Universidade Federal de São Carlos– UFSCar
Centro de Ciências Exatas e de Tecnologia– CCET
Departamento de Computação– DC
Programa de Pós-Graduação em Ciência da Computação– PPGCC

Augusto Rozendo Mendes

Understanding Depression Symptoms in Social Media Posts: a Fine-grained Approach under Limited Data Constraints

São Carlos
2024

Augusto Rozendo Mendes

Understanding Depression Symptoms in Social Media Posts: a Fine-grained Approach under Limited Data Constraints

Master Thesis presented to Programa de Pós-Graduação em Ciência da Computação do Centro de Ciências Exatas e de Tecnologia da Universidade Federal de São Carlos, as part of the requirements for obtaining a Master's degree in Computer Science

Line of research: Computation Methodologies and Techniques

Supervisor: Helena de Medeiros Caseli

São Carlos
2024

Agradecimentos

Agradeço a minha orientadora Helena de Medeiros Caseli (novamente) por sua paciência comigo. Agradeço também à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela bolsa recebida durante parte do desenvolvimento deste projeto, e ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos por fornecer estrutura e apoio durante toda minha experiência no ensino superior. Agradeço também à equipe do projeto Amive (FAPESP #20/05157-9) do qual esta pesquisa faz parte e à FAPESP pelo apoio a este projeto.

Abstract

This study investigated the identification of depression signs in online text, utilizing a set of fine-grained labels, composed of 21 distinct signs, in order to deepen the collective understanding of how depression is expressed online. Results indicated that emotional and external signs of depression are frequent in social media, while somatic signs are scarcely expressed; however this trend does not carry over to model performance, with models performing best in somatic sign classification and struggling with some of the most frequent signs.

Given these challenges regarding model performance, potentially related to data scarcity, a series of techniques were evaluated with the goal of improving model performance, including regularization techniques, data augmentation, prompt engineering and multi-task learning, among which multi-task learning proved to be the most promising. With the continuation of joint learning experiments, additional research questions concerning which auxiliary tasks lead to positive transfer - and why - were answered: 3 of the 7 auxiliary tasks led to positive transfer, including depression sign classification under a simplified taxonomy, fine-grained emotion classification and sentiment classification led to positive transfer, however none of a set of 12 task characteristics proved to be good predictors of said positive transfer

Keywords: Depression, NLP, social media, mental health, multi-task learning, machine learning,low resource.

List of Figures

List of Tables

List of Acronyms

AUC Area Under the receiver operating characteristic Curve

ANEW Affective Norms for English Words

AVPs AVerage Precision score

BDI Beck Depression Inventory

CNN Convolutional Neural Network

CES-D Center for Epidemiologic Studies Depression Scale

DSM Diagnostic and Statistical Manual of Mental Disorders

FiLaMTL Figurative Language enabled Multi-Task Learning framework

GRU Gated Recurrent Unit

GBD Global Burden of Disease

HAM-D Hamilton Scale for Assessment of Depression

HCI Human-Computer Interaction

ICC Intraclass Correlation Coefficient

LIWC Linguistic Inquiry and Word Count

LSTM Long short-term memory

LLM Large Language Model

MLP Multi-Layer Perceptron

MTAN Multi-task Attention Network

ML Machine Learning

NLP Natural Language Processing

PAL Projected Attention Layers

PHQ Patient Health Questionnaire

PANAS Positive Affect and Negative Affect Schedule

PDP Possible Depressive Profile

SVM Suport Vector Machine

SMART SMoothness- inducing Adversarial Regularization and BRegman pRoximal poinT opTimization

TF-IDF Term frequency-inverse document frequency

WHOQOL-Bref World Health Organization Quality-of-life Scale

Contents

Chapter 1

Introduction

Depression can be defined as a clinically significant form of psychic suffering that can lead to many impediments to functionality, reduction in quality of life and, in severe cases, death given the increased risk of suicide. The condition impacts a large share of the global population, being the second biggest cause of years lived with disability and the sixth biggest cause of years of life lost due to disability, according to the 2021 edition of the Global Burden of Disease (GBD) study [1]. Efforts to meet the great demand for mental health services are currently insufficient: the World Health Organization's mental health action plan for 2013-2030 has as one of its principles universal coverage for mental health services, reporting that "the gap between the need for treatment and its provision is large all over the world.", especially in low and medium income countries, where 76% to 85% of individuals with severe mental disorders do not receive any treatment (World Health Organization, 2021)

Expanding the tool set for identification of depressed individuals is one (of many) strategies that can help overcome this lack of proper care. Instruments aimed at identifying depression can not only support healthcare professionals in directing patients to mental health services, but also authorities in the construction of public health policies, through the identification of at-risk populations. Traditionally, diagnosis instruments aim to identify signs that are characteristic of depression based on different psychopathological theories, the most prevalent of which is the 9 items Patient Health Questionnaire (PHQ) (KROENKE; SPITZER; WILLIAMS, 2001), which is based on criteria for diagnosis of major depressive disorder established by the Diagnostic and Statistical Manual of Mental Disorders (DSM) (American Psychiatric Association, 2022). These tools, which include psychometric scales, self-guided tests and similar instruments, provide standarized meth-

[1] https://www.healthdata.org/research-analysis/gbd-data

ods for diagnosis, thus facilitating and speeding up the process of depression detection. Despite demonstrating their worth through decades of clinical practice, its important to recognize that such resources have limitations in terms of reach, since they require that i) healthcare professionals consider their application necessary or ii) that depressed people seek help by themselves, despite impacts to functionality and social stigma associated with the condition.

Given the ample adherence to social media by the global population, as well as the public nature of and easy access to data published in these platforms, research that investigates the use of content produced by social media users to combat depression has become an active area of interest, aimed at complementing traditional efforts for identifying, triaging and treating depressed individuals. Among such efforts, the use of Natural Language Processing (NLP) for the classification of depression is of particular note, given that texts produced by social media users can demonstrate either explicit signs of depression, in the form of self-declarations of depression, as well as more subtle characteristics like variations in language style or topics most often discussed.

The task of depression detection in the field of NLP applied to mental health is usually formulated as a binary classification problem, in which a document is labeled as positive (indicative of depression) or negative (without meaningful depression signifiers). A positive characteristic of this approach is its compatibility with automatic *corpus* labeling strategies, avoiding the need of manual annotation, which tends to lead to small datasets (especially when the task requires specialized annotators, which is often the case in the mental health domain). In these cases, automatic labeling of the positive class is typically done based on criteria such as participation in certain online communities, presence of metadata (e.g. tags) associated with depression, depression diagnosis in electronic medical records or occurrence of text snippets associated with depression self-reports (e.g. "I was diagnosed with depression"); while the negative class is attributed to a group of posts/users that do not meet such criteria, assuming not necessarily a complete lack of depression signs, but a lesser frequency of those in comparison to the positive class. These techniques result in bigger corpora, facilitating the use of deep learning techniques, and generally lead to well-performing models with good generalization capabilities.

Binary classification approaches to depression detection in social media posts have already demonstrated their usefulness in various contexts, such as automatic moderation, triaging and analysis of depression incidence in populations (CHOUDHURY et al., 2013; MILNE et al., 2016), being tools capable of reaching people that would be otherwise neglected. However, as opposed to traditional evaluation instruments, these models do not try to capture how depression is expressed in a case by case basis. This does not necessarily means that such models are incapable of identifying different signs of depression – various studies have observed that correlations between first person usage and depression found in models can be linked to the phenomenon of rumination, for example (NAMBISAN et

al., 2015; CHOUDHURY et al., 2013) – just that alternative approaches might help fill a knowledge gap in the collective understanding of how depression signs are expressed in online text and which of them can be effectively classified by techniques commonly found in the literature. Moving towards a better understanding of the topic can open the way to new applications for combating depression using NLP or – in case of the discovery of limitations of current approaches – point to future directions. Ideally, models should be capable of discerning between the various ways that depression can be expressed, mirroring the complex process of diagnosis and treatment.

Modeling the problem of depression classification as a multilabel/multiclass task is one possible way of bridging the gap between models and current medical practice, as well as providing the opportunity for finer grained (and thus presumably more informative) predictions. A smaller portion of studies investigate the application of NLP for the identification of depression symptoms (MOWERY; BRYAN; CONWAY, 2015; YADAV et al., 2020; YAZDAVAR et al., 2017; UBAN; CHULVI; ROSSO, 2022). These efforts typically base their label schema on the PHQ-9 (KROENKE; SPITZER; WILLIAMS, 2001). Using pre-established criteria validated by clinical practice ensures that resulting models deliver useful information about a user's mental state (provided that models are well performing), however it must be considered that instruments like the PHQ-9 were designed with clinical context in mind and might not be directly applicable to online text content (for example signs related to agitation cannot be evaluated by visual observation if the only provided information is text). It is also possible that online texts contain signs of depression that are not typically evaluated by traditional tools, such as external factors that can exacerbate/alleviate psychic suffering – for example financial issues and lack of a support network are potential risk factors, while ongoing treatment and self-care practices are protective factors.

Aiming at dealing with the limitations of previous works, this study evaluates a dataset of social media posts labeled with a set 21 signs of depression, designed by a collaborative process involving a multidisciplinary team of experts and aimed at providing a better understanding of how depression signs are commonly expressed in social media, and which of these signs can be predicted with sufficient performance. Of these signs, 18 are depression symptoms and 3 are additions signs, composing a fine grained label schema.

The resulting dataset is relatively small in comparison to other studies, with only 780 annotated posts. This low data volume is a result of the manual labeling by individuals with a proper degree of instruction in mental health topics, and raises additional questions with regards to the role of additional strategies such as data augmentation and semi-supervised learning as a way to make the training of classifiers with sufficient generalization capabilities viable, a common theme in studies that concern themselves with the multilabel/multiclass approach to depression detection (YAZDAVAR et al., 2017; YADAV et al., 2020) .

With these considerations in mind, the following sections contextualize and report observations obtained through experiments which aimed to answer the following research questions:

- ❏ RQ1: Which of the 21 signs of depression can be frequently found in online text?
- ❏ RQ2: Which of these signs can be most easily identified using machine learning techniques?

Research questions RQ1 and RQ2 are answered in Chapters 4 and 5 respectively.

In order to deal with the low data volume issues raised during the training and fine-tuning for the task of depression sign classification, multi-task emerged as an interesting approach to follow up. Thus, two research questions related to multi-task learning were defined:

- ❏ RQ3: Can multi-task learning improve performance for the fine-grained depression classification task? If so, are there particular signs that benefit from it? Which of the auxiliary tasks improve performance for the primary task?
- ❏ RQ4: Are there particular characteristics that good auxiliary tasks have in common? Can they be observed prior to training?

Research questions RQ3 and RQ4 are more thoroughly introduced in Chapter 2 and answered Section 5.4.

The remaining sections of this study are organized as follows: Chapter 3 describes relevant related work in the domain of NLP applied to mental health, Chapter 4 describes the methods and resources utilized in computational experiments, including an exploration of the collected dataset that aims to answer some of the research questions posed. Chapter 5 details experimental setup and results. Chapter 2 concerns key concepts, related work and discussion of computational experiments regarding the multi-task approach.

Chapter 2

Multi-task learning Key concepts

Multi-task learning (CARUANA, 1997) has proven to be an effective tool in many NLP applications, including sentiment classification (WANG et al., 2018), question answering (JOTY; MÀRQUEZ; NAKOV, 2018) among others (CLARK et al., 2019; LIU; JOHNS; DAVISON, 2019). It consists of the joint learning of two or more tasks by a single model in order to, by optimizing multiple loss functions, produce a model capable of extracting information from text that is relevant for all the tasks it was trained on. Since its inception, multi-task learning has been shown to have a positive effect on task performance when compared to a standard training regimen. While the mechanisms that dictate when and why certain task groupings benefit from multi-task learning remain unclear, there is consensus on the notions that this boost in performance comes from a regularizing effect (which leads to better generalization) and that a group of tasks tend to benefit each other in a multi-task regimen when they are sufficiently similar.

Given the potential benefits of multi-task learning, it is possible that its use for the task of fine-grained depression classification might not only mitigate problems associated with low data volume (by leveraging other datasets) but also improve model performance on common error cases, such as the difficulty in interpreting figurative language. As alluded to in section 5, multi-task learning has the potential to introduce varied and high-quality additional data in a model's training regimen without requiring time and resource intensive expert annotation efforts. In order to apply such techniques, it is necessary to define a set of auxiliary tasks that are potentially useful to the primary task of depression sign classification and do not require additional manual labeling. This additional data was looked for in either publicly available datasets, or by adopting automatic annotation policies to unlabeled data. The remaining sections give a brief introduction of the key concepts underpinning multi-task learning (Section 2.1), present additional related work

relevant to multi-task learning and it's application on NLP + mental health (Section 3.1, Section 3.2), introduce the selected auxiliary tasks (Section 5.3), and detail computational experiments on the impact of multi-task learning for the fine-grained depression sign task and their results (Section 5.4). Since this work represents a large portion of the efforts spent, and tackles a somewhat unexplored domain, an additional set of research questions was proposed:

- RQ3: Can multi-task learning improve performance for the fine-grained depression task? If so, are there particular signs that benefit from it? Which of the auxiliary tasks improve performance for the primary task?
- RQ4: Are there particular characteristics that good auxiliary tasks have in common? Can they be observed prior to training?

2.1 Key concepts

This section aims to introduce core theoretical foundations of multi-task learning. Since the goal is to improve model performance on the task of fine-grained depression sign classification, the reader should become familiarized with common challenges related to multi-task optimization, ways of interpreting the positive transfer phenomenon and some of the most common neural architectures and algorithms observed in the literature. Note that multi-task learning can be considered a joint learning based approach, alongside meta-learning and other approaches, but since these are not evaluated in this study, multi-task learning might be referred as joint learning interchangeably from here on out.

Multi-task learning can be understood as the training of a machine learning model through the optimization of multiple tasks simultaneously (RUDER, 2017). For the purpose of this study, a task is characterized by a loss function and a dataset (X). This dataset can in turn be understood as a combination of a value distribution ($p(X)$) and a distribution of label/result values (Y) associated to these input values ($p(Y|X)$). Thus, a task belonging to a task set $T_i \in T$ can be defined as:

$$T_i = \{p_i(X), p_i(Y|X), \mathcal{L}_i\} \tag{1}$$

This means that it is possible to build a set of tasks T in a myriad ways. Tasks can share the same input set X, but differ on label values. In this case, the label set is usually different, modeling different problems (e.g object segmentation and depth information for the same set of images). Tasks can also model the same problem, as long as their respective Y belong to different distributions (e.g. a recommendation system that models each user as a distinct task). It is possible to construct a set of tasks that do not share X, which is of particular interest to the depression sign classification task, since it enables the use of already collected and labeled data. Finally, it is possible for two tasks to

have identical datasets, but different loss functions associated with them, although this is an uncommon approach, since the choice of loss function is usually determined from the nature of a particular task (regression, classification, reinforcement) and data distribution.

The intuition that underpins multi-task learning as stated in the original paper (CARUANA, 1997) is that it "improves generalization by leveraging the domain-specific information contained in the training signals of related tasks". Note that since its introduction, multi-task learning was regarded as a form of regularization that is achieved by trying to obtain shared latent representations between tasks learned in parallel. A complementary framing involves understanding joint learning as a form of inductive transfer: similar to fine-tuning, Ruder (2017) affirms that:

> We can view multi-task learning as a form of inductive transfer. Inductive transfer can help improve a model by introducing an inductive bias, which causes a model to prefer some hypotheses over others. For instance, a common form of inductive bias is L1 regularization, which leads to a preference for sparse solutions. In the case of MTL, the inductive bias is provided by the auxiliary tasks, which cause the model to prefer hypotheses that explain more than one task. As we will see shortly, this generally leads to solutions that generalize better.

The decision in favor of a multi-task setup is often motivated by the need for improved performance in a particular task, in which case a distinction is made between the target task, called "primary task", and the tasks whose training signals might benefit the primary task through positive transfer, called "auxiliary tasks". This is the case of our proposed training setup, in which performance on the primary task of fine-grained sign classification is the main concern. It must be noted that auxiliary tasks might also benefit from the training signals from the primary task or other auxiliary tasks, and that in some cases no distinction is made between tasks and the goal is an improvement in overall performance.

It is possible to establish parallels between multi-task learning approaches and other techniques, such as transfer learning (in which the pre-training is executed with one task, and fine-tuning is done on another task), meta-learning (which is based on sampling instances from a set of tasks) and domain adaptation (which aims to improve performance using data from a different distribution to the source dataset). All these techniques are concerned with the phenomenon of positive transfer between tasks (that is, an improvement in performance resulting from the use of multiple tasks), given the task definition in Equation 1.

Because of these similarities, although it is possible to define multi-task learning succinctly as the process of optimizing for more than one task, it is important to note what distinguishes this approach from others, for the sake of disambiguation. Contrary to transfer learning, multi-task learning in done in parallel, which means models have ac-

cess to the training signals of all tasks during the whole training process. The difference between multi-task learning and meta-learning is a subtler one, and mainly derives from their objectives: while multi-task learning is concerned with improved generalization capabilities for the narrow set of tasks a model was trained on, meta-learning is concerned with improved generalization on unseen tasks (HOSPEDALES et al., 2022). In practice, meta-learning models learn meta-parameters to improve the ability of models to efficiently adapt to new tasks. Finally, with regards to domain adaptation: some techniques fit the criteria for multi-task learning given the proposed task definition, but only those that jointly learn patterns from both source and target domains simultaneously (PENG; DREDZE, 2017).

Certain factors can hamper a model's ability to learn latent representations that are useful for a given set of tasks. One of these factors is negative transfer, in which the training signals associated with a given task negatively impact model performance on other tasks (measured by ablating said task from training and comparing results). One way of interpreting this phenomenon is as a consequence of difficulties in the optimization process: if the gradients associated with a given task point to an opposite direction in relation to those from another task, it can "nudge" the model away from learning adequate representations for this other task, especially if said gradients are also of a bigger magnitude, dominating the training process (YU et al., 2020; WANG et al., 2021; LIU et al., 2021; NAVON et al., 2022).

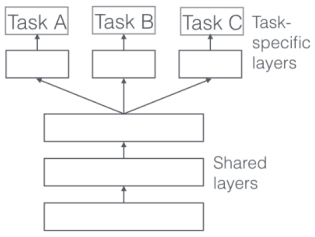
An essential design decision that leads to good performance in multi-task models is the choice of an appropriate task set. There is currently no consensus on what characteristics are associated with useful auxiliary tasks, save for the notion that these tasks should be in some way similar among themselves (and specially the primary task, if it exists). Some studies demonstrate correlation between positive transfer and task characteristics, like entropy and kurtosis (ALONSO; PLANK, 2017), similarity metrics among datasets based on covariance (WU; ZHANG; RE, 2020) and correlation analysis between label/embedding pairs across tasks (SCHRÖDER; BIEMANN, 2020), but these have a limited scope, investigating a small set of tasks/techniques. These tasks can exhibit different behavior, leading to different conclusions – for example, the experiments conducted by Alonso e Plank (2017) show correlation between information theoretical dataset characteristics and their associated task's utility in multi-task learning, while Bingel e Søgaard (2017) conclude that this type of information cannot be used to estimate behavior during joint learning, suggesting that features derived from gradient behavior during training are better predictors of performance. At time of writing, no work that provides theoretical guarantees about the impact of a task set on a multi-task model was found, save for training the model and analyzing the resulting performance, at least in the domain of deep learning. Yu et al. (2020) came the closest with proof of negative transfer under specific conditions, but with the caveat that such guarantees only apply to convex parameter

optimization landscapes and that models that do not meet the proposed criteria during training might still demonstrate negative transfer.

Another factor that influences the ease with which a model can learn adequate shared representations is its capacity: models with a sufficiently large number of parameters can learn distinct representations for each task that are completely segregated in latent space (WU; ZHANG; RE, 2020), in effect nullifying the regularizing effect of multi-task learning (which means that these occasions can be understood as the model "ignoring" the inductive bias introduced by the joint learning approach). The occurrence of such phenomenon is an indicator that tasks are incompatible, however under certain circumstances it is possible that lower capacity models can leverage the information from multiple tasks, since the smaller number of parameter "forces" the model to learn shared (and thus, potentially more general) representations and lead to better performance. Given these remarks, affirmations about the effectiveness of a given task set should take into account the impact that the number of trainable parameters has on model performance.

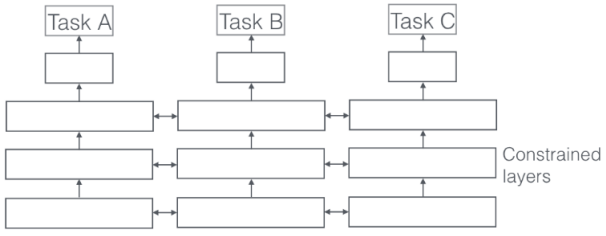
Among possible ways of categorizing multi-task learning techniques, the most common distinguishes between hard parameter sharing and soft parameter sharing techniques. Architectures that employ hard parameter sharing utilize the same subset of model parameters for different tasks. Soft parameter sharing techniques on the other hand, have segregated sets of parameters for each task, introducing information from other tasks indirectly, for example by penalizing the distance between the norms of parameter tensors in each layer. Hard parameter sharing approaches are the most common approach, possibly due to being more efficient in terms of the number of trainable parameters (especially important considering that jointly optimizing for multiple loss functions can be computationally demanding, depending on setup). Figures 1 and 2 bring examples of neural architectures for hard and soft parameter sharing respectively.

