esteem	XGBoost (word2vec)	0.391
Risk factor	BERTimbau	0.465
Tiredness/Discouragement/Fatigue	BERTimbau	0.483
Death/suicide of third party	BERTimbau	0.538
Loss/Diminishment of pleasure/libido	BERTimbau	0.540
Protective Factor	SVC (word2vec)	0.548
Helplessness/Social harm/Loneliness	MentalRoberta	0.614
Attention/memory deficit	XGBoost (word2vec)	0.645
Alteration in weight/eating habits	SVC (word2vec)	0.683
Physical symptom	BERTimbau	**0.700**
Suicide/Self-extermination	MentalRoBERTa	**0.712**
Sadness/Depressed mood	BERTimbau	**0.745**
Worry/Fear/Anxiety	MentalBERT	**0.794**
Sleep disorder	LogisticRegression (engineered)	**0.900**

Table 5: Average precision scores for each sign. Models which achieved more than 0.7 average precision score are in bold.

RQ1 Which of the 21 signs of depression are more frequent in online text?
We observed that, in general, categories related to **external factors** and **emotions** are frequently expressed in social media text, while direct reports of somatic symptoms are uncommon. Behavioral signs of depression vary in frequency on a case by case basis, with **"Suicide/Self-extermination"**, **"Alteration in efficiency/functionality"** and **"Tiredness/Discouragement/Fatigue"** among the most frequent signs (+100 instances). It must be considered that the high frequencies for "Suicide/Self-extermination" might be in part a consequence of the keywords used for data collection.

RQ2 Which of these signs are easily identified by machine learning techniques?
The evaluated models were capable of classifying somatic signs of depression with sufficient performance (here defined as 70%+ average precision score). Regardless of their category in the simplified taxonomy, most of the remaining assessed signs could not be predicted with sufficient precision utilizing common traditional and deep learning approaches. This result indicates that these techniques (alone) struggle to detect most signs of depression present in social media, at least in a fine-grained fashion, with the exception of the signs "Sadness/ Depressed mood" and "Worry/Fear/Anxiety", which most closely match existing work on depression detection and anxiety detection respectively. Its noteworthy that **some of the most frequent signs of depression expressed in the collected data (which could be indicative of general trends in social media text) are hard to classify**, demonstrating a clear direction for evaluation and improvement in future depression classification studies, specially concerning external factors, figurative language usage and better generalization to out of distribution data.

7. Conclusions and future directions

We collected and annotated a small corpus of 780 Facebook posts for the tasks of classifying a diverse set of 21 depression related signs, resulting in 2304 total positive instances. Signs related to emotional and external factors, as well as some behavior indicators of the condition occurred most frequently, but somatic signs of depression were overall easier to classify accurately.

Deep models achieved better performance than traditional ML techniques in most cases despite a lack of data and less robust hyperparameter search. These results, when paired with the observation that difficulties in model training seem to be driven mostly by the inherent complexity of classifying a given label rather than a mere product of small data volumes, suggest that a promising direction for future work on fine-grained depression sign classification is the application of more involved training methods, particularly those that can introduce new, task-relevant information to these deep representations without additional cost/time-intensive labeling, such as active, multi-task, transfer and

meta-learning approaches.

8. Ethics Statement

All social media content used was collected from publicly available pages following data collection protocols established by each platform. Care was taken to ensure that texts subject to manual annotation contained no identifiable information about their authors. Both collection and annotation of the depression sign corpus was approved by an ethics review board. Both data and generated models are intended exclusively for research purposes.

While the purpose of mental health classification tasks in general is to improve care, it is important to acknowledge that such models could potentially be used by malicious actors for purposes of harassment and discrimination.

Furthermore, we believe that if such models were to be applied to online texts produced by a general population with the express aim of conducting mental health interventions, participation should be strictly and clearly opt-in, regardless of the nature of such intervention.

Acknowledgments

This work is part of Amive project supported by FAPESP Grant #20/05157-9. We also thank the *Programa de Pós-graduação em Ciência da Computação* (PPGCC) from UFSCar and CAPES for the financial support.

American Psychiatric Association. 2022. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5-TR*. American Psychiatric Association Publishing.

Margaret M Bradley and Peter J Lang. 1999. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology ....

Erik Cambria, Jie Fu, Federica Bisio, and Soujanya Poria. 2015. Affectivespace 2: Enabling affective intuition for concept-level sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Vinicius Casani, Alinne C. Correa Souza, Rafael G. Mantovani, and Francisco Carlos M. Souza. 2021. DP-symptom-identifier: uma estratégia para classificar sintomas de depressão utilizando um conjunto de dados textuais na língua portuguesa. In *Annals of the XIII Brazilian Symposium of Information Tecnology and Human Language (STIL 2021)*. SBC.