Figure 1 – Hard parameter sharing network



Source: (RUDER, 2017)

Figure 2 – Soft parameter sharing network

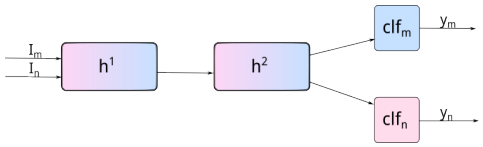


Source: (RUDER, 2017)

A common characteristic of joint learning approaches is their sensitivity not only to the task set and model capacity, but to the chosen architecture and set of hyper-parameters. This phenomenon is clearly illustrated by a simple "tree-like" architecture. As the name suggests, these models are composed of a "trunk" of shared parameters and branches

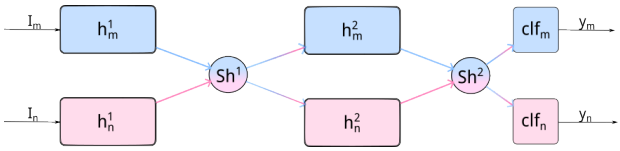
out into task specific layers for each task (generally a single dense layer for classification purposes)[1]. The choice of which layer of the network will be the splitting point between shared and task-specific parameters has great impact on the performance of these models. Intuitively, if too many parameters are shared, a model might be unable to extract task-specific characteristics from the shared latent representations, and if too few parameters are shared, the potential benefits of joint learning are diminished, limited to a small part of the network. In order to avoid the search of an appropriate architecture, some techniques aim to learn which parameters should be shared during training. Sluice (RUDER et al., 2019) and cross-stitch (MISRA et al., 2016) networks use separate models for each task, in a similar manner to soft parameter sharing, but connect them through linear combinations between representations. A similar approach reduces the overhead brought about by the training of various models in parallel by selecting features for each task using attention mechanisms (LIU; JOHNS; DAVISON, 2019), though this approach is still significantly more costly than single task training. Figures 3 and 4 illustrate a tree-like architecture and a cross-stitch network respectively. It should also be noted that there is no guarantee that the aforementioned techniques will converge to optimal parameter sharing, since they are subject to the same optimization difficulties as other approaches.

Figure 3 – "Tree-like" network



Produced by author

Figure 4 – Cross stitch network



Produced by author

Independently of the selected technique, a multi-task model has to optimize for multiple loss functions. How to treat this optimization also has a big impact on a model's overall performance. Typically, a joint loss is optimized, which is calculated by aggregating separately computed task losses. This joint loss is usually a weighted sum of each loss function, such that the weights associated to each task loss are considered hyper-parameters. It is also possible to treat each task loss independently, calculating model updates with regards to each task loss according to some logic (for example by alternating between tasks in each batch, or sampling from a task distribution). These task sampling approaches are functionally similar to a weighted sum joint loss if gradient accumulation is used. More complex joint loss functions aim to determine task weights automatically based on each task's characteristics, such as uncertainty (KENDALL; GAL; CIPOLLA, 2018), average loss during training (LIU; JOHNS; DAVISON, 2019), and even change

[1] It should be noted that there is no properly defined nomenclature for this kind of architecture (since it is the most basic one, it is often simply denoted as "multi-task" in work that is not concerned with other joint learning techniques), so the term "tree-like" will be adopted for the remainder of this study.

task weights dynamically via gradient descent, with the objective of normalizing all losses in order to make the training robust to outliers (LIU et al., 2020), avoiding the need to conduct hyper-parameter search (which can be costly).

Chapter 3

Related work

Among the various studies that investigate the application of NLP in the mental health domain, those that report insights concerning the ways in which social media users talk about depression in text were selected for reporting in this Chapter, specially those studies that aim to explicitly model signs of depression established by the clinical practice and literature. By identifying characteristics in text and user behavior that can distinguish users afflicted by mental health disorders from healthy individuals, it is possible to obtain a firmer grasp on what expressions of depression might be more frequent in social media (and should therefore be properly covered by depression classifiers). Additionally, this chapter aims to exemplify common techniques found in the broader literature of NLP + mental health.

Choudhury et al. (2013) is a pioneering work in the field of automatic depression detection in social media, and is illustrative of some of the key discoveries concerning how depressed users express themselves online. The study collected Twitter posts authored by participants of a Mechanical Turk crowdsourcing task, accompanied by associated scores in psychometric scale questionnaires. Since crowdsourcing task data is noisy – participants can answer too quickly in order to maximize the financial return when engaging with the platform, producing potentially inconsistent data – and aiming at mitigating this noise the authors applied two questionnaires: Center for Epidemiologic Studies Depression Scale (CES-D) (RADLOFF, 1977) and Beck Depression Inventory (BDI) (BECK; STEER; BROWN, 1996). If the questionnaire results were not congruent, user data was not collected. After filtering, 476 users with more than a year of post history were taken into account and all of their posts were collected. 171 of the selected users were considered as depressed for the purposes of the study. It is worth noting that despite the fact that the chosen instruments evaluate a set of depression symptoms, this information was not

taken into account during data analysis or modeling.

The authors utilized a feature engineering approach for model training. The feature set includes (i) behavioral factors: frequency of posts, time of posting and engagement; (ii) social graph data: number of followers/followed accounts, density and reciprocity of connections with other users; (iii) emotion: positive and negative sentiment, dominance and arousal values, extracted with Linguistic Inquiry and Word Count (LIWC) and Affective Norms for English Words (ANEW) respectively; (iv) linguistic style: occurrence of LIWC categories, primarily related to syntax; (v) occurrence of words related to depression - extracted from a Term frequency-inverse document frequency (TF-IDF) representation fitted on a supplemental corpus of Yahoo Answers posts tagged as "mental health" - and antidepressives (manually constructed).

Analysis of linguistic features indicates that depressed users tend to write in first person more often (possibly indicative of excessive focus on oneself, a phenomenon called rumination), express negative sentiment and use swear words frequently and utilize depression related terms more than non-depressed users. Common themes expressed by depressed users were extracted through a crowdsourcing task where participants were asked to manually assign posts to different groups. The resulting grouping consists of symptoms (e.g. anxiety, nausea, insomnia, irritability), self-reports of depression, discussions about ongoing treatment and descriptions of relationships or other aspects related to a user's personal life.

Other studies established novel relations between social media content and the behavior of depressed patients observed in clinical practice. Schwartz et al. (2014) demonstrated that it is possible to observe seasonal patterns of depression through a regression model capable of outputting scores for 6 personality aspects of neuroticism (tension, depression, frustration, guilt, self-consciousness) in Facebook posts. Additionally, the authors modeled 10 topics from posts produced by depressed users, reporting language indicative of despair (e.g "hopeless", "helpless"), solitude (e.g. "lonely", "rejected") and corroborating the link between depression and swearing and insomnia.

A relevant factor to consider when interpreting results is that most datasets are collected from a single source of data, be it one social media platform or specific online communities. Studies often present conclusions about depression in social media without acknowledging that their findings might not generalize to other domains. While these works often do corroborate each other, that is not always the case and a potential cause of such discrepancies is this difference in domain. Ji et al. (2018) evaluated two datasets extracted from different social media platforms (Reddit and Twitter) with different methodologies for automatic labeling – participation in certain communities for labeling depression related posts in the Reddit dataset and presence of keywords for the Twitter dataset – in order to train a suicidal ideation classifier. The authors observed differences in the ways users express their symptoms: while Twitter users utilize direct

and aggressive language, making direct mention of their suicidal ideation and somatic symptoms, Reddit users tented to talk about their depression in narrativized form, contextualizing their sentiments and discussing their financial, familial and social situation and reporting their subjective experience of depression such as anhedonia. The study also points to similarities between the two datasets, such as the increased frequency of first person usage in suicidal users, suggesting that despite domain differences, there is a set of common characteristics to all text produced under depression. Understanding how different social media platforms might influence the content produced by its user base is especially relevant considering that Twitter is by far the most frequent data source for dataset construction.

Some studies evaluate the application of NLP in the domain of mental health for the Portuguese language. Santos, Oliveira e Paraboni (2023) developed a publicly available corpus for the classification of depression and anxiety at a user level, called SetembroBR. This corpus consists of posts from Twitter users accompanied by information regarding depression diagnosis and social network information (the users mentioned/followed in the collected tweets). The authors used depression and anxiety diagnosis self-reports as a criteria for a user's inclusion in the positive class, but only after manual validation of these reports. For control group data collection, random users were sampled while matching dates of publication, gender and number of posts of the positive class cohorts and keeping a proportion of 7 control group users to every positive class user. After data collection, only users for which the date of diagnosis could be determined were maintained, in order to filter out any posts made post-diagnosis. Furthermore, popular accounts (those with more than 10000 followers and individuals that self-reported other conditions (such as bipolar disorder, borderline syndrome, autism or schizophrenia) were discarded. Among the evaluated methods – logistic regression, Convolutional Neural Network (CNN), Long short-term memory (LSTM) and fine-tuned BERT models with a Bi-LSTM classification head – the best performance was obtained with the fine-tuned BERT model (macro averaged F1-score of 63% for the depression classification task and 61% for the anxiety classification task).

Additionally, the authors were concerned with the presence of explicit mentions of health-related information, such as "I need to see a psychiatrist" or "I am having anxiety episodes on a daily basis", which might simplify the task at hand in an undesired manner, biasing the model. Thus, they removed all tweets that contained mentions of depression, anxiety, treatment or medication, and reran the experiments, achieving comparable results, with a small decrease in performance, suggesting that resulting models do not rely on explicit self-reports in order to properly identify anxious and depressed individuals.

Casani et al. (2021) researched multi-class depression symptom classification in Portuguese Twitter posts. The authors proposed a simplified taxonomy on which the classification task is based. This taxonomy consists of 3 symptom categories: psychological,

physiological and behavioral symptoms, as well as a neutral category. This approach to symptom classification is, based on the evaluated information, novel in the domain of NLP applied to mental health, and in this present work it is also followed as a potentially useful way of grouping and interpreting results in more complex symptom sets. An initial set of 200 sentences representative of the 3 symptom categories was built by a mental health specialist, and this dataset was used to train the model. Additionally, Twitter posts were collected for the purposes of model training. These posts were manually labeled by a mental health professional, resulting in 2008 annotated posts. The evaluated approaches consisted of combinations in a simple set of features – TF-IDF and bag-of-words – and techniques – Suport Vector Machine (SVM), Multi-layer Perceptron and Naive Bayes. These models still achieved good performance, with Area Under the receiver operating characteristic Curve (AUC) above 0.9. No information concerning performance for individual symptoms was reported, but considering that the data is mostly balanced (save for a moderate imbalance in physiological symptoms, which has around 300 instances instead of 500), it can be inferred that the model performs well for all symptoms.

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 needs 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. This approach proved to be competitive with fully supervised approaches, achieving better performance than fully supervised baselines in 5 out of 9 symptoms, as well as better general accuracy, with improvements for “decreased pleasure”, “feeling down”,“sleep disorder”, “loss of energy” and “change in appetite”. The proposed approach only struggled with the “hyper/lower activity” and “concentration problems” symptoms, achieving F1 scores of 30% and 38% respectively, however baseline models did not demonstrate these same difficulties, which suggests that all symptoms can be discerned via text processing. The authors conclude that there are significant differences in topic preference between depressed and non-depressed users, suggesting that the PHQ-9 categories are, as a whole, useful for discerning depression in social media text.

Lee et al. (2021) introduced a micromodel based architecture that focuses on explainability, utilizing mental health tasks as a case study. Inspired by microservices architectures, the model is composed of a collection of classifiers capable of identifying individual relevant characteristics in single utterances that are then aggregated and used as input for a task-specific Explainable Boosting Machine classifier. For user level depression

classification, those included PHQ-9 symptoms and cognitive distortions. To avoid manually annotated data for each micromodel, the authors built a collection of representative utterances for each task taking into account lexicon based queries, and then utilized a pretrained BERT (DEVLIN et al., 2018) model to find semantically similar utterances, thus generating the necessary datasets for each task. The resulting models achieved competitive results for the CLPych2015 and CLPsych2019 shared tasks (COPPERSMITH et al., 2015; ZIRIKLY et al., 2019) while remaining easily interpretable. These results indicate that highly curated, low volume datasets can be used to bootstrap classification models for a wide variety of mental health related tasks, simply by looking for similar answers in a large unlabeled corpora, which consequently implies that users tend to express the same symptom in similar ways.

Table 1 – Summary of related work

Work	Texts	Language	Main results
(CHOUDHURY et al., 2013)	Twitter	English	Frequent use of first person, negative sentiments, swear words, discussion of relashionships and treatment
(SCHWARTZ et al., 2014)	Facebook	English	Association with insomnia, swear words, despair and solitude. Positive correlation with aspects of neuroticism
(JI et al., 2018)	Reddit, Twitter	English	Ways of expressing depression can vary between platforms, but there are commonalities between them
(SANTOS; OLIVEIRA; PARABONI, 2023)	Twitter	Portuguese	Depression and anxiety can be identified without relying on explicit mention of health-related terms
(CASANI et al., 2021)	Twitter	Portuguese	Depression symptoms can be grouped into psychological, physiological and behavioral, each one frequently observed in social media posts
(YAZDAVAR et al., 2017)	Twitter	English	Significant differences in PHQ-9 based topics between depressed and non-depressed users
(LEE et al., 2021)	CLPysch 2019 and 2015	English	Expressions of a given symptom or cognitive distortion are expressed in similar ways

Source: Produced by author

3.1 Multi-task learning in mental health applications

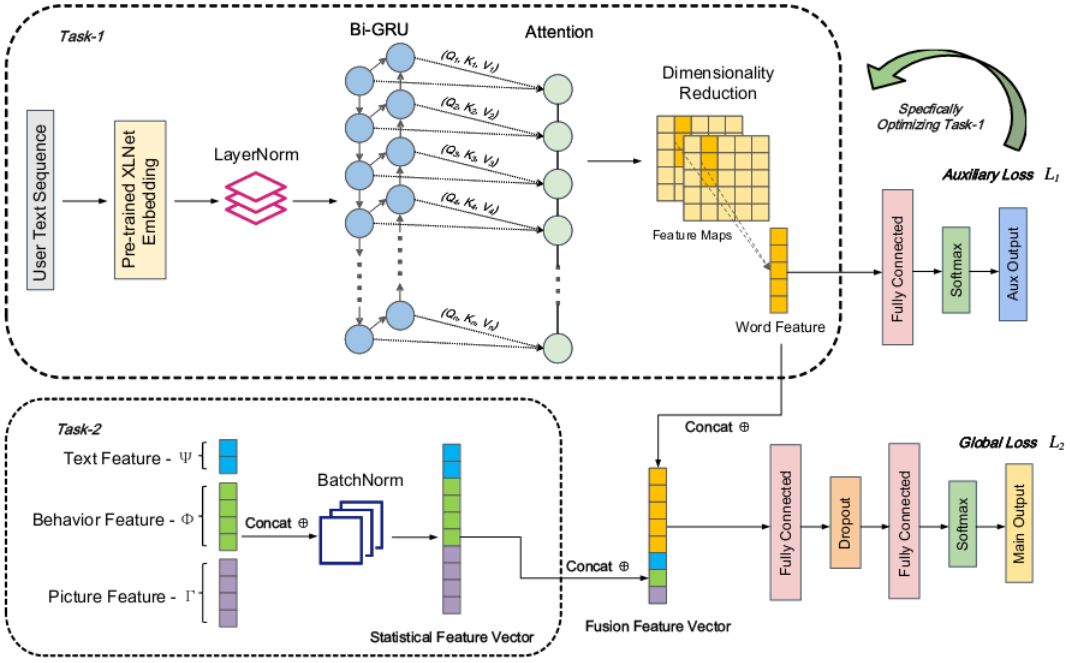
Some studies investigate how multi-task learning can be used in the mental health domain. This section reports related work in order to obtain a better understanding of which auxiliary tasks and techniques can improve results.

Wang et al. (2020) propose a model for the binary user-level depression classification in the social media platform Sina Weibo (a micro blogging platform, similar to Twitter/X) using a simple tree-like multi-task architecture. The task set is somewhat unusual, consisting of two tasks with identical loss functions and label data, but different features. The rationale behind this architecture was using the regularization properties of multi-task learning in order to avoid losing information from the pre-trained model during fine-tuning (commonly referred to as catastrophic forgetting). In this case, the auxiliary task had to predict a user's depression status using only embedding data – extracted from XLNet (YANG et al., 2020) followed by a single Bi-GRU block – while the primary task utilized additional features, including visual (tone, saturation and brightness of profile picture), behavioral (time and frequency of posts, ratio of original vs shared posts, ratio of posts published at late hours) and textual (proportion of texts with negative sentiment, frequency of depression-related words). Both the XLNet and Bi-GRU blocks are shared between tasks. Authors report a gain of 1.21% in comparison to single-task baselines, achieving a F1-score of 97.72%. Figure 5 is a representation of the proposed model.

Xu et al. (2019) investigated multi-task learning for early prediction of depression in the elderly population. Features extracted from questionnaire data, including general health information (if an individual smokes, has a mental health record, diabetes, etc.), demographic (gender, age, marital status) and socioeconomic data (status of retirement) collected over time. The task set consisted of two tasks: binary classification of individuals as at-risk of depression (would develop depression in up to two years after prediction) and a regression task of CES-D score at a particular point in time. The chosen architecture was a simple tree-like multi-task network, with a shared LSTM block (since both tasks receive time series data) and fully connected dense layers for the task-specific parameters. This architecture resulted in small, but statistically significant improvement over single-task baselines (SVM, Multi-Layer Perceptron (MLP), dynamic bayesian networks and LSTM), with a gain of 0.5% AUC for the binary classification task (from 86.8% to 87.3%) and a reduction of 0.035 in the mean absolute error for the score regression task (from 1.322 to 1.287).

Similar to this study's proposed use case for multi-task learning, Liu et al. (2021) aimed to identify depression in undergraduate students with limited data – albeit using audio clips instead of text. The authors utilized the classification of a speaker's gender as the auxiliary task, in order to introduce an inductive bias that would facilitate the learning of latent representations capable of distinguishing between genders despite low data volume. This paper utilizes a modified version of the usual tree-like multi-task architecture, with

Figure 5 – Network architecture (WANG et al., 2020)



(WANG et al., 2020)

shared layers consisting of one CNN and one Bi-LSTM block, but no explicitly separate task-specific layers, instead utilizing an attention block that takes the current task as a parameter and acts as a feature selector – somewhat similar to the approach proposed by Liu, Johns e Davison (2019), but restricted to the last layer.

Yadav et al. (2020) propose a multi-task learning architecture in order to improve performance in the primary task of PHQ-9 symptom classification in online posts. Particularly, the author's goal was to reduce the frequency of a particular error case: posts that utilize figurative language, since the semantic content of words in these posts can differ significantly from common parlance, and sarcasm in particular can invert the meaning of expressions. As mentioned in Section 4, the use of figurative language is a common way for users to express their feelings online. The authors chose a multi-class figurative language classification task as an auxiliary task, with a label set consisting of "metaphor", "sarcasm" and a generic category for other figures of speech. Additionally, the binary classification of posts as depressed/non-depressed was also utilized as an auxiliary task. A set of 12155 tweets was collected and subsequently annotated with the PHQ-9 symptoms by 4 annotators. These same annotators labeled the set according to the 3 figurative language categories and the binary depression classification task. Among the collected posts, 3738 (30.75%) were considered indicative of depression, among which 1485 (~40%) contained

figures of speech.

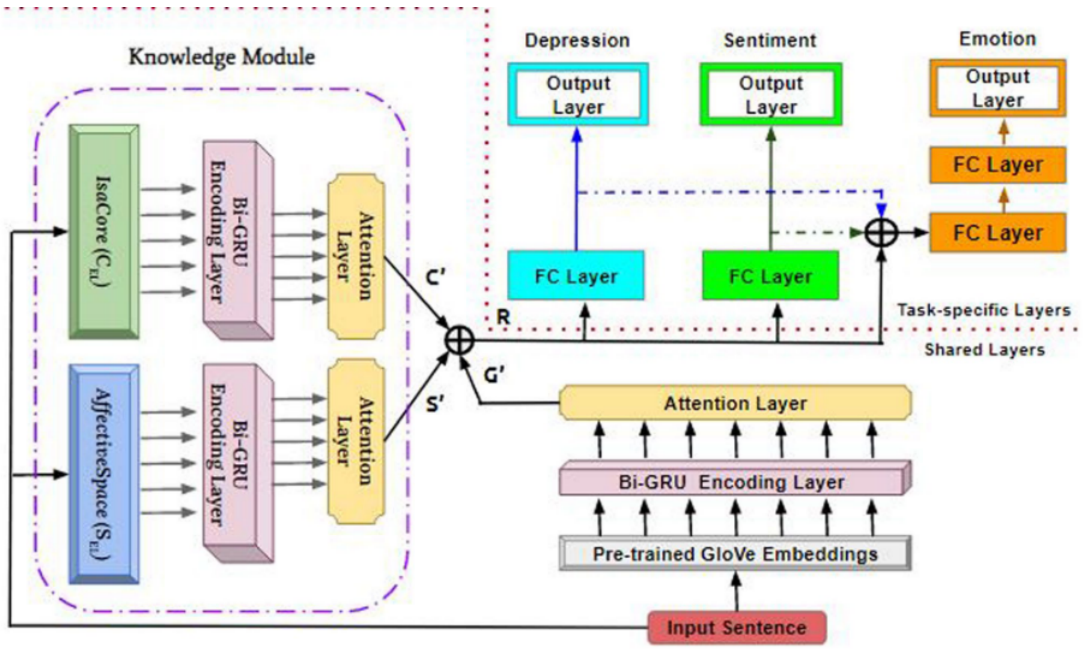
The proposed multi-task neural architecture, Figurative Language enabled Multi-Task Learning framework (FiLaMTL) is similar in nature to a sluice network (RUDER et al., 2019), but adapted for the fine-tuning of a BERT model. It consists of two models trained in parallel with parameter sharing being governed by learnable parameters that linearly combine the outputs of the self-attention layers of both models. This architecture was compared to single-task baselines (standard fine-tuning) and other established multi-task neural architecture baselines ("tree-like", cross-stitch, co-attention), and achieved better performance in terms of F1-score, precision and recall for all 3 tasks. The figurative language task achieved the biggest performance improvement (from F1-score of 67.09% in a single-task setting to 75.67%), but the primary task of symptom classification also improved (from 73.02% to 75.03%).

Ghosh, Ekbal e Bhattacharyya (2022) proposed the use of multi-task learning in order to improve a emotion classification task in the domain of mental health. The primary task consists of a multi-label emotion classification task in suicide notes, and two auxiliary tasks were proposed: binary depression classification and multi-class sentiment classification (with the standard positive, negative and neutral categories). The corpus was based on CEASE, a corpus of 2393 sentences extracted from suicide notes, which was expanded by the authors with 2539 new sentences collected from suicide notes available online, totaling 4932 annotated samples. The authors considered the following emotions and additional categories: Forgiveness, Happiness, Hopefulness, Love, Pride, Thankfulness, Abuse, Anger, Blame, Fear, Guilt, Hopelessness, Sorrow, Information and Instructions. The labels for the sentiment classification task were automatically labeled by mapping the emotion categories as positive, negative or neutral.

Three multi-task neural architectures were evaluated, which vary in terms of complexity. One model is a "tree-like" architecture, with the shared parameters consisting of a sequence of bidirectional Gated Recurrent Unit (GRU) blocks and an attention layer, and the task-specific parameters consisting of fully connected linear layers, with all tasks receiving pre-trained GloVe embeddings as input. The second architecture, referred to as "cascaded" introduces hierarchical relations between the auxiliary tasks (which are "simpler") and the primary task by feeding the output of some auxiliary task layers into the primary task as additional output. Finally, the authors propose introducing external knowledge to the model using IsaCore (CAMBRIA et al., 2012) and AffectiveSpace2 (CAMBRIA et al., 2015a) representations, which model common sense information that might not be represented in the collected texts, concatenating these features as additional input. Both the introduction of hierarchical relations and IsaCore and AffectiveSpace2 embedding led to better performance. Figure 6 showcases the final model. All models achieved better performance in comparison to a single-task baseline in the case of the primary task and the sentiment classification task, but negatively impacted the binary

classification task. Subsequent experiments demonstrated that the sentiment classification task had a positive impact in performance (+0.44% F1-score) and the binary depression classification task had a negative impact on performance (-0.22% F1-score).

Figure 6 – Final architecture proposed by (GHOSH; EKBAL; BHATTACHARYYA, 2022)



(GHOSH; EKBAL; BHATTACHARYYA, 2022)

Li, Braud e Amblard (2022) also investigated multi-task learning for the explicit purpose of mitigating issues related to lack of data using external corpora. The primary task consist of a binary classification task for dialogues between patient and therapist, for which three auxiliary tasks were selected: (i) identification of emotions according to the Ekman taxonomy for a list of categories) (ii) identification of dialog acts (which can be informative, question, directive or commissive) and (iii) identification of conversation topics (school life, culture/education, attitude/emotional, relationships, tourism, health, work, politics, finance and daily life) at a document level (the set of conversation turns associated with a given patient). All auxiliary tasks were associated with a corpus – DailyDialog, by Li et al. (2017) - distinct from the primary task corpus – DAIC-WOZ, by Gratch et al. (2014).

A hard parameter sharing approach was adopted, and since tasks are supervised at different level (either turn-level or document-level), the shared parameters were divided in two blocks: a Bi-LSTM block for processing turn-level information, and a subsequent block that processes the resulting states from the first block in order to obtain document-level information. Experiments demonstrated that all auxiliary tasks benefit the primary

task in comparison to a single-task baseline (F1-score of 43.9%): the addition of topic classification improves performance by a total of 11.5%, and act classification improved performance by 16.9%. The proposed model utilizes all tasks and achieves significantly better performance than the previous state of the art for emotion classification in this particular dataset (XEZONAKI et al., 2020), from 70% to 70.6% F1-score.

3.2 Related work on multi-task learning and what makes a good auxiliary task

This section presents related works that are in some way concerned with what constitutes an appropriate task set and what types of challenges these models can face during training.

Yu et al. (2020) propose a set of 3 factors that together lead to difficulties in the joint optimization process for multi-task models, called the "tragic triad". The authors state that a task can dominate the gradient descent based learning process of other tasks if and only if: (i) its gradient is in conflict with the gradients of the negatively impacted tasks, (ii) this gradient has a bigger magnitude than other tasks and (iii) the optimization space has a large curvature. In other words, a task can cause negative transfer if it leads the model to a minimum in the parameter space that is not shared by other tasks, and the model cannot "escape" this inadequate minimum because the gradient from this harmful task dominates the optimization process.

Two gradients are conflicting when the cosine of their angle is less than 0. Given two task gradients g_i and g_j, a multi-task loss function L and parameter vectors pre and post-update θ and θ', their magnitude similarity measures and curvature of the optimization space can be respectively defined as:

$$\phi(g_i, g_j) = \frac{2||g_i||_2||g_j||_2}{||g_i||_2^2 + ||g_j||_2^2} \tag{2}$$

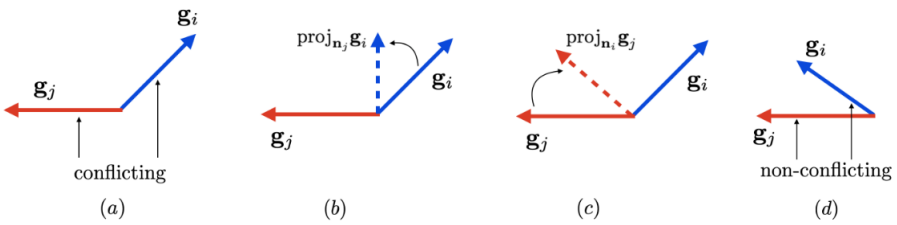
$$H(L; \theta, \theta') = \int_0^1 \nabla L(\theta)^T \nabla^2 L(\theta + \alpha(\theta' - \theta)) \nabla L(\theta) \tag{3}$$

The scale of ϕ is 1 when the magnitudes are exactly the same and tends to 0 as they differ. The curvature value H, however, can only be considered large or small depending on comparisons to other curvatures, with the authors defining a large curvature as a positive constant C without further elaboration.

The authors then propose a model agnostic gradient correction algorithm called PCGrad, which aims to eliminate the first condition in the tragic triad (conflicting gradients) by projection gradients onto the normal vectors of their conflicting pairs, as shown in Figure 7. Experiments with the datasets NYUv2 and CIFAR100 (both multi-label datasets adapted as multi-task) demonstrated PCGrad's ability to complement other

multi-task learning techniques, surpassing baselines in three strong baseline approaches: cross-stitch,Multi-task Attention Network (MTAN) and routing-based networks.

Figure 7 – Examples of PCGrad application a) shows conflicting gradients, b) and c) their projections, and d) the resulting gradients



(YU et al., 2020)

Alonso e Plank (2017) were interested in understanding what characteristics are indicative of a useful task. They selected a set of primary tasks (frame detection, named entity recognition and sentiment classification) and a set of simpler auxiliary tasks, which would normally be extracted as features (noun and verb phrase chunking, part of speech tagging, dependency extraction and log frequencies of words treated as discrete labels). An analysis of each task's characteristics was carried out by extracting information theory metrics from their respective datasets, including entropy (both all labels and excluding neutral category, if applicable) and kurtosis (which measures the assymetry of label distributions). Note that this choice of task characteristics implies a hidden assumption: the most important information contained in an auxiliary task is provided by its labels, or at least by the input/label relations. Experiments were conducted on a single "tree-like" multi-task architecture with LSTM layers as their shared component and dense layers for all task-specific parameters. It was observed that, in all cases, tasks that had low kurtosis and high entropy benefited the primary tasks, suggesting that measures that model in some way distributional information about an auxiliary task's dataset show correlation with improved performance. While these experiments are limited to a narrow set of tasks and a single architecture, they provide evidence that under certain conditions, it might be possible to identify useful auxiliary tasks before training.

Liu, Johns e Davison (2019) propose a meta-learning technique that generates a set of auxiliary tasks in order to improve performance of image classification models through multi-task learning without manually searching for an auxiliary task set. The proposed algorithm, MAXL, is similar to a popular meta-learning technique – MAML (FINN; ABBEEL; LEVINE, 2017) – being based on gradient descent and model agnostic, but is designed for a specific meta-representation: training an automatic classifier capable of generating new auxiliary task data using the same dataset of the primary task.

Initial experiments suggest that models trained with MAXL benefit from the introduction of a entropy regularization component in the meta-objective, encouraging high entropy in the auxiliary tasks, corroborating observations made by Alonso e Plank (2017), this time in the domain of computer vision. These experiments additionally indicate that introducing hierarchies between tasks leads to better performance. This hierarchical tasks are based on the primary tasks label schemes, subdividing a label into more specific sub-labels, and are constructed by a multiplicative gating mechanism in the softmax function of auxiliary task losses. Given a binary mask M_i associated with an auxiliary task $_i$ and the predicted label of the primary task $\hat{y}_i$, we have:

$$p(\hat{y}_i) = \frac{expM_i \odot \hat{y}_i}{\sum_i expM_i \odot \hat{y}_i} \tag{4}$$

The proposed algorithm was tested on a robust set of 7 image classification benchmarks: CIFAR-100, MNIST, ImageNet, CIFAR-10, CINIC-10, UCF-101. Only CIFAR-100 has a two-level hierarchical structure to its label set, and therefore was the only one that could be used to evaluate the quality of the automatically labeled auxiliary data. Three multi-task neural architectures were evaluated, with all being "tree-like" networks with different "trunks": a 4 layer CNN and two pre-trained models, VGG-16(SIMONYAN; ZISSERMAN, 2015) and ResNet-32 (HE et al., 2016). MAXL provided a significant improvement over baseline single-task performance for all datasets. The automatically labeled data also demonstrated itself competitive in comparison to manually labeled data in the case of CIFAR-100, and better than other baseline approaches (random classification and a K-means clustering strategy).

Ni et al. (2023) shared a similar preoccupation to the ones of RQ 4 and 5: do the insights reported in the literature apply to a more complex domain, and if so, which ones and how so? The authors investigated NLP applied to the financial domain, with a set of 6 tasks grouped into 4 distinct ability types: 2 sentiment classification tasks specific to financial news – Malo et al. (2014) classifies a news article as negative, positive or neutral and Cortis et al. (2017) models sentiment as a regression task from -1 to 1 – 2 numeric understanding tasks - one aimed at classifying a number as monetary value, percentage, time, etc. (CHEN; HUANG; CHEN, 2020) and another concerned only with classifying if a given number in a tweet is a monetary value (CHEN; HUANG; CHEN, 2019) - 1 that models causal relations between financial sector facts (LAMM et al., 2018) and 1 for semantic role tagging (MARIKO et al., 2020). It is possible to establish both qualitative parallels between tasks - for example, semantic role tagging can improve the number understanding tasks by improving the model's ability to discern what is a quantity - and quantitative ones - the authors compared cosine similarity of embeddings extracted from single-task models for each task and their gradient behavior during training, looking for similarities as proposed by Vu et al. (2020).

A pre-trained model specific to the financial domain, P-FinBVERT (ARACI, 2019)

was selected for baseline single-task training. Experiments were carried out with multiple task combinations, grouped by abilities. The results demonstrated that positive transfer does occur between abilities - the best results were obtained when multiple abilities were jointly learned. Additionally, the semantic role tagging task benefited all other tasks. 4 of 6 tasks were improved in comparison to the single-task baseline: the sentiment classification task (from 86.61% to 87.06% accuracy), both number understanding tasks (87.06%-87.79% accuracy for the multi-class task and 85.52%-86.52% accuracy for the binary classification task) and the causal relation task (77.12%-78.40% accuracy).

Task embedding analysis shows that the sentiment regression task had the lowest cosine similarity with other tasks, which could explain it being one of two tasks that did not improve with multi-task learning. The other task, however, was its complete opposite: semantic role tagging had high cosine similarity with other tasks (and improved them) but did not benefit from multi-task learning, illustrating that there might be asymmetrical relations of positive transfer that are not properly captured by cosine similarity, which is symmetric. The authors conclude that similar gradient behavior is not a good predictor of positive transfer. Finally, they investigate the impact of model capacity on different task sets by freezing model weights and training only attention-based adapters (STICKLAND; MURRAY, 2019a) of different sizes. They conclude that small capacities can hinder a model's ability to learn many tasks jointly (performance improves as more tasks are removed), suggesting that training with a large task set requires larger capacities, but also indicate that these multi-task models can be very efficient: a multi-task model trained with 99,8% fewer parameters still showed better performance in comparison to its single-task fully tuned baseline, concluding that this approach can have a regularizing effect, inducing more general representations that can be utilized by all tasks (a similar argument as the one proposed in Section 2.1).

Chapter 4

Resources and Techniques

This chapter describes resources and techniques employed in computational experiments reported on chapter 5, including the elaboration of the proposed label set for signs of depression, data collection and subsequent annotation of a corpus for the task of fine grained multi-label classification of these signs (section 4.1); an analysis of the gathered data (section 4.2); and introduction to the pre-trained models and techniques to be evaluated (section 4.3) .

4.1 Data collection and annotation

Posts were collected from public Facebook pages in order to build a dataset for the task of fine grained depression sign classification. These pages were of a particular kind: "university secrets" pages, in which students from Brazilian universities can vent their feelings and discuss their day-by-day undergraduate routine in a semi-anonymous fashion, through a process of submission of textual content via form, followed by approval and publication by a centralized account (usually operated by page moderators). This process is intended to obfuscate the original author. Considering the hypothesis that users can express depression in different ways depending on the platform, this data source has some characteristics that might encourage varied expressions of depression signs, such as the possibility of posting long-form content, the degree of anonymity and the fact that such pages are designed - in part - for emotional expression (though not necessarily negative emotions).

However, it is important to note that collecting from a single source of data is a limitation, one that is exacerbated by the fact that this dataset is focused on a particular cohort of the population (university students). On a related note, university students are a

demographic of particular interest given their higher incidence of depression (CAVESTRO; ROCHA, 2006; EVANS et al., 2018) and in the scope of the Amive[1] project it was the subject population.

Posts were collected using Crowdtangle (Crowdtangle Team, 2021) , a public insights tool owned and operated by Meta. A manual search for "university secrets" pages was carried out by going through a list of Brazilian universities. Afterwards a search was conducted, filtering relevant posts by keyword search ("suicide", "depression", "kill myself", "will to live", "cut myself", "want to die"). The interval for search was set to cover posts published from 2012 up to 2021, and all collected posts contained exclusively Brazilian Portuguese text (with the exception of the occasional use of loanwords). Given the sensitive nature of the content, a manual process of anonymization was conducted, substituting identifying information – including references to places, institutions, events, courses, usernames, links and dates – with generic tags (e.g. <university>, <city>). In the case of links and usernames, this process serves a dual function: while this information is generally removed from text as part of pre-processing, there is evidence that suggests that interaction between users can be a relevant factor for depression detection (CHOUDHURY et al., 2013), in which case links to other posts and user mentions can carry useful information and should preferably not be removed but rather standardized.

As a byproduct of the manual nature of the anonymization process, it was possible to identify three broad categories in the collected posts: i) posts where users talk about their own struggles with depression, ii) posts that tackle the theme impersonally, be it in relation to a particular event like suicide of a third party or a general worry with the health of a student body stemming from stressors such as pressure to perform, being apart from family, social isolation, unreasonable teachers, etc. and iii) posts that didn't contain mention of depression despite keyword filtering, often due to links to some pages that contained the substring "*da depressão*" (of depression) as part of their name. While it can be argued that posts in this last category can help with performance in texts from domains outside the topic of mental health, they carry little useful information about how users talk about depression online and were thus removed. In total 780 posts were selected for annotation.

The collected posts were relatively lengthy in comparison to Twitter posts, which are the most common data source found in recent literature (LIU et al., 2022a), with an average length of 178 words[2], keeping in mind that until recently tweets had a limit of 280 characters. These posts can vary greatly in size, with a standard deviation of 160 words. The shortest collected post is only 4 words long, while the longest has 949. The resulting dataset is small in terms of number of posts, but these posts are richer with information than average, assuming that longer texts provide more opportunities for depression signs

[1] <https://www.amive.ufscar.br/>
[2] As estimated using the Natural Language Toolkit (https://www.nltk.org/)

to be expressed.

The Amive project aimed to develop a set of solutions for depression detection and automatic interventions for depressed people. As part of Amive, a set of 21 signs related to depression was devised. Given the multidisciplinary nature of the project (and of the domain of NLP + mental health as a whole), both this set of signs and the accompanying annotation guidelines were produced by a committee of psychology, psychiatry, occupational therapy, NLP and Human-Computer Interaction (HCI) specialists. This committee was responsible for surveying the literature in search of tools and theoretical resources, in order to define as broad a set of depressions signs as possible, basing themselves not only on material explicitly designed for depression evaluation such as the PHQ-9 and Hamilton Scale for Assessment of Depression (HAM-D) (FREIRE et al., 2014), but also on other relevant information, such as the World Health Organization Quality-of-life Scale (WHOQOL-Bref) (SKEVINGTON; LOTFY; O'CONNELL, 2004) for quality of life measurement (which is impacted by depression) and Positive Affect and Negative Affect Schedule (PANAS) (WATSON; CLARK; TELLEGEN, 1988) for evaluating affect. Additionally, committee members discussed the viability of candidate sets of signs by analyzing sampled collected posts and drawing from their own professional experience, especially as it relates to the student population. Tables 2 and 3 demonstrate, respectively, signs of depression and other correlated signs accompanied by brief descriptions. An additional concern during the development of the label set and annotation guidelines was the need to take into account material that draws from a specifically Brazilian context, such as the book "Religião, Psicopatologia e Saúde Mental" (DALGALARRONDO, 2009), which facilitated the design of annotation guidelines which were both more directly relevant to the annotated texts and more easily comprehended by annotators.

The data was labeled by 4 annotators familiar with the domain of mental health (psychology, psychiatry and occupational therapy students). Annotation was carried out on any given span of text, as opposed to the more typical labeling on a per-post basis. This means that multiple instances of the same sign can (and often do) occur on the same post. Annotators were also free to have overlapping spans, assign multiple signs to the same span and did not have to limit span annotation to sentence boundaries, the only agreed upon requirement was that annotated spans had to carry enough context with regards to their assigned label such that the non-included text was unnecessary (that is, they had to "make sense by themselves"). This permissive annotation setup defers judgment on what constitutes an expression of a given sign onto the specialist annotators on a case by case basis, providing autonomy to carry their analysis in however way they see fit. Constructing a corpus in such a way has drawbacks: there might be inconsistencies in how different annotators label posts (e.g. one annotator might annotate a paragraph containing descriptions of social isolation during the COVID-19 pandemic as a single occurrence of "Helplessness/Social harm/Loneliness", while another might

Table 2 – 18 depression symptoms and their descriptions defined in the Amive project

Symptom	Description
Alteration in sleeping patterns and sleep disorders	Not sleeping adequately (sleeping too little or too much, the resulting tiredness, insomnia)
Alteration in efficiency/functionality	Reduction in efficiency during execution of some task
Sadness/Depressed mood	Negativity, pessimism, melancholia, despondency
Helplessness/ Social harm/ Loneliness	Sentiment of not having support from others, disinterest in being with other people
Suicide/Self-extermination	Suicidal ideation, including passive desire for death, planning and attempted execution of suicide
Worry/Fear/Anxiety	Grouping of similar symptoms. Exaggerated and continuous worry, anticipation, fear and anguish over something that is yet to happen.
Despair	Inability to adopt positive perspectives concerning the present and future. Sentiment that there is no way out.
Feeling of worthlessness/low self-esteem	Aversion to oneself, characterized by sentiment of uselessness, depreciation and diminishment of one's personal worth.
Irritation/aggressiveness	Verbal or physical aggression. Provocation, hostility. Persistent rage, fits of rage.
Physical symptom	Whatever form of physical symptom, such as headaches, sweating, nausea and alterations in heart rate
Feelings of guilt	Feeling of being responsible for some perceived harm against another person or oneself e regret for the decision or behavior associated with it.
Difficulty in decision making	Uncertainty, difficulty in making choices.
Tiredness/Discouragement/Fatigue	Grouping of similar symptoms, characterized by excessive tiredness, lack of morale and motivation.
Attention/memory deficit	Excessive distraction, reduction in capacity for concentration, loss of memory, difficulties in remembering what needs to be done (or was done)
Feelings of emptiness	Sensation of lack of something, inability to feel any emotions
Alteration in weight/eating habits	Report of weight gain or loss, eating in excess or a lot less that usual
Loss/Diminishment of pleasure/libido	Diminished pleasure in life or previously pleasurable activities, diminished sexual interest
Agitation/Restlessness	Agitated and unquiet mind, need of doing something/moving.