Glen Coppersmith, Mark Dredze, and Craig Harman. 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, Baltimore, Maryland, USA. Association for Computational Linguistics.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013a. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th annual ACM web science conference*, pages 47–56.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013b. Predicting depression via social media. In *Proceedings of the international AAAI conference on web and social media*, volume 7, pages 128–137.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly available pretrained language models for mental healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190, Marseille, France. European Language Resources Association.

Christian Haag Kristensen, Carlos Falcão de Azevedo Gomes, Alice Reuwsaat Justo, and Karin Vieira. 2011. Normas brasileiras para o affective norms for english words. *Trends in Psychiatry and Psychotherapy*, 33:135–146.

Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613.

Andrew Lee, Jonathan K. Kummerfeld, Larry An, and Rada Mihalcea. 2021. Micromodels for Efficient, Explainable, and Reusable Systems: A Case Study on Mental Health. In *Findings of the Association for Computational Linguistics:*

EMNLP 2021, pages 4257–4272, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Danxia Liu, Xing Lin Feng, Farooq Ahmed, Muhammad Shahid, Jing Guo, et al. 2022. Detecting and measuring depression on social media using a machine learning approach: systematic review. *JMIR Mental Health*, 9(3):e27244.

Paulo Mann, Aline Paes, and Elton H. Matsushima. 2020. See and read: Detecting depression symptoms in higher education students using multimodal social media data. *Proceedings of the International AAAI Conference on Web and Social Media*, 14:440–451.

Augusto R Mendes, Rafael VP Passador, and Helena M Caseli. Identificando sintomas de depressao em postagens do twitter em português do brasil. In *Annals of the XIII Brazilian Symposium of Information Tecnology and Human Language (STIL 2021), pages=162–171, year=2021, organization=SBC*.

Danielle L. Mowery, Albert Park, Craig Bryan, and Mike Conway. 2016. Towards automatically classifying depressive symptoms from Twitter data for population health. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 182–191, Osaka, Japan. The COLING 2016 Organizing Committee.

Ivandré Paraboni, Amanda Maria Martins Funabashi, and Wesley Ramos dos Santos. 2020. Searching Brazilian Twitter for signs of mental health issues. *LREC*, page 7.

Takaya Saito and Marc Rehmsmeier. 2015. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432.

Wesley Ramos dos Santos, Rafael Lage de Oliveira, and Ivandré Paraboni. 2023. Setembrobr: a social media corpus for depression and anxiety disorder prediction. *Language Resources and Evaluation*, pages 1–28.

Ji Shoxiong, Celina Ping Yu, Fung Sai-fu, Pan Shirui, and Long Guodong. 2018. Supervised Learning for Suicidal Ideation Detection in Online User Content. *Complexity, Hidawi Wiley*, page 11.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

World Health Organization. 2021. Comprehensive mental health action plan 2013–2030. Technical report, World Health Organization.

Shweta Yadav, Jainish Chauhan, Joy Prakash Sain, Krishnaprasad Thirunarayan, Amit Sheth, and Jeremiah Schumm. 2020. Identifying depressive symptoms from tweets: Figurative language enabled multitask learning framework. *arXiv preprint arXiv:2011.06149*.

Amir Hossein Yazdavar, Hussein S Al-Olimat, Monireh Ebrahimi, Goonmeet Bajaj, Tanvi Banerjee, Krishnaprasad Thirunarayan, Jyotishman Pathak, and Amit Sheth. 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.

9. Language Resource References

Cambria, Erik and Liu, Qian and Decherchi, Sergio and Xing, Frank and Kwok, Kenneth. 2022. *SenticNet*. PID https://sentic.net/downloads/.

Filho, Pedro P. Balage and Pardo, Thiago Alexandre Salgueiro and Aluísio, Sandra M. 2013. *Brazilian Portuguese LIWC Dictionary for Sentiment Analysis*. PID http://nilc.icmc.usp.br/portlex/index.php/pt/projetos/liwc.

Ji, Shaoxiong and Zhang, Tianlin and Ansari, Luna and Fu, Jie and Tiwari, Prayag and Cambria, Erik. 2022. *MentalBERT*. PID https://huggingface.co/mental.

Leal, Sidney Evaldo and Duran, Magali Sanches and Scarton, Carolina Evaristo and Hartmann, Nathan Siegle and Aluísio, Sandra Maria. 2023. *NILCMetrix*. PID http://fw.nilc.icmc.usp.br:23380/nilcmetrix-en.

Nivre, Joakim and De Marneffe, Marie-Catherine and Ginter, Filip and Goldberg, Yoav and Hajic, Jan and Manning, Christopher D and McDonald, Ryan and Petrov, Slav and Pyysalo, Sampo

and Silveira, Natalia and others. 2016. *Universal dependencies v1*. ISLRN 586-682-285-530-1.