Source: Produced by author

separate different expressions of the same sentiment in multiple spans), which potentially hinders a machine learning algorithm's ability to extract patterns from the data. While these are relevant concerns, this work's assumption is that mitigating the complex and often subjective nature of psychological analysis by simplifying or limiting the decision

Table 3 – Additional signs associated with depression and their descriptions

Sign	Description
Risk Factor	Circumstances that make treatment of depression more difficult and may cause or make the condition worse.
Protective Factor	Resources that help a person face their problems
Death/Suicide of third party	Reports of suicide attempts (be they sucessful or not) by another person. Grief, loss of loved ones

Source: Produced by author

making process during annotation runs contrary to making use of annotator expertise and answering the proposed research questions (in order to properly determine if these depression signs can be automatically classified as they naturally occur, should we concern ourselves with how well the data lends itself to machine learning algorithms)? Regardless of the merit of such annotation design decisions, the first 100 posts were collectively annotated by the annotators in the presence of members of the specialist committee, in order to foster a better understanding on what was expected of the labeling process.

It is possible that texts that do not appear to be written by depressed individuals nonetheless contain some of the proposed signs: diagnosis is often defined based on the number and severity of presented symptoms, as is the case of the DSM-5 which (at time of writing) defines major depressive disorder as the presence of 5 or more symptoms, among which "Depressed mood" and/or "Loss of pleasure/interest" must be present. This presents an opportunity with regards to understanding what characterizes depression in text: are there differences between how a depressed person might express a given sign as opposed to a (presumably) non-depressed one? In order to better understand which posts were credibly written by a depressed person, special tags were appended to each post in order to label them as a <Possible Depressive Profile (PDP)>. This provides a post-level binary classification dataset associated with the fine-grained set, while incurring little additional effort on the part of annotators since it is a natural conclusion of the process required for sign annotation, in which an annotator will weight in the intensity, amount and risk posed by the perceived signs of depression.

Each annotator was assigned a subset of posts, with no overlap between each subset, plus a common subset of 34 posts which were annotated by all in order to measure inter annotator agreement. Two measures were chosen for agreement measurement: two-way randomized effects Intraclass Correlation Coefficient (ICC), which was considered an appropriate tool for evaluating rater-based clinical assessment (KOO; LI, 2016) and Krippendorf's nominal alpha (KRIPPENDORFF, 1995), which complements the former measurement (that only takes into account the number of times each of the signs was identified in a post, rather than how much the spans labeled by each annotator are aligned). The sampled posts have a moderate to high ICC of 0.731 (95% confidence interval: 0.691

$< ICC < 0.768$) but only a moderate nominal alpha of 0.424 (95% confidence interval: $0.412 < \alpha < 0.439$). This means that while annotators tended to agree on which signs were present in each post, and even on the number of distinct expressions of each sign in a post, they tended to annotated different (though often overlapping) spans of text as indicative of these signs. This subset of posts was additionally validated by two members of the specialist committee through a curation process, in order to achieve a more rigorous gold-standard. This curated data would compose a test set in addition to the previously mentioned collectively annotated set of 100 posts. In total, 780 posts were annotated.

4.2 Data analysis

Table 4 shows the number of instances for each sign in the train and test sets defined for the experiments in this work. In addition to the specified signs associated with depression, an additional neutral class was included in order to have a better understanding on how much of the collected texts did not contain any depression sign. Given the annotator agreement analysis reported in section 4.1, posts that were not tagged as PDP were considered the most reliable source of spans that did not contain any signs, since the moderate Krippendorf's nominal alpha values suggests that non-annotated spans would be somewhat noisy across the board, using annotator consensus on the lack of depressive profile helps mitigate cases where neutral spans contain some indications of signs of depression.

The collected data has a long-tail label distribution, which is a possible consequence of the fine-grained label set, with some signs being much more frequent than the rest ("Sadness/Depressed mood" and "Helplessness/Social harm/Loneliness", for example) while others are scarce (such as "Sleep Disorder" and "Agitation/Restlessness"). This property is also illustrated in Figure 8, 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). These similar distributions are an additional evidence that texts annotated by single annotators – as opposed to collective annotation or data validated by curation – are reliable, since the distribution was not swayed by the individual annotator's bias. Even though a lot of signs are infrequent, only one is scarce enough as to be ineligible for model training: "Agitation/Restlessness" does not occur in the test set.

With regards to annotated span length, similar distributions can be observed by comparing box plots of the number of characters present in instances from both train and test sets Figure 9, which is expected, since the curation process is primarily one of solving conflicts between existing annotations, although the option to label additional examples

Table 4 – Number of positive instances for each sign

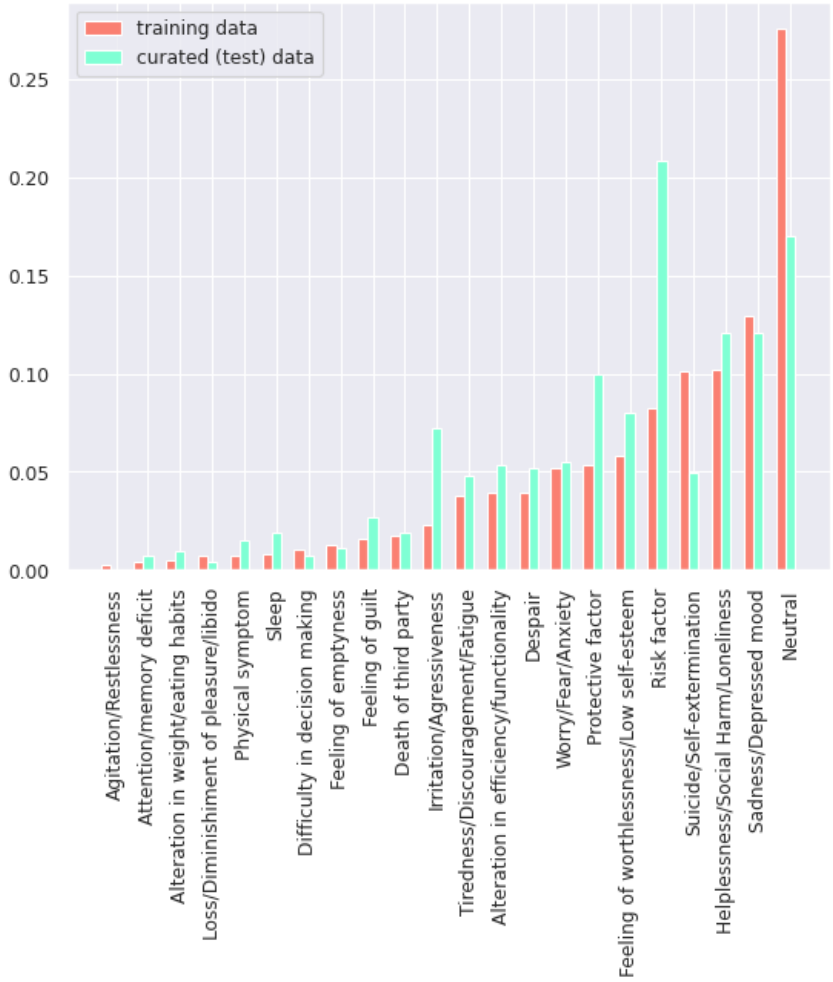
Sign	**Train**	**Test**
Agitation/Restlessness	5	0
Attention/memory deficit	9	4
Alteration in weight/eating habits	10	5
Loss/Diminishment of pleasure/libido	15	2
Physical symptom	16	8
Difficulty in decision making	22	4
Sleep disorder	17	10
Feeling of emptiness	28	6
Death/suicide of third party	37	10
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

Source: Produced by author

was available, if the curators noticed omissions on the annotated data. 162 span instances (comprising ~6% of total annotated data) can be considered outliers given an upper limit of 195 characters (Figure 10). These outliers mostly belong to the neutral category and often consist of whole posts that were not labeled with a single sign. During model training these instances should be broken up into smaller chunks as a pre-processing step, in order to prevent text length from being a relevant factor for determining neutral posts, as in this case longer spans are an artifact of the annotation process.

While reporting on the number of observed instances for each sign is a good starting point for answering RQ1, it does not explain why some signs are more frequent than others. Answering this question is not necessarily the aim of this study, however it is a natural consequence of looking for patterns and insights in the collected data. Possible insights can be gained by adopting a simplified taxonomy that groups signs into a smaller number of categories. A taxonomy inspired by the one proposed by Casani et al. (2021), consisting of behavioral, emotional, somatic and external categories was mapped onto the fine-grained sign set (the mapping was conducted by the author and validated by one member of the specialist committee). By grouping the signs, its possible to observe

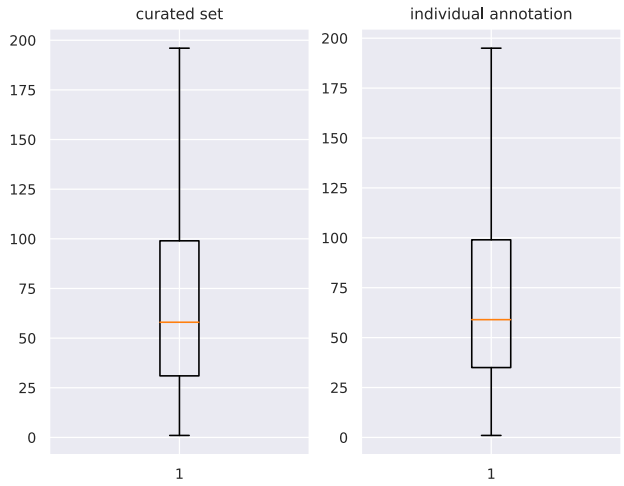
Figure 8 – Label distribution for train and test sets



Source: Produced by author

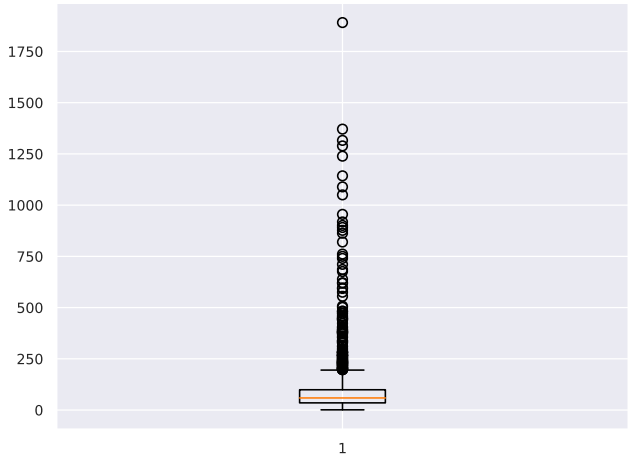
broader patterns: somatic signs are rare in online text, emotional and external signs are frequent, while behavioral sign frequency varies on a case by case basis. One possible explanation is that students prefer to externalize their inner state of mind and discuss everyday events that impacted them through text, rather than state physical discomfort (even as an off-hand remark) while discussing the broader topic of depression. It is worth pointing out that infrequent signs are not necessarily difficult to automatically classify and models that try to identify them are potentially useful for depression detection and treatment, but low frequency can pose a challenge for both data collection and modeling if the trend observe in the collected data applies to online text as a whole, potentially indicating that somatic signs might by better served by alternative data sources such as biometric sensor data. The mapping between the fine grained and simplified taxonomies is detailed in Table 5, which additionally orders the signs into different brackets of frequency based on the number of observed instances.

Figure 9 – Text length boxplots for the training and text datasets



Source: produced by author

Figure 10 – Text length boxplot with combined sets and outliers



Source: produced by author

Another potential avenue of analysis that can yield additional information concerns whether the proposed signs are expressed differently by depressed and healthy users. In this case, the frequency and co-occurrence of these signs can by themselves be sufficient as a diagnosis criteria, as outlined by resources like the DSM (American Psychiatric Association, 2022). Additionally, the intensity and subject matter surrounding these utterances may also contain useful information (for example, while it is perfectly healthy to feel sadness or anger, outbursts of these emotions can not only be more frequent in

Table 5 – Mapping of depression related signs to a simplified taxonomy

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

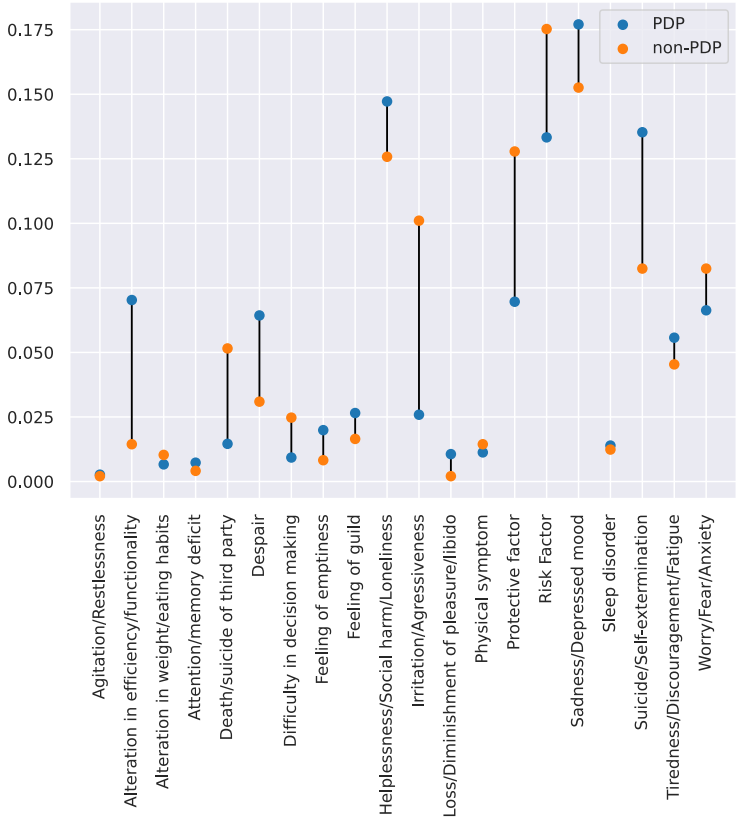
Source: Produced by author

depressed individuals, but also more intense and focused inwards).

A visual comparison of label distributions for signs in the PDP and non-PDP posts (Figure 11) shows that most of the emotional and behavioral signs are more frequent in PDP posts, while external factors are more frequent in non-PDP posts. Somatic signs occur with similar frequency in both post categories. Of special interest is the observation that external factor signs are more common in non-PDP posts, which potentially contradicts the hypothesis that depressed users like talking about everyday events that impacted them, or at least more so than healthy users, leading to additional inquiries on how these two groups talks about this subset of signs. Extracting features from the data can help shed light on why so many non-PDP users are talking about external factors: taking for example the 20 most frequent terms for the sign "Risk Factor" (obtained by TF-IDF, illustrated in Figure 12), it is possible to observe that depressed individuals tend to write about their familial situation ("mother", "father" and "family" are respectively the 1°, 3° and 4° most frequent terms) and about themselves ("I" is the 2° most frequent term); while non-PDP posts discuss the topic of mental health in an impersonal manner - "we (as in, the student body) get sick", "depression", "anxiety" are all among the top 5

most frequent terms - with a particular focus on university day-to-day life ("university", "class" and "professors" are the 9°, 11° and 14° most frequent terms). A similar case of differences between the two post types can be observed in Figure 13.

Figure 11 – Sign distribution between PDP and non-PDP users



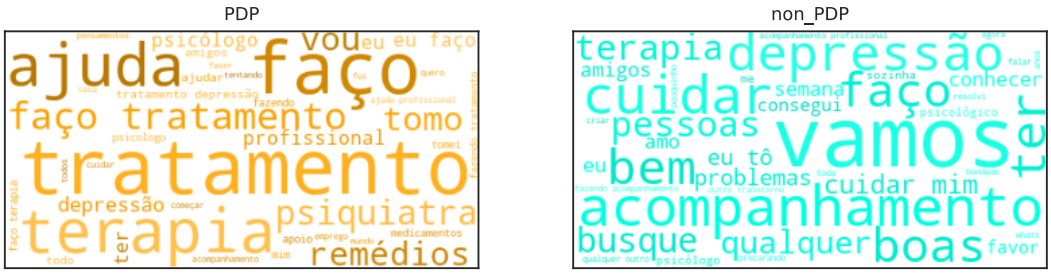
Source: produced by author

In conclusion, this analysis successfully answered RQ1. Depression signs are unevenly represented in online text, with a substantial portion of the proposed signs being seldom expressed in social media, while others are ubiquitous, including some that are often not emphasized when discussing depression classification efforts, such as "Risk Factor", "Protective Factor" and "Alteration in efficiency/functionality". By mapping these signs onto a simplified taxonomy, it is possible to determine that somatic signs are the most infrequent, and that emotional and external signs of depression are frequently expressed in social media. Finally, PDP and non-PDP users tend to express some signs differently, even if they are frequent for both cohorts.

Figure 12 – Word clouds for instances of the Risk Factor class for PDP and non-PDP posts

Source: produced by the author

Figure 13 – Word clouds for instances of the Protective Factor class for PDP and non-PDP posts



Source: produced by author

4.3 Pretrained models and feature engineering

This section describes a collection of pretrained models and other resources selected for training classification models, in order to answer RQ2. Most of these resources were chosen given their reported usefulness for the related tasks of binary depression detection and multi-label/multi-class symptom classification (see Chapter 3 for references), while some are, to the best of our knowledge, novel approaches for depression detection and will be introduced as such.

4.3.1 Pretrained language models

Since the advent of large pretrained language models based on the transformer architecture (VASWANI et al., 2017) such as BERT (DEVLIN et al., 2018), XLNet (YANG et al., 2020) and RoBERTa (LIU et al., 2019), it has become common to use them for downstream tasks either by fine-tuning or by utilizing their contextual embeddings as inputs for other models. The following language models were selected for the task of fine grained depression sign classification in this work: mBERT (DEVLIN et al., 2018), BERTimbau

(SOUZA; NOGUEIRA; LOTUFO, 2020), MentalBERT and MentalRoBERTa (JI et al., 2022). All selected models belong to the BERT family of models, but present trade-offs in terms of the domain they were trained on.

mBERT is a multilingual model that, in contrast to the original BERT, was trained on a corpus composed of the complete Wikipedia dump for 100 different languages, including Portuguese. BERTimbau was trained exclusively on Brazilian Portuguese text, utilizing the BrWaC corpus (FILHO et al., 2018). This corpus was constructed by crawling web pages in accordance with the WaCky methodology (BARONI et al., 2009) and thus has a larger breath of data sources (and presumably, was trained on more varied domains in comparison to wiki data). Lastly, MentalBERT and MentalRoBERTa were trained with the aim of providing contextualized representations that were useful to the specialized domain of mental health, in a similar way to specialized models such as clinicalBERT and biomedicalBERT, with the only difference between MentalBERT and MentalRoBERTa being the underlying architecture. The training data was collected from Reddit communities related to mental health discussion (r/depression, r/SuicideWatch, r/Anxiety, r/offmychest, r/bipolar, r/mentalillness and r/mentalhealth), which is a domain closely aligned with the fine grained sign classification dataset, except for the fact that it was trained exclusively on English data, requiring automatic translation to Portuguese, which can lead to additional noise in the resulting dataset. During planning and execution of experiments detailed in Chapter 5, no specialized model trained in Portuguese was known by the author.

4.3.2 Additional resources

In addition to fine-tuning approaches, resources were selected for the extraction of 9 different feature types to be used in the training of machine learning models. These resources can be divided according to the type of feature extracted: those that inform traditional feature engineering approaches and those used that were used for embedding extraction.

4.3.2.1 Engineered features

Five resources were used to extract features for training traditional machine learning models:

1. **LIWC**: The Brazilian Portuguese version of the Linguistic Inquiry and Word Count (FILHO; PARDO; ALUÍSIO, 2013). Comprises a set of 72 categories, including 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.). The feature extracted

from this resource is the number of occurrences of words pertaining to each of the LIWC categories in a given annotated span.

2. **AnewBR**: a Brazilian Portuguese version of ANEW (KRISTENSEN et al., 2011). Contains dominance (intensity) and arousal (pleasantness) values of words. These values are averaged in order to obtain a feature that captures the general tone of 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; PASSADOR; CASELI, 2021). For the purposes of this study, this lexicon was refined in a process of validation and localization carried out by a member of the Amive's specialist committee and a medical student.[3] It was used to calculate the number of words pertaining to each of the symptoms in the annotated spans.

4. **POS+morph**: A set of course-grained POS (part-of-speech) and morphological categories following the Universal Dependencies (NIVRE et al., 2016) UPOS and FEATS guidelines, as implemented by the spaCy library[4]. 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., 2024), that includes descriptive, cohesion, readability, morphosyntatic and psycholinguistic metrics. This is a seldom considered type of feature for depression classification, given the investigated literature, which was included because it could potentially capture information related to cognitive impairment associated with depression and consequently inform the classification of signs such as "Attention/memory deficit" and "Difficulty in decision making".

4.3.2.2 Embedding features

Four types of word embeddings were investigated in this work:

1. **Static embeddings**: Word2Vec, FastText, GloVe and LexVec models trained from social media posts in Portuguese and in the mental health domain, sourced from the dataset introduced by Paraboni, Funabashi e Santos (2020). A pre-trained version of Word2Vec[5] was also used in order to evaluate if there's benefit in utilizing in-domain

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

[4] https://spacy.io/

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

data as opposed to general resources trained on larger corpora. The average of the embeddings of every token in each text instance was extracted from these resources in order to obtain a span embedding.

2. **TF-IDF**: Term frequency-inverted document frequency values. The term frequencies were fitted on the training set. This feature aims to capture term frequency information while penalizing terms that are frequent in all of the evaluated documents (and thus might not be informative about a given text's content).

3. **Contextualized embeddings**: Internal representations of the special token [CLS] produced by a pre-trained model, BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020). This feature has the same overall goals of the static embedding features, adequately representing information about a text in a dense manner. These embeddings are treated differently because they represent transformer-based models, which are able to generate different embeddings for the same token depending on surrounding context.

4. **AffectiveSpace embeddings**: Portuguese version of AffectiveSpace (CAMBRIA et al., 2015a), a graph embedding trained on the SenticNet knowledge graph (CAMBRIA et al., 2015b), which aims to model common sense knowledge as it pertains to emotions. This is another rarely explored approach in the domain of mental health, though it was used in a similar domain before (GHOSH; EKBAL; BHATTACHARYYA, 2022). As with the previous static embedding resources, the token embeddings in a given annotated span were averaged.

Chapter 5

Single-task Experiments

This chapter details the setup and results of computational experiments designed to answer RQ2 by measuring model performance of machine learning techniques commonly seen in the literature and thus identifying which of the fine grained signs can be accurately classified by the resulting models. An additional set of experiments was conducted in order to better determine the role that data imbalance and small data volume play in this task, and the potential for regularization techniques to address these issues.

5.1 Baseline experiments

The initial set of experiments aimed to identify which of the proposed signs could be discerned automatically from text using strategies that have previously proven useful in the literature (exemplified in chapter 3), in order to better understand which type of content these models can identify, and which they tend to miss. These experiments can be understood as a comprehensive set of baselines, though the intent is not to compare these results to a novel network architecture or algorithm, but between the performance observed in traditionally reported task setups (binary classification, PHQ-9 symptom classification) to our proposed fine-grained classification task.

As mentioned in section 4.3, both pretrained models and resources appropriate for feature extraction were selected for evaluation, which were respectively used for the training of deep learning and traditional Machine Learning (ML) models. Five different traditional ML algorithms were chosen: SVMs, logistic regression and XGBoost, because they have proven adequate for the tasks of binary depression and symptom classification; Naive Bayes and Random Forest, which were also frequently evaluated in the cited work (PARABONI; FUNABASHI; SANTOS, 2020; CHOUDHURY et al., 2013; JI et al., 2018;

LIU et al., 2022b; CASANI et al., 2021). Training for deep learning models consisted of the standard fine-tuning setup, in which all pre-trained model parameters are adjusted to the downstream task, with the addition of a linear classification head in order to properly model the classification task. For this set of experiments, separate classification models were trained for each sign in a one-vs-all setup.

Section 4.3 also introduces two groupings of feature sets that can be used as input for traditional ML training: "engineered" features, which are generally sparse and encode information concerning a particular aspect of a text (e.g. sentiment, syntax, particular topics or concepts); and embedding features, which are dense representations that are expected to carry all sorts of information about a given piece of text. In order to better understand what type of feature is most useful for the sign classification task, traditional ML models were trained on the whole engineered feature set, akin to (as the grouping name implies) feature engineering approaches (LIWC, NilcMetrix, ANEW, POS+morph), but were trained on each embedding type separately (since all embedding strategies are designed to achieve the same overall goal of packing as much meaningful information as possible in a dense representation). NilcMetrix is an outlier in the engineered feature set: it is an untested approach in the mental health domain, in comparison to the often reported success of employment of other features, and it is composed of comparatively complex metrics (which were often designed with larger documents in mind). These factors, combined with the fact that NilcMetrix features comprise the largest share of all engineered features (roughly a third of all data at 72 unique features), it was decided that ablating NilcMetrix features was a necessary step in order to adequately measure its impact on performance.

Traditional ML algorithms hyper-parameters were fitted through a 60-iteration random search with 5-fold cross validation (stratified, given the imbalance in sign frequency). In the case of deep learning models, the only adjusted hyper-parameters were the choice of pre-trained model and the loss function, which could be either cross entropy loss or one of its adaptations, focal loss (LIN et al., 2017). Focal loss was originally designed for object detection, and was included in order to potentially help the models cope with the label imbalance, specially considering the large number of neutral instances present in the dataset. It penalizes easy examples (those that the model gets right with high confidence) with a factor γ, and also emphasizes samples of the positive class by a factor α. Given p is the probabilities outputted by a model:

$$p_t = \begin{cases} p & y = 1 \\ 1 - p & y = 0 \end{cases} \tag{5}$$

and

$$\alpha_t = \begin{cases} \alpha & y = 1 \\ 1 - \alpha & y = 0 \end{cases} \tag{6}$$

we have:

$$Focal = -log(p_t) * \alpha_t * (1 - p_t)^{\gamma} \quad (7)$$

α can be calculated as the inverse frequency of the positive class or as a hyper-parameter to be optimized alongside γ. Given time constraints, α and γ values were set to 0.3 and 0.8 respectively for this initial run of experiments. Remaining fine-tuning parameters were fixed as per the original BERT paper: AdamW (LOSHCHILOV; HUTTER, 2019) optimizer with a learning rate of 5e-5, early stopping after 4 epochs, constant learning schedule with linear warmup for the first 10% of training steps, 10% dropout before classification head and random sampling during training.

The primary metric chosen for evaluation was the AVPs, which is the average precision across different classification thresholds, varying from 0 to 100% recall. This decision was made considering a couple of factors: the performance metric should be robust to label imbalance, akin to f1-score, but should ideally also not depend on a decision threshold, being able to capture the discriminative ability of a model without further parametrization, akin to AUC. AVPs meets both criteria (SAITO; REHMSMEIER, 2015), and a threshold of what could be considered sufficiently good performance was empirically set to 70%.

Table 6 shows the best results for each sign on the curated test set. These were obtained on a single training run on the whole training data after hyper-parameter fitting. It is possible to observe that some signs can be classified with relatively good performance given the 70% AVPs criteria, but most models cannot accurately discriminate most categories. One possible common sense explanation for poor performance is the small number of positive training instances for some categories, since models might be unable to generalize to new examples of a depression sign if it has seen only a scant few during training. This phenomenon can be observed in Figure 14, however this visualization also showcases signs that contradict this simple explanation: signs that occur infrequently can still be predicted more accurately than other with an order of magnitude more data. For example, “Sleep disorder” reaches 90% AVPs with only 17 training instances of the positive class, while the best model for “Risk Factor” only reaches 46% with 177 instances (being the second most frequent sign, behind only “Sadness/Depressed mood”). This somewhat unexpected result might point to new understandings of what constitutes an “easy” to classify depression sign, and in particular which common signs (such as the aforementioned “Risk Factor”) might not be adequately identified by current modeling approaches.

Taking into account the simplified taxonomy proposed in section 4, it is possible to find additional evidence that sign frequency does not necessarily mean ease of classification: somatic signs are in general more easily discernible than other groupings, reaching a macro-averaged AVPs of 76.14% (as shown in Table 7). One possible explanation for these results is that model performance is primarily dependent on the inherent complexity of each sign classification task, and as such even infrequent signs can be predicted if they

Table 6 – Average precision scores for each sign. Models which achieved more than 70% average precision score are in bold.

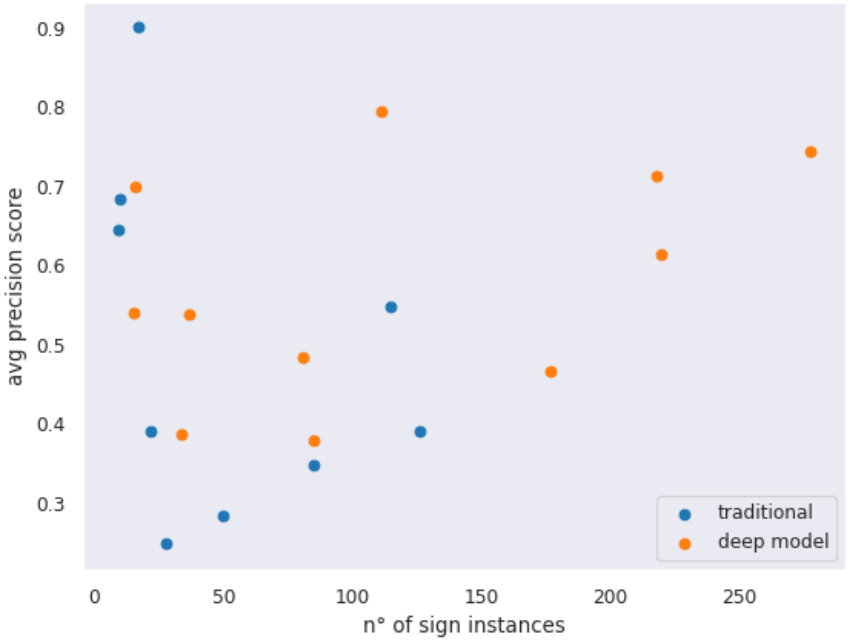
Sign	Model	AVP
Feeling of emptiness	XGBoost (word2vec)	24.8%
Irritation/Aggressiveness	SVC (word2vec)	28.4%
Alteration in efficiency/functionality	SVC (word2vec)	34.8%
Despair	BERTimbau	37.9%
Feeling of guilt	BERTimbau	38.7%
Difficulty in decision making	RandomForest (engineered)	39.1%
Feeling of worthlessness/Low self-esteem	XGBoost (word2vec)	39.1%
Risk factor	BERTimbau	46.5%
Tiredness/Discouragement/Fatigue	BERTimbau	48.3%
Death/suicide of third party	BERTimbau	53.8%
Loss/Diminishment of pleasure/libido	BERTimbau	54.0%
Protective Factor	SVC (word2vec)	54.8%
Helplessness/Social harm/Loneliness	MentalRoberta	61.4%
Attention/memory deficit	XGBoost (word2vec)	64.5%
Alteration in weight/eating habits	SVC (word2vec)	68.3%
Physical symptom	BERTimbau	**70.0**%
Suicide/Self-extermination	MentalRoBERTa	**71.2**%
Sadness/Depressed mood	BERTimbau	**74.5**%
Worry/Fear/Anxiety	MentalBERT	**79.4**%
Sleep disorder	LogisticRegression (engineered)	**90.0**%

Source: Produced by author

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 manner (e.g. "I didn't sleep well last night"), "Despair", in turn, can be expressed in so many ways as illustrated in Table 8. This complexity can also potentially mislead models: 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 have a more negative affect (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 talk about the protective factor indirectly (e.g. "I couldn't have done it by myself", implying a healthy support network). These examples are true cases from our corpus and were gathered during an evaluation of model errors.

A limitation that should be noted is that this analysis of model error cases was not comprehensive and it is difficult to quantitatively measure this type of semantic variety: approaches to measure lexical diversity such as type/token ratio and measuring semantic similarity by comparing embeddings would, I argue, not adequately capture this variety, since the focus is on figurative language and what is left unsaid, not necessarily which

Figure 14 – Visualization of model performance per number of training instances.



Source: Produced by author

Table 7 – Macro average precision scores for each type of sign

Sign	**AVPs**
Somatic	76.14% ± 12.08%
Behavioral	52.01% ± 14.18%
External	51.73% ± 4.50%
Emotional	48.07% ± 20.90%

Source: Produced by author

particular set of words were used (for example, "I haven't done it yet, because of my daughter" is more similar to "I have yet to kill myself" than "My head hurts" is to "I feel like my heart is about to explode"). Despite these limitations, it is worth pointing out that similar observations have been reported, Yadav et al. (2020) affirm that a common error case for depression symptom classification models is the use of figurative language, which is commonly employed since it helps people express feelings that would be otherwise hard to communicate.

The results also demonstrate that despite a general lack of data (fewer than 300 positive instances seen during training for every sign), limited hyper-parameter tuning (in comparison to the traditional ML models) and minimal adjustments to the training regi-

Table 8 – Examples for the "Despair" sign instances. Note the use of figurative language and the varied ways users can broach the subject.

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...

Source: Produced by author

men that account for this low data volume, deep models achieved better performance than other ones in 11 out of 20 cases (Table 6), even in some of the least frequent categories ("Physical symptom", "Loss/Diminishment of pleasure/libido"). Accounting for label imbalances proved to be relevant, with all of the best models found through hyper-parameter search utilizing focal loss.

With regard to traditional models, engineered features achieved best performance in 9 of 20 signs when compared with their embedding counterparts, and 2 out of 20 when compared to all models. While feature engineering by itself is 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 across the board. Traditional models benefited from accounting for label imbalance, with most models utilizing some sort of class weight balancing strategy.

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 out of 20 signs despise noise introduced by automatic translation, points to the potential benefit of a language model for Portuguese specific to the mental health domain.

Finally, since accounting for label imbalance proved to be relevant, an additional set of experiments utilizing common sampling (random sampling, near miss sampling and sampling of centroid clusters) and regularization techniques (SMOTE) was conducted in order to balance out the data. These techniques all led to a drop in performance for all classes. Table 9 details these results for traditional ML. Note that of these selected techniques, only random sampling was applicable to deep learning approaches, thus a different approach will be explored in section 5.2.

Table 9 – Best AVPs for each depression sign using the selected sampling and regularization techniques

AVPs	**Technique**	**Sign**	**Difference**
0.813	ClusterCentroids	Sleep disorder	-0.087247
0.456	ClusterCentroids	Difficulty in decision making	0.065
0.342	SMOTE	Death/Suicide of third party	-0.188
0.038	RandomUnderSampler	Loss/Diminishment of pleasure/libido	-0.114
0.749	RandomUnderSampler	Worry/Fear/Anxiety	-0.015
0.670	NearMiss	Sadness/Depressed mood	-0.061
0.254	RandomOverSampler	Alteration in weight/eating habits	-0.429
0.316	ClusterCentroids	Alteration in efficiency/functionality	-0.031
0.291	RandomOverSampler	Tiredness/Discouragement/Fatigue	-0.018
0.457	RandomUnderSampler	Helplessness/Social harm/Loneliness	-0.079
0.260	SMOTE	Despair	-0.044
0.425	SMOTE	Feeling of worthlessness/low self-esteem	0.033
0.323	RandomOverSampler	Attention/memory deficit	-0.321
0.383	RandomOverSampler	Risk Factor	0.007
0.577	RandomOverSampler	Protective Factor	0.029
0.188	ClusterCentroids	Irritation/aggressiveness	-0.096
0.251	SMOTE	Feelings of guilt	-0.055
0.058	SMOTE	Feelings of emptiness	-0.190
0.311	SMOTE	Physical symptom	-0.065
0.660	RandomUnderSampler	Suicide/Self-extermination	0.065

Source: Produced by author

Additionally, performance was evaluated in a multi-label training setup (as opposed to the one-versus-all approach). For traditional ML approaches, all multi-label models performed worse than their one-vs-all counterparts (in terms of macro-averaged AVPs). Surprisingly, model performance was significantly improved by this approach in deep learning models (see Table 10), despite multi-label classification being at first glance a more challenging task (intuitively, discerning between the presence or absence of a category should be easier than picking the appropriate category from a set of 21 signs). Two candidate explanations for this behavior are listed below:

- ❑ Label distribution argument: The multi-label setup, despite still suffering from imbalance problems (see the long-tailed distribution in Chapter 4), has a less pronounced version of this problem, particularly concerning the more frequent cate-

gories, which helps with discerning useful patterns from data.

- Introduction of additional information: Having more label information to work with, the model has an easier time discerning between categories. For example, spans that portray negative sentiment but are not classified under the "Sadness/Depressed Mood" class can be accurately classified as pertaining to a different sign (say for example "Dimishment/Loss of pleasure/libido"). In other words, instead of potentially misleading the model, the larger number of contrasting sign examples help the models more accurately represent what characterizes any particular sign.

Table 10 – Average precision scores for each sign in the multi-label model and difference over one-vs-all approach.

Sign	**AVP**	**Difference**
Feeling of emptiness	0.593	+0.493
Irritation/Aggressiveness	0.429	+0.199
Alteration in efficiency/functionality	0.409	+0.075
Despair	0.436	+0.056
Feeling of guilt	0.586	+0.199
Difficulty in decision making	0.436	+0.328
Feeling of worthlessness/Low self-esteem	0.53	+0.164
Risk factor	0.625	+0.16
Tiredness/Discouragement/Fatigue	0.569	+0.086
Death/suicide of third party	0.583	+0.044
Loss/Diminishment of pleasure/libido	0.077	-0.463
Protective Factor	0.722	+0.16
Helplessness/Social harm/Loneliness	0.692	+0.078
Attention/memory deficit	0.316	+0.021
Alteration in weight/eating habits	0.546	-0.01
Physical symptom	0.489	-0.211
Suicide/Self-extermination	0.771	+0.059
Sadness/Depressed mood	0.873	+0.129
Worry/Fear/Anxiety	0.909	+0.114
Sleep disorder	0.877	+0.1

Source: Produced by author

5.2 Alternative strategies for deep model training

Since deep learning approaches have demonstrated their capabilities despite the data scarce scenario, a set of experiments was designed to evaluate potential avenues for model improvement. These approaches fall into two broader categories: reducing the number of trainable model parameters in order to reduce the risk of overfitting and generating new

examples that are semantically diverse. All hyper parameters were fixed according to section 5.1, the only pretrained model evaluated was BERTimbau, since it was the approach that achieved best performance for a plurality of signs compared to other techniques, and only utilizing binary cross entropy as the loss function. The macro-averaged AVPs across 10 fold cross validation was used. These techniques included:

- ❏ Freezing layers: In this experiment the embedding layer + the first n encoding layers of a pre-trained model were frozen. This can potentially help mitigate overfitting and prevent catastrophic forgetting in which a pre-trained model loses the capabilities it acquired during pre-training because of overly aggressive parameter updating. Three different configurations were tested: freezing only the embedding layer, embedding + first 3 layers and embeddings + first 6 layers.

- ❏ Utilizing contextualized embeddings as features of a lower capacity model: Despite fine-tuning pre-trained transformer models being a common approach, it is possible that training a lower capacity model using pre-trained embeddings as input might be less prone to overfitting. In order to test this hypothesis, BERTimbau embeddings were utilized as features for the training of a relatively simple BiLSTM architecture, with two layers of hidden layers. The final state of the special token [CLS] produced by this BiLSTM was then fed to a final linear classification layer.

- ❏ Pruning: Pruning techniques consist in removing a large number of model parameters according to certain criteria. These approaches are based on the notion of "lottery tickets" in overparametrized networks (FRANKLE; CARBIN, 2019), subnetworks with performance equal to the original network. This phenomenon has been observed in pre-trained BERT models (CHEN et al., 2020), and in such cases models are capable of retaining transfer learning capabilities despite the dropped weights. It is another candidate approach to reduce model capacity and thus prevent overfitting.

- ❏ Distilled models: Model distillation techniques are designed to train lower capacity models based on the outputs of high capacity models while preserving performance. A publicly available distilled version of BERTimbau[1] with half the number of encoder layers was utilized.

- ❏ SMART: SMoothness- inducing Adversarial Regularization and BRegman pRoximal poinT opTimization (SMART) (JIANG et al., 2020) is a regularization techniques specifically designed for the fine-tuning of language models. It is composed of an adversarial regularization component, which induces model output to stay consistent across similar input, and a component that penalizes large updates in the parameter space.

[1] https://huggingface.co/adalbertojunior/distilbert-portuguese-cased

- ❏ Automatic annotation based on embedding similarity: Akin to the strategy used by Lee et al. (2021) (see Chapter 3), a sentence transformer (REIMERS; GUREVYCH, 2019) model (in this case, `paraphrase-xlm-r-multilingual-v1`[2]) was used to find novel text similar to labeled data in an unlabeled corpus and automatically annotating these examples. The publicly available code for the authors micromodel approach was adapted for the sign classification task, including new examples from the unlabeled corpus collected by Paraboni, Funabashi e Santos (2020) that met the minimum cosine similarity threshold criteria of 0.85. This process was iteratively repeated until no new examples were included.

- ❏ Back translation: Consists in the automatic translation of a dataset to other languages, and then back to the original language, in order to obtain examples that vary in syntax while preserving semantic content (and thus, are still aligned with their original labels). The model utilized for automatic translation was $\texttt{m2m}_1\texttt{00}(FANetal., 2021), utilizingt$
$english, french, spanish, italian, mandarin, germanandrussian.$

- ❏ Mixup: MixUp is a data augmentation technique that aims to create new training examples by directly manipulating the latent representations. New instances are the result of a linear combination of latent representations from two original instances. Given α as a value sampled from a given beta distribution with values ranging from 0 to 1 (for this experiment a uniform distribution was chosen):

$$\hat{X} = X_1 * \alpha + X_2 * (1 - \alpha) \tag{8}$$

$$\hat{Y} = Y_1 * \alpha + Y_2 * (1 - \alpha) \tag{9}$$

Table 11 shows the performance of the resulting models in comparison to a 10-fold cross validation training run of the baseline multi-label model reported in section 5.1, since it was the model with achieved best overall performance in the previous experiments. None of the evaluated techniques produced a statistically significant gain in performance[3]. Techniques that heavily reduced model capacity (aggressive pruning, distilled model and LSTM model) showed a substantial loss in overall performance. It is worth pointing out that models resulting from less aggressive weight freezing/dropping strategies do not suffer from significant performance loss, suggesting that the evaluated pre-trained model is indeed over-parametrized for the depression sign classification task, and as a result such techniques or similar ones like model adapters (HOULSBY et al., 2019) might provide other benefits like faster training times and reduced resource usage.

Of particular interest is the fact that the label propagation technique, back translation and MixUp data augmentation approach did not yield any significant improvement in performance, and in fact slightly tended to worsen it (though not to a statistically

[2] https://huggingface.co/sentence-transformers/paraphrase-xlm-r-multilingual-v1
[3] Utilizing the Wilcoxon paired t-test.

Table 11 – Macro-averaged AVPs for each model and comparison with baseline

Model	**Performance**	**Gain/Loss**	**p-value**
Distillation	48.1%	-4.7%	0.005
Freezing(embeddings)	53.5%	0.7%	0.92
Freezing(3 encoders)	52.5%	-0.2%	0.59
Freezing(6 encoders)	53.9%	1.1%	0.375
SMART	54.4%	1.65%	0.137
LSTM+embeddings	32.2%	-20.5%	0.002
Pruning(20%)	51.6%	-1.1%	0.43
Pruning(40%)	50.6%	-2.1%	0.105
Pruning(70%)	44.6%	-8.2%	0.001
Back translation	53.7%	1%	0.32
Label Expansion	51.4%	-1.3%	0.492
MixUp	52.4%	-0.3%	0.767

Source: Produced by author

substantial degree). Two possible explanations for the inability of the generated data to improve model performance is that: I) the generated examples are noisy, making it more difficult for the model to learn meaningful patterns from the data, and II) the generated data falls into the same distribution as the labeled data it was generated from, which means it could not help improve a model's generalization capabilities. While it is not possible to properly evaluate these hypothesis for the MixUp approach (given the difficulty in interpreting latent space representations), observing samples of the silver dataset obtained by the micromodel automatic labeling strategy seems to corroborate hypothesis II.

These results are evidence of a major barrier in the pursuit of fine-grained depression sign classifiers, since the necessity for expert labelers limits the collection of diverse and high quality data, and thus also limits the capacity to train accurate deep learning models. An ideal candidate solution to this problem would have to be able to provide this diverse and trustworthy information without additional effort on the part of annotators. Two candidate solutions that might be able to meet these criteria are multi-task learning and few-shot learning by Large Language Model (LLM) prompting: multi-task learning can leverage data from publicly available datasets, preferably from those that model a closely related task to depression sign classification, providing better generalization capabilities, and LLMs have proven capable of generalizing well with small amounts of data, being mainly dependent on high quality examples and prompting.

Preliminary experiments utilizing the open source LLMs llama2-7b, llama2-13b and

platypus7b were also carried out. 8 bit quantization and 4 bit quantization were utilized for the 7 billion and 13 billion parameter models respectively, given computational resource limitations. This study had a rich source for in-context learning: the annotation guidelines, which were designed to be clear, unambiguous, concise and provide informative examples to annotators. Thus, the evaluated LLMs had access to the same guidance as human annotators, though no counter-examples were provided, as they were not present in the guidelines. These models struggled with most categories, generally achieving low precision (high type I error, which means false positives compose a large amount of overall predictions of the positive class). Two categories ("Physical symptom", "Attention deficit or memory impairment") could not be accurately classified at all, with reported precision and recall of 0. Consequently, this approach was reserved for future work, as there was a lack of resources required for the use of larger capacity models.

In conclusion, these experiments answer RQ2: "Sleep Disorder", "Worry/Fear/Anxiety", "Sadness/Depressed mood", "Suicide/Self-extermination", "Protective Factor" and "Physical Symptom" can be accurately classified by the evaluated models. By grouping these signs in the aforementioned simplified taxonomy, it was observed that somatic signs were the most easily classified. Several signs still present inadequate performance despite their frequency and a breath of different approaches evaluated. It is possible that multi-task learning can help these models achieve better generalization capabilities and thus, better performance for some os these signs. In Chapter 2, experiments carried out on multitask learning are presented.

5.3 Auxiliary tasks

This section details the datasets associated with the auxiliary tasks selected for multi-task learning. Since there is no consensus on what constitutes similarity between tasks, tasks were selected based on data source (social media, preferably long form), task structure (post or span-level, preferably classification) and a loose sense of perceived usefulness ("is it likely that the same competency is needed for both this task and the primary task?") A notable omission is the SetembroBR dataset (SANTOS; OLIVEIRA; PARABONI, 2023), mentioned in Chapter 3, which would fit most of these criteria, however the user-level nature of the classification task proved difficult to model without adapting the multi-task architecture, which would confound comparisons between tasks.

5.3.1 GoEmotions

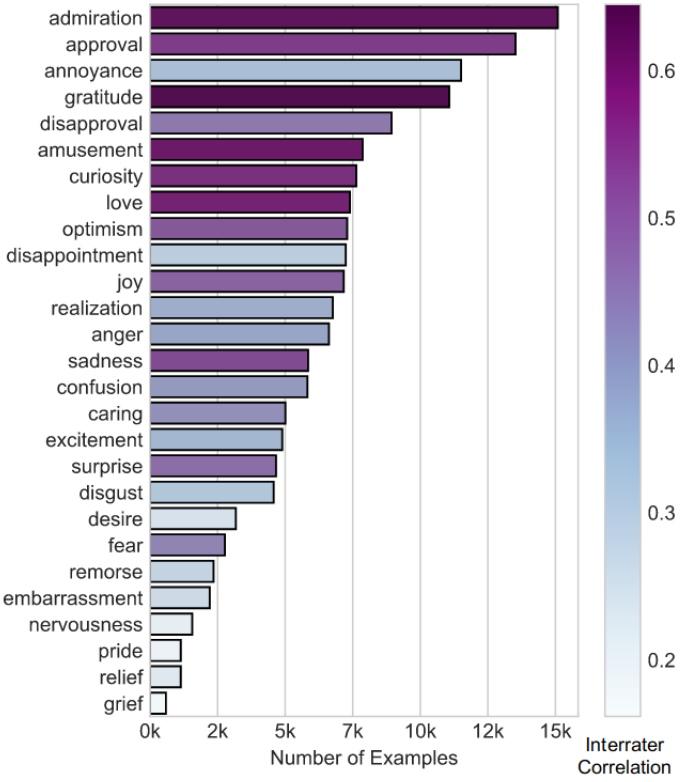
GoEmotions (DEMSZKY et al., 2020) is a publicly available dataset for fine-grained emotion classification. In a similar sense to the primary task dataset, GoEmotions was designed to fill a knowledge and resource gap in emotion detection, since publicly available resources were limited to coarse classification schemes. It is composed of 54263 comments extracted from 482 Reddit communities. Their fine-grained label set consists of 27 categories: neutral, admiration, approval, annoyance, gratitude, disapproval, amusement, curiosity, love, optimism, disappointment, joy, realization, anger, sadness, confusion, caring, excitement, surprise, disgust, desire, fear, remorse, embarrassment, pride, relief and grief.

This dataset mirrors the fine-grained nature of the primary task dataset, has a similar data source (social media, short spans extracted from long form text, semi-anonymous), models a related task - it is possible to draw a direct parallel between some emotions and the proposed depression signs, mostly obviously in those that are grouped under the "Emotional" category in the simplified taxonomy. The dataset also has a long tailed label distribution, as can be observed in Figure 15, which additionally presents the difficulties associated with the annotation efforts, with some categories suffering from poor inter rater correlation.

The GoEmotions authors also propose mappings between their fine-grained taxonomy and 2 more course grained taxonomies: Ekman taxonomy, comprised of 7 categories (anger, disgust, fear, joy, sadness, surprise and neutral) and Sentiment (neutral, positive, negative and ambiguous). The mappings are detailed in Table 12. Each taxonomy can be treated as a task, totaling 3 auxiliary tasks.

It should be noted that there are aspects that make this task dissimilar to the primary task. These posts were collected from a large variety of communities and were not restricted to a given subject matter, as opposed to the primary task dataset which was constructed from social media posts from university students, after keyword filter-

Figure 15 – Number os instances and inter-rater agreement for each emotion in the GoEmotions dataset



(DEMSZKY et al., 2020)

ing targeting depression related posts. Besides, comments were pre-processed in order to not surpass 30 tokens and were all in English, necessitating automatic translation which might result in noise. The process of automatic translation was replicated from Hammes e Freitas (2021), who utilized the itranslate library and reported similar performance to the original study, suggesting that the translation process did not negatively impact performance.

5.3.2 Occupational Therapy annotated data

The dataset for this task was built from a subset of data from the primary task dataset, at the request of one of the members of the annotation team for their own research efforts. It consists of text spans annotated in the same manner as the primary task dataset, with a label set consisting of 7 different types of occupation according to the American Occupational Therapy Association, which defines an occupation as “activities that people do every day to give their live meaning and purpose”. The categories are: activities of daily living, health management, rest and sleep, education, work, play, leisure, and social

Table 12 – Mapping of GoEmotions categories

Emotion	Ekman	Sentiment
Amusement	Joy	Positive
Excitement	Joy	Positive
Joy	Joy	Positive
Love	Joy	Positive
Desire	Joy	Positive
Optimism	Joy	Positive
Pride	Joy	Positive
Admiration	Joy	Positive
Graditude	Joy	Positive
Relief	Joy	Positive
Approval	Joy	Positive
Realization	Surprise	Ambiguous
Surprise	Surprise	Ambiguous
Curiosity	Surprise	Ambiguous
Confusion	Surprise	Ambiguous
Fear	Fear	Negative
Nervousness	Fear	Negative
Remorse	Sadness	Negative
Shame	Sadness	Negative
Disappointment	Sadness	Negative
Grief	Sadness	Negative
Disgust	Disgust	Negative
Anger	Anger	Negative
Annoyance	Anger	Negative
Disapproval	Anger	Negative
Sadness	Sadness	Negative

Source: Produced by author

participation. The annotation effort resulted in a small dataset of 363 annotated spans. Table 13 reports the number of instances for each occupation.

Since this data was annotated at a later date than the primary task set and was not immediately considered as a candidate auxiliary task (at the time of annotation, multi-task learning was not considered as a potential avenue for model improvement), additional care had to be taken in order to ensure that data leakage did not occur: both direct span

Table 13 – Number of instances for each occupation category

Occupation	Instances
Leisure	9
Health management	13
Work	18
Activities of daily living	19
Rest and sleep	22
Education	129
Social participation	153

Source: Produced by author

matches and fuzzy matches[4] were detected in order to appropriately assign instances to the test and train sets. The majority of annotated spans were wholly original despite the presence of "Alteration in efficiency/functionality" being a sign category, with ~ 34% of training instances and ~ 13% of test instances from the primary task matching the auxiliary task's corpus. It is worth pointing out that there is a subtle difference between the aforementioned sign category and the auxiliary label set: text concerning a particular occupation does not necessarily reports on alteration in an person's behavior regarding that occupation. Regardless, parallels can be drawn between this auxiliary task and some primary task signs ("Alteration in efficiency/functionality", "Helplessness/Social harm/Loneliness", "Loss/Diminishment of pleasure/libido").

5.3.3 Automatically labeled tasks extracted from the primary task dataset

Some tasks can be automatically extracted from the primary task dataset. These include simplified taxonomy data (which can be obtained by mapping each fine-grained sign to its corresponding coarse-grained category, as detailed in Chapter 4) and binary classification of PDP, which closely resembles binary classification tasks and can be modeled as span-level and post-level.

5.3.4 Classification of r/Desabafos tags

The Reddit community r/Desabafos is an online space where users can vent about whatever topic in semi-anonymous fashion, in a manner similar to the primary task's "university secrets" pages. The community has a collection of tags for the purpose of

[4] fuzzy matches were computed using `thefuzz` library, https://github.com/seatgeek/thefuzz

labeling posts, useful for filtering and organizing posts into specific topics. These topics include “Good news”, “Depression”, “Daily life”, “Bad News” and “Relationships” (a subset of categories which are among the most frequent and potentially useful for the task of depression sign classification) and can be provided by the author or added by community moderators. These tags can therefore be used as a way to gather high quality annotated data automatically, since these tags were produced by human annotators with a vested interest in accurately classifying posts. Posts were collected by means of the third party Reddit API Pushift 2023’s dump of Reddit data, Table 14 showcases the categories alongside an example and the number of instances for each of them. Since post titles often contained necessary context for the comprehension of the post text body, these were prepended to each instance. It should be noted for the sake of result interpretation that this is a post-level classification task, which results in the average instance for this auxiliary task being much longer than those of the primary task (which are text spans).

Table 14 – Number of instances for each r/desabafos tag

Tag	Instances
Good news	7944
Depression	9762
Bad News	10267
Relationship	31217
Daily life	31466

Source: Produced by author

5.4 Multi-task Experiments

This section details the setup and results of computational experiments designed to answer RQ3 and RQ4 by measuring both model performance of multi-task models and a selection of task characteristics that, according to the literature, are indicative of positive transfer (see the studies cited in section 3.1).

5.4.1 Auxiliary task baselines

While improvement on any of the auxiliary tasks performance is not the motivation for utilizing multi-task learning in this particular study, understanding the results a target neural architecture can achieve when modeling a candidate task can provide relevant context about the model's capabilities and potentially be a useful predictor of positive transfer, since works like Alonso e Plank (2017) provide evidence that "simpler" tasks are more likely to benefit a given primary task. Therefore, baseline experiments were ran for each auxiliary task.

Baseline models were trained with the same setup described in Section 5, however hyper parameter search was conducted using the `optuna`[5] library with a small number of trials (30), instead of executing a more through random search, since doing the latter proved to be excessively slow for the bigger datasets. Just as with the primary task results, the AVPs of each label were reported, alongside a macro-average. It must be acknowledged that AVPs is not necessarily the best fit for some of these tasks, as they do not share the same label distribution challenges of the primary task. However, there is no inherent disadvantage in utilizing AVPs as the optimization target of hyper-parameter search, nor as a general descriptor of model performance, since it is robust to label imbalances and takes into account both precision and recall (just like F1-score). Thus, it was kept as the main performance evaluation criteria. In the case of GoEmotions, additional performance information was added, for the sake of comparison with the results of previous studies. What follows are tables reporting the performance for each task, followed by brief analyses of model performance, except for the depression sign classification task under a simplified taxonomy, as these results were already reported and discussed in Chapter 5.

Table 15 reports results for the occupation classification task. It can be observed that most occupations are identified by the model with sufficient precision despite the scarce data (given the previously established threshold of 70% AVPs) save for the "Leisure" occupation, the least frequent category in the dataset. This overall good performance, combined with the fact that its associated dataset was produced from the same corpus and under a similar annotation process are indicative of a good candidate task, as it is quite "similar" to the primary task in these regards. However, it should be considered that occupation category information is only directly relevant to a small number of

[5] <https://optuna.org/>

signs ("Alteration in efficiency/functionality", "Protective Factor", "Helplessness/Social harm/Loneliness").

Table 15 – Average precision scores for the occupation classification task

Occupation	AVPs
Activities of daily living	84%
Health management	79.6%
Rest and Sleep	82.4%
Work	81.7%
Education	91.9%
Social Participation	96.7%
Leisure	37.1%

Source: Produced by author

Tables 16, 17 and 18 report results for the emotion classification tasks using the fine-grained, Ekman and sentiment taxonomies respectively. The results for the fine-grained taxonomy are in line with the expected values as reported by (DEMSZKY et al., 2020) and (HAMMES; FREITAS, 2021): the model struggled to distinguish between the fine-grained signs, achieving a macro-averaged AVPs of 41.7% and a macro-average F1 score of 40.3%. In the case of the simplified taxonomies, however, the model fares worse than in the original article, achieving a macro-averaged F1 score of 34% and 38%, and a macro-average AVPs of 32% and 42.2% for the Ekman and sentiment tasks respectively, showing no improvement in performance when compared to the fine-grained approach, when the expected result (as per the original article) was an improvement. One possible reason for this discrepancy is noise introduced by the translation to Portuguese, as Hammes e Freitas (2021) do not report results for the simpler taxonomies. The general poor performance of the chosen pre-trained model on this task is a potential concern: since the model struggles to learn adequate representations for these auxiliary tasks, how can the primary task benefit from these non-optimal representations? However, it should be noted that the model is capable of accurately distinguishing some emotion categories, which might prove useful. Another relevant factor regarding performance analysis concerns the fact that these tasks all share the same input data, which can potentially help to answer RQ4, by providing evidence pertaining to label distribution as a predictor of positive transfer, since $p(X)$ in Equation 1 is identical between these tasks and the models struggle with each of them, so both factors are controlled for.

Table 19 reports results for the binary post-level classification task for the depression sign dataset, utilizing the presence of the <PDP> label in a post as the labeling criterion. The model scores a macro-average AVPs of 83.8%, which is not surprising, as it mirrors

Table 16 – Average precision scores for the fine-grained classification task

Emotion	AVPs
Admiration	71.2%
Amusement	80.1%
Anger	41.4%
Annoyance	31.4%
Approval	36.1%
Caring	37.0%
Confusion	42.2%
Curiosity	53.7%
Desire	46.3%
Disappointment	24.9%
Disapproval	35.5%
Disgust	49.9%
Embarrassment	30.6%
Excitement	39.9%
Fear	70.4%
Gratitude	95.4%
Grief	41.2%
Joy	51.6%
Love	84.1%
Nervousness	27.7%
Optimism	51.6%
Pride	43.7%
Realization	20.5%
Relief	16.7%
Remorse	54.5%
Sadness	56.9%
Surprise	50.8%
Neutral	69.8%

Source: Produced by author

the common binary depression classification setup commonly seen in the literature, which often achieves good performance.

Table 20 reports results for the r/desabafos tag classification task. The model achieves adequate performance for most categories, with the exception of posts tagged as “Good News”. The chosen architecture can adequately model this task, and the associated dataset is the largest one among other auxiliary tasks. If the chosen tags are indeed

Table 17 – Average precision scores for the Ekman emotion classification task

Emotion (Ekman)	AVPs
Joy	10.3%
Anger	30.6%
Surprise	54.2%
Sadness	55.7%
Disgust	50.4%
Fear	24.8%
Neutral	69.2%

Source: Produced by author

Table 18 – Average precision scores for the sentiment classification task

Sentiment	AVPs
Positive	1.4%
Negative	32.5%
Ambiguous	32.6%
Neutral	46.6%

Source: Produced by author

Table 19 – AVPs for the binary depression classification task

Category	AVPs
Positive	81.7%
Negative	85.8%
Macro average	83.8%

Source: Produced by author

useful for the sign classification task, and the dataset is "similar" given the domain (see Section 5.3), then this task is a potentially adequate candidate for positive transfer.

5.4.2 Evaluating primary task performance

Multi-task models were trained is a similar fashion to the auxiliary task baselines, utilizing `optuna` for hyper parameter search. All models were trained with the same "tree-like" multi-task neural architecture, with shared parameters consisting of a pre-trained BERTimbau language model and single linear layers for each task's dedicated parameters.

Table 20 – AVPs for the r/desabafos tag classification task

Category	AVPs
Good news	49.3%
Depression	69.9%
Daily life	75.2%
Bad news	78.0%
Relationships	91.4%
Macro average	72.8%

Source: Produced by author

Hyper parameter search was expanded in order to include the task weights, the selection of individual task losses (either focal loss or binary cross entropy, as per Section 5) and the utilization of Projected Attention Layers (PAL) (STICKLAND; MURRAY, 2019b) adapters, which are all potentially relevant to proper positive transfer (as explained in Section 2.1).

More involved multi-task neural architectures and learning algorithms cited in previous sections were not considered in order to simplify the experiment, as they proved both too computationally expensive and could potentially mislead analysis of results, since the introduction of bespoke components to the base architecture would have to be accounted for in ablation experiments, which fall outside the scope of this study [6]. A notable exception is the application of PCGrad, since it is a technique that targets the exclusively multi-task phenomenon of gradient conflict, and introduces no new dynamic to the training process that could be accounted for in a single-task baseline. PCGrad results are reported and discussed separately.

Tables 21, 22 and 23 detail the performance of multi-task models for each of the task pairs. It can be observed that only the simplified sign taxonomy and GoEmotions sentiment taxonomy tasks led to successful positive transfer under traditional joint learning, with all other candidate tasks resulting in an overall performance loss. With regards to RQ3, it is indeed possible to improve performance for the fine-grained depression sign classification task, albeit only slightly, and with potential severe downside to some categories: performance for the categories “Feelings of guilt”, “Protective Factor”, “Sadness/Depressive mood”, “Worry/Fear/Anxiety”,“Sleep disorders” was universally impacted by negative transfer regardless of the chosen task. Conversely, “Alteration in Efficiency/Functionality” and “Loss/Diminishment of pleasure/libido” improved for

[6] for example, the introduction of additional parameters in architectures such as Cross-stitch and Sluice networks would have to be accounted for in the single-task scenario, resulting in an effective rerun of a substantial portion of Section 5 experiments

the majority of auxiliary tasks.

No other pattern could be discerned for the other sign categories, as most of these were only improved by some tasks. It is possible to draw parallels between some tasks and improvements to particular signs. For example, the occupation classification task led to the biggest improvement for the "Alteration in efficiency/functionality" sign, and the r/desabafos tasks, which successfully modeled the "Relationships" category in a single task setting led to the biggest gain in performance for the "Loss/Dimishment of pleasure/libido" category. However, there are just as many instances where these "intuitive" parallels, often invoked in the literature when discussing task "similarity", break down. For example, none of the tasks derived from GoEmotions led to an improvement for the "Feelings of guilt" sign, despite remorse being one of the emotions covered by the fine-grained taxonomy (and thus was also present in the other two tasks).

Table 21 – Average precision scores for the Simplified taxonomy and r/desabafos tasks. Improved scores in bold

Sign	**Single-task**	**Simplified**	**r/desabafos**
Feeling of emptiness	59.3%	51.8%	46.9%
Irritation/Aggressiveness	42.9%	39.4%	**45.8%**
Alteration in efficiency/functionality	40.9%	**46.1%**	**41.6%**
Despair	43.6%	**55.06%**	**50.4%**
Feeling of guilt	58.6%	40.4%	42.4%
Difficulty in decision making	43.6%	**91.6%**	43.3%
Feeling of worthlessness/Low self-esteem	53%	43.6%	50.2%
Risk factor	62.5%	60.3%	60.6%
Tiredness/Discouragement/Fatigue	56.9%	57.8%	**59.2%**
Death/suicide of third party	58.3%	**60.9%**	46.4%
Loss/Diminishment of pleasure/libido	0.7%	**71.4%**	**98%**
Protective Factor	72.2%	68.3%	67%
Helplessness/Social harm/Loneliness	69.2%	64.8%	69.2%
Attention/memory deficit	31.6%	0.09%	0.08%
Alteration in weight/eating habits	54.6%	52.2%	**69.7%**
Physical symptom	48.9%	**52.7%**	21.4%
Suicide/Self-extermination	77.1%	71.5%	**79.1%**
Sadness/Depressed mood	87.3%	77.8%	75.8%
Worry/Fear/Anxiety	90.9%	82.2%	81.1%
Sleep disorder	87.7%	79.7%	67.6%
Macro average	56.9%	**58.4%**	55.7%

Source: Produced by author

These same models were retrained with the addition of PCGrad gradient conflict correction, in order to better understand the role of gradient behavior for multi-task learning in this domains, since there is conflicting evidence regarding gradient conflict as a predictor of positive transfer (NI et al., 2023; YU et al., 2020). Table 24 summarizes the results for the PCGrad models in comparison to the single-task baseline and their multi-

Table 22 – Average precision scores for the binary depression classification and occupation classification tasks. Improved scores in bold

Sign	Single-task	Binary	Occupation
Feeling of emptiness	59.3%	52.1%	**64.7%**
Irritation/Aggressiveness	42,9%	**46.8%**	42.9%
Alteration in efficiency/functionality	40.9%	38.1%	**53.1%**
Despair	43.6%	**49.4%**	40.4%
Feeling of guilt	58.6%	37.1%	41.1%
Difficulty in decision making	43.6%	**71.2%**	33%
Feeling of worthlessness/Low self-esteem	53%	40.5%	38.4%
Risk factor	62.5%	52.9%	57.9%
Tiredness/Discouragement/Fatigue	56.9%	47.9%	44.4%
Death/suicide of third party	58.3%	**68.9%**	**74.3%**
Loss/Diminishment of pleasure/libido	0.7%	**51.2%**	**52.1%**
Protective Factor	72.2%	60.6%	68%
Helplessness/Social harm/Loneliness	69.2%	61.6%	66.1%
Attention/memory deficit	31.6%	**53.04%**	15.6%
Alteration in weight/eating habits	54.6%	43.8%	**63.1%**
Physical symptom	48.9%	**62.5%**	**51.4%**
Suicide/Self-extermination	77.1%	74%	70%
Sadness/Depressed mood	87.3%	73.8%	73.2%
Worry/Fear/Anxiety	90.9%	84.8%	81.4%
Sleep disorder	87.7%	68.5%	74.6%
Macro average	56.9%	56.9	55.2%

Source: Produced by author

task counterparts. Most tasks suffered a loss in performance after the application of PCGrad, save for the Ekman and fine-grained emotion classification tasks. These results imply that gradient conflict (either by itself, or as part of the "tragic triad" criteria) is not the primary driver of negative transfer for most of the evaluated tasks, although it is still a potentially relevant factor.

These experiments successfully answered RQ3, however no better understanding of what leads to positive transfer was achieved. These concerns were addressed by a subsequent set of experiments aimed at RQ4. It should be noted that the experiments in this section serve as the foundation for analysis of task characteristics, since they determine a benchmark against which candidate task characteristics should be evaluated: they should be particularly pronounced or exclusive to the three tasks that led to positive transfer.

5.4.3 Analyzing task characteristics for predictors of positive transfer

A set of possibly predictive task characteristics was selected in order to answer RQ4 and further understand how multi-task models behave on the complex domain of mental

Table 23 – Average precision scores for the GoEmotions tasks. Improved scores in bold

Sign	**Single-task**	**Emotions**	**Ekman**	**Sentiment**
Feeling of emptiness	59.3%	41.9%	39.5%	**62.6%**
Irritation/Aggressiveness	42.9%	36.7%	**46.3%**	**43%**
Alteration in efficiency/functionality	40.9%	39.1%	**42.4%**	**52.2%**
Despair	43.6%	**46.4%**	**50.3%**	**49.8%**
Feeling of guilt	58.6%	54.2%	29.9%	42.4%
Difficulty in decision making	43.6%	**49.2%**	43.5%	39.5%
Feeling of worthlessness/Low self-esteem	53.0%	48.3%	45.9%	46.5%
Risk factor	62.5%	**64.3%**	50.9%	59.7%
Tiredness/Discouragement/Fatigue	56.9%	52.9%	46.8%	51.6%
Death/suicide of third party	58.3%	39.3%	37.7%	47.0%
Loss/Diminishment of pleasure/libido	0.7%	**14.6%**	**12.0%**	**64.3%**
Protective Factor	72.2%	72.1%	70.5%	68.6%
Helplessness/Social harm/Loneliness	69.2%	**70.0%**	58.6%	64.7%
Attention/memory deficit	31.6%	**33.2%**	13.3%	**64.5%**
Alteration in weight/eating habits	54.6%	43.6%	**61.8%**	**63.0%**
Physical symptom	48.9%	40.0%	29.1%	32.6%
Suicide/Self-extermination	77.1%	76.1%	69.0%	72.8%
Sadness/Depressed mood	87.3%	83.7%	73.6%	74.9%
Worry/Fear/Anxiety	90.9%	84.9%	77.5%	86.7%
Sleep disorder	87.7%	49.5%	54.2%	80.1%
Macro Average	56.9%	52.0%	47.0%	**58.32%**

Source: Produced by author

Table 24 – Macro AVPs comparison between multi-task models

Task	Multi-task	PCGrad
Binary	56.9%	54.5%
Ekman	47.0%	50.1%
Emotion	52.0%	58.6%
r/desabafos	55.7%	51.0%
Occupation	55.2%	53.4%
Sentiment	58.32%	55.3%
Simplified	58.4%	52.6%

Source: Produced by author

health. These were gathered based on the studies cited in Section 3.1, and can be broadly divided into two categories: characteristics that can be extracted prior to the training of any model, be it multi-task or it's associated primary and auxiliary single-task baselines, and those that require prior training. The former consists of information theoretical

metrics proposed by Alonso e Plank (2017) – label entropy and kurtosis – which capture the label distribution of a given dataset; characteristics that model how closely data from different tasks "resemble" each other by some mean: computing the average task embeddings from each dataset and comparing them through cosine similarity – denoted as "TextEmb", as per Vu et al. (2020) and employed in a multi-task context by Ni et al. (2023) – checking vocabulary overlap and absolute differences in some text properties, such as average length and token-type ratio, and evaluating the impact of the different dataset sizes associated with each task. Characteristics that require the training of models aim to capture the behavior of said models during their training, and were represented by the "tragic triad" criteria introduced by Yu et al. (2020) and discussed in Section 3.1: conflicting gradients, gradient magnitude and curvature.

Table 25 details the entropy, kurtosis and dataset size values for each task. Note that these differ from the other characteristics, as they do not attempt to model the relationship between the tasks, be it during training or by trying to model the notion of "similarity", which means there is an assumption that some tasks are inherently more appropriate for joint learning regardless of primary task. The Spearman rank correlation coefficient between all of these values and the performance of the multi-task model was calculated, but no statistically significant correlation was found, failing to disprove the null hypothesis (reported in Table 26). It can be observed that the two of the successful auxiliary tasks (sentiment classification and simplified depression sign classification) have similar entropy values, and are in the "middle ground" in terms of kurtosis, potentially distinguishing them from other tasks. Additionaly, the sentiment classification task has the lowest kurtosis value in comparison to other GoEmotions tasks, and is the only one that caused positive transfer without the use of PCGrad. However it cannot be definitively stated that there is a range of ideal values for these two characteristics, given the lack of correlation and the counterexample of r/desabafos, which has the highest entropy and lowest kurtosis and still led to negative transfer. Dataset size has no correlation with positive transfer in the evaluated tasks, which is corroborated by the literature on transfer learning (VU et al., 2020).

With regards to task characteristics extracted from text, there is also no statistically significant correlation with model performance. Tables 27 and 28 detail the respective values and Spearman rank correlation statistics for the TextEmb, token-type ratio, text length and vocabulary overlap characteristics[7]. These results imply that notions of "similarity" between task inputs, at least as it pertains to the proposed depression sign classification task, is not an adequate predictor of positive transfer.

Finally, the "tragic triad" information was collected for the best performing model of each task. Table 29 details the values of characteristics based on gradient information and Table 30 report the Spearman rank correlation coefficient and p-values between these

[7] All tasks derived from GoEmotions share the same corpus, so the are simply reported as "GoEmotions"

Table 25 – Values for entropy, kurtosis and dataset size

Task	Entropy	Kurtosis	Dataset size
Emotions	9.74	31.7	54263
Ekman	8.68	30.2	54263
Sentiment	7.98	21.5	54263
r/desabafos	11.24	-0.94	90656
Occupation	5.78	-0.14	363
Simplified	7.47	12.3	2675
Binary	5.73	-1.95	2675

Source: Produced by author

Table 26 – Spearman correlation coefficients and p-values for each task with regards to similarity-agnostic characteristics

Characteristic	Coefficient	p-value
Entropy	-0.39	0.38
Kurtosis	-0.42	0.33
Dataset size	-0.18	0.69

Source: Produced by author

Table 27 – Values for TextEmb, Token-type ratio, text length and vocabulary overlap

Task	TextEmb	Text length	Token-type ratio	Overlap (primary->auxiliary)	Overlap (auxiliary->primary)
GoEmotions	0.95	2.8	13.5	11%	74%
r/desabafos	0.89	186	256	0.2%	0.98%
Occupation	0.98	5	3.8	87.2%	19%
Simplified	1	0	0	100%	100%
Binary	0.9	156	2.89	63.9%	99%

Source: Produced by author

characteristics and the performance of the multi-task models. As was the case with other task characteristics, none of the gradient-based ones demonstrated a correlation that disproves the null hypothesis. When observing the listed values, one notable outlier is the simplified taxonomy task, which is the only one that does not suffer from almost any

Table 28 – Spearman correlation coefficients and p-values for each task with regards to text-based characteristics

Characteristic	**Coefficient**	**p-value**
TextEmb	0.18	0.69
Text length	-0.18	0.69
Token-type ratio	-0.48	0.27
Overlap (primary->auxiliary)	0.47	0.36
Overlap (auxiliary->primary)	0.63	0.12

Source: Produced by author

gradient conflict during training. One potential explanation, which would also align with observations made by Liu, Davison e Johns (2019), is that the hierarchical structure of this task pair does not lend itself to gradient conflict, since the input data for samples in each batch is shared between tasks, and these tasks tend to produce gradients with similar directionality, since they model the same task at different granularity levels. If this explanation is correct, utilizing tasks associated with disparate datasets for joint learning purposes is inherently harder, which means it is also harder to utilize multi-task learning for overcoming data scarcity constraints, the main motivation for this study's approach, although the improved performance for the emotion and sentiment classification tasks is evidence that this approach is potentially viable.

In conclusion, with regards to RQ4, the auxiliary tasks that resulted in positive transfer do not share task characteristics that distinguish them from other candidate tasks, among 12 different characteristics, which include both those that can be observed prior to training and after the fact.

Table 29 – Joint learning gradient-based characteristics

Task	Conflict ratio	Dominant primary task	Magnitude similarity	Average curvature
Emotions	42.4%	27.1%	0.74	$-1.19e^{-09}$
Ekman	44.2%	17.64%	0.71	$-4.87e^{-10}$
Sentiment	46.5%	8.68%	0.574	$-1.23e^{-10}$
r/desabafos	44.1%	6.53%	0.578	$-7.21e^{-11}$
Occupation	38.2%	73.5%	0.457	$9.51e^{-12}$
Simplified	0.8%	71.4%	0.606	$-2.66e^{-10}$
Binary	38.1%	69.9%	0.32	$-1.2e^{-09}$

Source: Produced by author

Table 30 – Spearman correlation coefficients and p-values for each task with regards to gradient-based characteristics

Characteristic	Coefficient	p-value
Conflict ratio	-0.35	0.43
Dominant primary task	0.11	0.81
Curvature	0.11	0.81
Magnitude Similarity	-0.42	0.33

Source: Produced by author

Chapter 6

Conclusion

This study investigated the identification of signs of depression in a fine-grained label set consisting of 18 distinct symptoms and 3 additional categories associated with depression, elaborated with the help of a multidisciplinary team of mental health experts. This resource was then use to inform the annotation of a dataset of anonymous Facebook posts obtained from "university secrets" pages. Although it is not possible to share the corpus because it contains sensitive data (even anonymized), the annotation guidelines are **one of the contributions** of the Amive project.

In order to better understand how depressed people express their condition in social media platforms, through text, RQ1 was defined: **Which of the 21 signs of depression can be frequently found in online text?** It was observed that the proposed signs followed a long-tailed distribution, with a substantial portion of the evaluated signs being infrequent in social media text, while some are ubiquitous. Of particular importance is the frequency of the "Risk Factor", "Protective Factor" and "Alteration in efficiency/functionality" categories, which are seldom drawn attention to in the investigated literature. Additionally, by adopting a simplified taxonomy consisting of somatic, behavioral, emotional and external signs, **it becomes clear that emotional and external signs of depression are frequently expressed, while somatic features are not**.

The dataset annotated with the 21 signs of depression was subsequently used to train machine learning models for the task of depression sign classification, in order to answer **RQ2: Which of these signs can be most easily identified using machine learning techniques?** This task proved challenging, with only six of the signs being robustly discerned by any of the evaluated approaches – "Sleep Disorder", "Worry/Fear/Anxiety", "Sadness/Depressed mood", "Suicide/Self-extermination", "Protective Factor" and "Physical symptom". Notably, two of these are among the most

infrequent signs, and some frequent signs such as "Risk Factor" and "Alteration in efficiency/functionality" could not be easily discerned, suggesting that label imbalances were not solely to blame. An analysis of the data suggested that not only was the available data scarce (given the necessity of expert involvement during annotation which tends to be slow and costly in comparison to other forms of modeling depression detection), but that some of the signs were also inherently harder to classify than others, with frequent employment of figurative and indirect language. By grouping these signs in the aforementioned simplified taxonomy, **it was observed that somatic signs were the most easily classified, despite also being the most infrequent**. These results were reported in Mendes e Caseli (2024), another **contribution** of this work.

A diverse set of techniques that could potentially mitigate issues regarding data scarcity and improve the generalization of the models was selected, including various data augmentation, regularization and semi-supervised techniques, a few-shot prompt engineering approach and multi-task learning. Given preliminary results, the fine-tuning of pre-trained language models combined with multi-task learning was chosen as the avenue to be explored in future computational experiments. With this focus on a particular set of techniques came two new research questions (RQ3 and RQ4) revolving around the mechanisms that underpin joint learning, which are still a subject of debate in the literature. **RQ3** concerns the impact of joint learning for the fine-grained depression sign classification task: **Can multi-task learning improve performance for the fine-grained depression task? If so, are there particular signs that benefit from it? Which of the auxiliary tasks improve performance for the primary task?**

A set of 7 different tasks were selected as candidate auxiliary tasks. Three tasks pertain to emotion classification, all derived from GoEmotions and distinct only in terms of label distribution, as they were based on emotion taxonomies with differing granularity (fine-grained, Ekman and sentiment); one task shared the same input data of the primary task, but grouped sign categories into the simplified 4 sign taxonomy; 2 additional tasks utilized the same dataset under similar annotation guidelines, one modeling different occupation categories and the other potential depressive profiles; and finally another was a post category classification task collected from the r/desabafos community, which is analogous to the "university secrets" pages in several ways. **3 of the evaluated auxiliary tasks led to positive transfer**, an improvement in overall performance relative to the single-task baseline: **the fine-grained emotion classification task, sentiment classification task and simplified sign classification task**. However, this improvement was slight and came with tradeoffs, since it resulted in loss of performance for several signs. Only **two signs were improved by the majority of auxiliary tasks "Alteration in efficiency/functionality" and "Loss/Diminishment of pleasure/libido"**. No other pattern could be discerned from the joint learning results, which means that **most signs are not inherently benefited by the multi-task learning setup**.

Finally, **RQ4** was intended to further the understanding of indicators of positive transfer in the complex domain of mental health: **are there particular characteristics that good auxiliary tasks have in common? Can they be observed prior to training?** A collection of 12 characteristics of task-specific characteristics was extracted, consisting of: label entropy and kurtosis, dataset size, TextEmb (cosine similarity of averaged contextual embeddings), absolute differences in mean length and token type ratio, overlaps in vocabulary, the ratio of gradient conflict, the ratio with which the primary task dominated the optimization process, the magnitude similarity between task gradients and the average curvature during training. None of these measures achieved statistically significant correlation with positive transfer, thus **the answer to RQ4 is that there is no discernible predictor of good auxiliary tasks among evaluated task characteristics, be it prior of after training, under the studied scenario**.

In conclusion, while the proposed fine-grained depression sign classification approach allowed a more thorough understanding of online depression, and has the potential to provide more detailed information about a social media user's mental state, this approach comes with serious challenges. Multi-task learning, which was the only alternative approach that led to any improvement over baseline, has several challenges of its own, with many open questions, specially regarding the mechanisms that lead to positive transfer, selecting appropriate auxiliary tasks and solving optimization problems exclusive to joint learning approaches. Directions for future work include checking for potential gaps in current binary and symptom-based depression classification models with regards to the proposed fine-grained taxonomy, conducting a more comprehensive set of experiments on an expanded set of task and task characteristics in the mental health domain – as there is a dearth of information regarding joint learning behavior in complex domains (NI et al., 2023) – and a reevaluation of the prompt engineering approach with closed source, larger capacity and more recent LLMs.

References

ALONSO, H. M.; PLANK, B. When is multitask learning effective? Semantic sequence prediction under varying data conditions. In: **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, EACL, Long Papers**. Association for Computational Linguistics (ACL), 2017. Disponível em: <https://arxiv.org/abs/1612.02251>.

American Psychiatric Association. **Diagnostic and Statistical Manual of Mental Disorders: DSM-5-TR**. American Psychiatric Association Publishing, 2022. ISBN 9780890425756. Disponível em: <https://books.google.com.br/books?id=kYyizgEACAAJ>.

ARACI, D. Finbert: Financial sentiment analysis with pre-trained language models. **arXiv preprint arXiv:1908.10063**, 2019.

BARONI, M. et al. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. **Language resources and evaluation**, Springer, v. 43, p. 209–226, 2009.

BECK, A. T.; STEER, R. A.; BROWN, G. **Beck Depression Inventory–II**. American Psychological Association (APA), 1996. Disponível em: <https://doi.org/10.1037/t00742-000>.

BINGEL, J.; SØGAARD, A. Identifying beneficial task relations for multi-task learning in deep neural networks. In: **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers**. Valencia, Spain: Association for Computational Linguistics, 2017. p. 164–169. Disponível em: <https://aclanthology.org/E17-2026>.

CAMBRIA, E. et al. Affectivespace 2: Enabling affective intuition for concept-level sentiment analysis. In: **Proceedings of the AAAI conference on artificial intelligence**. [S.l.: s.n.], 2015. v. 29, n. 1.

_____. Senticnet. **Sentic computing: a common-sense-based framework for concept-level sentiment analysis**, Springer, p. 23–71, 2015.

_____. Semantic multidimensional scaling for open-domain sentiment analysis. **IEEE intelligent systems**, IEEE, v. 29, n. 2, p. 44–51, 2012.

CARUANA, R. Multitask learning. **Machine learning**, v. 28, n. 1, p. 41–75, 1997. Publisher: Springer. Disponível em: <https://link.springer.com/article/10.1023/A:1007379606734>.

CASANI, V. et al. DP-symptom-identifier: uma estratégia para classificar sintomas de depressão utilizando um conjunto de dados textuais na língua portuguesa. In: **Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL 2021)**. Sociedade Brasileira de Computação, 2021. Disponível em: <https://doi.org/10.5753/stil.2021.17794>.

CAVESTRO, J. d. M.; ROCHA, F. L. Prevalência de depressão entre estudantes universitários. **Jornal brasileiro de psiquiatria**, SciELO Brasil, v. 55, p. 264–267, 2006.

CHEN, C.-C.; HUANG, H.-H.; CHEN, H.-H. Numeral attachment with auxiliary tasks. In: **Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval**. [S.l.: s.n.], 2019. p. 1161–1164.

_____. Numclaim: Investor's fine-grained claim detection. In: **Proceedings of the 29th ACM International Conference on Information & Knowledge Management**. [S.l.: s.n.], 2020. p. 1973–1976.

CHEN, T. et al. The lottery ticket hypothesis for pre-trained bert networks. **Advances in neural information processing systems**, v. 33, p. 15834–15846, 2020.

CHOUDHURY, M. D. et al. Predicting depression via social media. In: **Proceedings of the international AAAI conference on web and social media**. [S.l.: s.n.], 2013. v. 7, n. 1, p. 128–137.

CLARK, K. et al. BAM! born-again multi-task networks for natural language understanding. In: **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**. Florence, Italy: Association for Computational Linguistics, 2019. p. 5931–5937. Disponível em: <https://aclanthology.org/P19-1595>.

COPPERSMITH, G. et al. Clpsych 2015 shared task: Depression and ptsd on twitter. In: **Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality**. [S.l.: s.n.], 2015. p. 31–39.

CORTIS, K. et al. Semeval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In: **Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)**. [S.l.: s.n.], 2017. p. 519–535.

Crowdtangle Team. **Crowdtangle**. 2021. List ID: 1479661. Disponível em: <https://www.crowdtangle.com/>.

DALGALARRONDO, P. **Religião, psicopatologia e saúde mental**. [S.l.]: Artmed Editora, 2009.

DEMSZKY, D. et al. GoEmotions: A Dataset of Fine-Grained Emotions. In: **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**. Online: Association for Computational Linguistics, 2020. p. 4040–4054. Disponível em: <https://www.aclweb.org/anthology/2020.acl-main.372>.

DEVLIN, J. et al. BERT: pre-training of deep bidirectional transformers for language understanding. **CoRR**, abs/1810.04805, 2018. Disponível em: <http://arxiv.org/abs/1810.04805>.

EVANS, T. M. et al. Evidence for a mental health crisis in graduate education. **Nature biotechnology**, Nature Publishing Group US New York, v. 36, n. 3, p. 282–284, 2018.

FAN, A. et al. Beyond english-centric multilingual machine translation. **J. Mach. Learn. Res.**, JMLR.org, v. 22, n. 1, jan 2021. ISSN 1532-4435.

FILHO, J. A. W. et al. The brwac corpus: a new open resource for brazilian portuguese. In: **Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)**. [S.l.: s.n.], 2018.

FILHO, P. P. B.; PARDO, T. A. S.; ALUÍSIO, S. M. An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In: **Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology**. [s.n.], 2013. Disponível em: <https://aclanthology.org/W13-4829>.

FINN, C.; ABBEEL, P.; LEVINE, S. Model-agnostic meta-learning for fast adaptation of deep networks. In: PMLR. **International conference on machine learning**. [S.l.], 2017. p. 1126–1135.

FRANKLE, J.; CARBIN, M. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In: **International Conference on Learning Representations**. [s.n.], 2019. Disponível em: <https://openreview.net/forum?id=rJl-b3RcF7>.

FREIRE, M. et al. Escala hamilton: estudo das características psicométricas em uma amostra do sul do brasil. **Jornal Brasileiro de Psiquiatria**, Instituto de Psiquiatria da Universidade Federal do Rio de Janeiro, v. 63, n. 4, p. 281–289, Oct 2014. ISSN 0047-2085. Disponível em: <https://doi.org/10.1590/0047-2085000000036>.

GHOSH, S.; EKBAL, A.; BHATTACHARYYA, P. A Multitask Framework to Detect Depression, Sentiment and Multi-label Emotion from Suicide Notes. **Cognitive Computation**, v. 14, n. 1, p. 110–129, jan. 2022. ISSN 1866-9964. Disponível em: <https://doi.org/10.1007/s12559-021-09828-7>.

GRATCH, J. et al. The distress analysis interview corpus of human and computer interviews. In: REYKJAVIK. **LREC**. [S.l.], 2014. p. 3123–3128.

HAMMES, L. O. A.; FREITAS, L. A. de. Utilizando bertimbau para a classificação de emoções em português. In: SBC. **Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana**. [S.l.], 2021. p. 56–63.

HE, K. et al. Deep residual learning for image recognition. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2016. p. 770–778.

HOSPEDALES, T. et al. Meta-Learning in Neural Networks: A Survey. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 44, n. 9, p. 5149–5169, set. 2022. ISSN 1939-3539. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

HOULSBY, N. et al. Parameter-efficient transfer learning for NLP. In: CHAUDHURI, K.; SALAKHUTDINOV, R. (Ed.). **Proceedings of the 36th International Conference on Machine Learning**. PMLR, 2019. (Proceedings of Machine Learning Research, v. 97), p. 2790–2799. Disponível em: <https://proceedings.mlr.press/v97/houlsby19a.html>.

JI, S. et al. Supervised learning for suicidal ideation detection in online user content. **Complexity**, v. 2018, p. 1–10, 09 2018.

_____. MentalBERT: Publicly available pretrained language models for mental healthcare. In: **Proceedings of the Thirteenth Language Resources and Evaluation Conference**. Marseille, France: European Language Resources Association, 2022. p. 7184–7190. Disponível em: <https://aclanthology.org/2022.lrec-1.778>.

JIANG, H. et al. SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization. In: **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**. Online: Association for Computational Linguistics, 2020. p. 2177–2190. Disponível em: <https://www.aclweb.org/anthology/2020.acl-main.197>.

JOTY, S.; MÀRQUEZ, L.; NAKOV, P. Joint multitask learning for community question answering using task-specific embeddings. In: **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**. [S.l.: s.n.], 2018. p. 4196–4207.

KENDALL, A.; GAL, Y.; CIPOLLA, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2018. p. 7482–7491.

KOO, T. K.; LI, M. Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. **Journal of chiropractic medicine**, Elsevier, v. 15, n. 2, p. 155–163, 2016.

KRIPPENDORFF, K. On the reliability of unitizing continuous data. **Sociological Methodology**, JSTOR, p. 47–76, 1995.

KRISTENSEN, C. H. et al. Normas brasileiras para o affective norms for english words. **Trends in Psychiatry and Psychotherapy**, SciELO Brasil, v. 33, p. 135–146, 2011.

KROENKE, K.; SPITZER, R. L.; WILLIAMS, J. B. The phq-9: validity of a brief depression severity measure. **Journal of general internal medicine**, Wiley Online Library, v. 16, n. 9, p. 606–613, 2001.

LAMM, M. et al. Textual analogy parsing: What's shared and what's compared among analogous facts. In: **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**. Brussels, Belgium: Association for Computational Linguistics, 2018. p. 82–92. Disponível em: <https://aclanthology.org/D18-1008>.

LEAL, S. E. et al. Nilc-metrix: assessing the complexity of written and spoken language in brazilian portuguese. **Language Resources and Evaluation**, Springer, v. 58, n. 1, p. 73–110, 2024.

LEE, A. et al. Micromodels for Efficient, Explainable, and Reusable Systems: A Case Study on Mental Health. In: **Findings of the Association for Computational Linguistics: EMNLP 2021**. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. p. 4257–4272. Disponível em: <https://aclanthology.org/2021.findings-emnlp.360>.

LI, C.; BRAUD, C.; AMBLARD, M. Multi-task learning for depression detection in dialogs. In: **Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue**. [S.l.: s.n.], 2022. p. 68–75.

LI, Y. et al. DailyDialog: A manually labelled multi-turn dialogue dataset. In: **Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**. Taipei, Taiwan: Asian Federation of Natural Language Processing, 2017. p. 986–995. Disponível em: <https://aclanthology.org/I17-1099>.

LIN, T.-Y. et al. Focal loss for dense object detection. In: **Proceedings of the IEEE international conference on computer vision**. [S.l.: s.n.], 2017. p. 2980–2988.

LIU, B. et al. Conflict-averse gradient descent for multi-task learning. **Advances in Neural Information Processing Systems**, v. 34, p. 18878–18890, 2021.

LIU, D. et al. Detecting and measuring depression on social media using a machine learning approach: systematic review. **JMIR Mental Health**, JMIR Publications Inc., Toronto, Canada, v. 9, n. 3, p. e27244, 2022.

_____. Detecting and Measuring Depression on Social Media Using a Machine Learning Approach: Systematic Review. **JMIR Mental Health**, v. 9, n. 3, p. e27244, 2022. Publisher: JMIR Publications Inc., Toronto, Canada.

LIU, L. et al. Towards impartial multi-task learning. In: **International Conference on Learning Representations**. [S.l.: s.n.], 2020.

LIU, S.; DAVISON, A.; JOHNS, E. Self-supervised generalisation with meta auxiliary learning. **Advances in Neural Information Processing Systems**, v. 32, 2019.

LIU, S.; JOHNS, E.; DAVISON, A. J. End-to-end multi-task learning with attention. In: **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**. [S.l.: s.n.], 2019. p. 1871–1880.

LIU, Y. et al. Improved Depression Recognition Using Attention and Multitask Learning of Gender Recognition. In: **2021 International Conference on Asian Language Processing (IALP)**. [S.l.: s.n.], 2021. p. 57–61.

_____. **RoBERTa: A Robustly Optimized BERT Pretraining Approach**. 2019. Disponível em: <https://arxiv.org/abs/1907.11692>.

LOSHCHILOV, I.; HUTTER, F. Decoupled weight decay regularization. In: **International Conference on Learning Representations**. [s.n.], 2019. Disponível em: <https://openreview.net/forum?id=Bkg6RiCqY7>.

MALO, P. et al. Good debt or bad debt: Detecting semantic orientations in economic texts. **Journal of the Association for Information Science and Technology**, Wiley Online Library, v. 65, n. 4, p. 782–796, 2014.

MARIKO, D. et al. The financial document causality detection shared task (FinCausal 2020). In: **Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation**. Barcelona, Spain (Online): COLING, 2020. p. 23–32. Disponível em: <https://aclanthology.org/2020.fnp-1.3>.

MENDES, A. R.; CASELI, H. Identifying fine-grained depression signs in social media posts. In: CALZOLARI, N. et al. (Ed.). **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**. Torino, Italia: ELRA and ICCL, 2024. p. 8594–8604. Disponível em: <https://aclanthology.org/2024.lrec-main.754>.

MENDES, A. R.; PASSADOR, R. V.; CASELI, H. M. Identificando sintomas de depressao em postagens do twitter em português do brasil. In: SBC. **Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana**. [S.l.], 2021. p. 162–171.

MILNE, D. N. et al. CLPsych 2016 shared task: Triaging content in online peer-support forums. In: HOLLINGSHEAD, K.; UNGAR, L. (Ed.). **Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology**. San Diego, CA, USA: Association for Computational Linguistics, 2016. p. 118–127. Disponível em: <https://aclanthology.org/W16-0312>.

MISRA, I. et al. Cross-stitch networks for multi-task learning. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2016. p. 3994–4003.

MOWERY, D. L.; BRYAN, C.; CONWAY, M. Towards developing an annotation scheme for depressive disorder symptoms: A preliminary study using twitter data. In: **Proceedings of the 2nd workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality**. [S.l.: s.n.], 2015. p. 89–98.

NAMBISAN, P. et al. Social media, big data, and public health informatics: Ruminating behavior of depression revealed through twitter. In: **Proceedings of the 2015 48th Hawaii International Conference on System Sciences**. USA: IEEE Computer Society, 2015. (HICSS '15), p. 2906–2913. ISBN 9781479973675. Disponível em: <https://doi.org/10.1109/HICSS.2015.351>.

NAVON, A. et al. Multi-task learning as a bargaining game. In: CHAUDHURI, K. et al. (Ed.). **Proceedings of the 39th International Conference on Machine Learning**. PMLR, 2022. (Proceedings of Machine Learning Research, v. 162), p. 16428–16446. Disponível em: <https://proceedings.mlr.press/v162/navon22a.html>.

NI, J. et al. When does aggregating multiple skills with multi-task learning work? a case study in financial NLP. In: **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Toronto, Canada: Association for Computational Linguistics, 2023. p. 7465–7488. Disponível em: <https://aclanthology.org/2023.acl-long.412>.

NIVRE, J. et al. Universal Dependencies v1: A multilingual treebank collection. In: CALZOLARI, N. et al. (Ed.). **Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)**. Portorož, Slovenia: European Language Resources Association (ELRA), 2016. p. 1659–1666. Disponível em: <https://aclanthology.org/L16-1262>.

PARABONI, I.; FUNABASHI, A. M. M.; SANTOS, W. Ramos dos. Searching Brazilian Twitter for signs of mental health issues. **LREC**, p. 7, maio 2020.

PENG, N.; DREDZE, M. Multi-task domain adaptation for sequence tagging. In: **Proceedings of the 2nd Workshop on Representation Learning for NLP**. Vancouver, Canada: Association for Computational Linguistics, 2017. p. 91–100. Disponível em: <https://aclanthology.org/W17-2612>.

RADLOFF, L. S. A self-report depression scale for research in the general population. **Applied psychol Measurements**, v. 1, p. 385–401, 1977.

REIMERS, N.; GUREVYCH, I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: INUI, K. et al. (Ed.). **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**. Hong Kong, China: Association for Computational Linguistics, 2019. p. 3982–3992. Disponível em: <https://aclanthology.org/D19-1410>.

RUDER, S. An overview of multi-task learning in deep neural networks. **arXiv preprint arXiv:1706.05098**, 2017. Disponível em: <https://arxiv.org/abs/1706.05098>.

RUDER, S. et al. Latent multi-task architecture learning. In: **Proceedings of the AAAI Conference on Artificial Intelligence**. [S.l.: s.n.], 2019. v. 33, n. 01, p. 4822–4829.

SAITO, T.; REHMSMEIER, M. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. **PloS one**, v. 10, n. 3, p. e0118432, 2015. ISSN 1932-6203. Disponível em: <https://europepmc.org/articles/PMC4349800>.

SANTOS, W. R. d.; OLIVEIRA, R. L. de; PARABONI, I. Setembrobr: a social media corpus for depression and anxiety disorder prediction. **Language Resources and Evaluation**, Springer, p. 1–28, 2023.

SCHRÖDER, F.; BIEMANN, C. Estimating the influence of auxiliary tasks for multi-task learning of sequence tagging tasks. In: **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**. Online: Association for Computational Linguistics, 2020. p. 2971–2985. Disponível em: <https://aclanthology.org/2020.acl-main.268>.

SCHWARTZ, H. A. et al. Towards Assessing Changes in Degree of Depression through Facebook. In: **Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality**. Baltimore, Maryland, USA: Association for Computational Linguistics, 2014. p. 118–125. Disponível em: <http://aclweb.org/anthology/W14-3214>.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. In: COMPUTATIONAL AND BIOLOGICAL LEARNING SOCIETY. **3rd International Conference on Learning Representations (ICLR 2015)**. [S.l.], 2015.

SKEVINGTON, S. M.; LOTFY, M.; O'CONNELL, K. A. The world health organization's whoqol-bref quality of life assessment: psychometric properties and results of the international field trial. a report from the whoqol group. **Quality of life Research**, Springer, v. 13, p. 299–310, 2004.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Bertimbau: pretrained bert models for brazilian portuguese. In: SPRINGER. **Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9**. [S.l.], 2020. p. 403–417.

STICKLAND, A. C.; MURRAY, I. Bert and pals: Projected attention layers for efficient adaptation in multi-task learning. In: PMLR. **International Conference on Machine Learning**. [S.l.], 2019. p. 5986–5995.

_____. BERT and PALs: Projected attention layers for efficient adaptation in multi-task learning. In: CHAUDHURI, K.; SALAKHUTDINOV, R. (Ed.). **Proceedings of the 36th International Conference on Machine Learning**. PMLR, 2019. (Proceedings of Machine Learning Research, v. 97), p. 5986–5995. Disponível em: <https://proceedings.mlr.press/v97/stickland19a.html>.

UBAN, A. S.; CHULVI, B.; ROSSO, P. Multi-aspect transfer learning for detecting low resource mental disorders on social media. In: **Proceedings of the Thirteenth Language Resources and Evaluation Conference**. [S.l.: s.n.], 2022. p. 3202–3219.

VASWANI, A. et al. Attention is all you need. In: **Proceedings of the 31st International Conference on Neural Information Processing Systems**. Red Hook, NY, USA: Curran Associates Inc., 2017. (NIPS'17), p. 6000–6010. ISBN 9781510860964.

VU, T. et al. Exploring and predicting transferability across nlp tasks. In: **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. [S.l.: s.n.], 2020. p. 7882–7926.

WANG, W. et al. Personalized microblog sentiment classification via adversarial cross-lingual multi-task learning. In: **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**. Brussels, Belgium: Association for Computational Linguistics, 2018. p. 338–348. Disponível em: <https://aclanthology.org/D18-1031>.

WANG, X. et al. TransVae:A Novel Variational Sequence-to-Sequence Framework for Semi-supervised Learning and Diversity Improvement. In: **2021 International Joint Conference on Neural Networks (IJCNN)**. [S.l.: s.n.], 2021. p. 1–8. ISSN: 2161-4407.

WANG, Y. et al. **A Multitask Deep Learning Approach for User Depression Detection on Sina Weibo**. arXiv, 2020. ArXiv:2008.11708 [cs]. Disponível em: <http://arxiv.org/abs/2008.11708>.

WATSON, D.; CLARK, L. A.; TELLEGEN, A. Development and validation of brief measures of positive and negative affect: the panas scales. **Journal of personality and social psychology**, American Psychological Association, v. 54, n. 6, p. 1063, 1988.

World Health Organization. **Comprehensive mental health action plan 2013–2030**. [S.l.]: World Health Organization, 2021.

WU, S.; ZHANG, H. R.; RE, C. UNDERSTANDING AND IMPROVING INFORMATION TRANSFER IN MULTI-TASK LEARNING. 2020.

XEZONAKI, D. et al. **Affective Conditioning on Hierarchical Networks applied to Depression Detection from Transcribed Clinical Interviews**. 2020.

XU, Z. et al. Individualized prediction of depressive disorder in the elderly: A multitask deep learning approach. **International Journal of Medical Informatics**, v. 132, p. 103973, dez. 2019. ISSN 13865056. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S1386505619303314>.

YADAV, S. et al. Identifying depressive symptoms from tweets: Figurative language enabled multitask learning framework. **arXiv preprint arXiv:2011.06149**, 2020.

YANG, Z. et al. **XLNet: Generalized Autoregressive Pretraining for Language Understanding**. 2020. Disponível em: <https://arxiv.org/abs/1906.08237>.

YAZDAVAR, A. H. et al. 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**. [S.l.: s.n.], 2017. p. 1191–1198.

YU, T. et al. Gradient surgery for multi-task learning. **Advances in Neural Information Processing Systems**, v. 33, p. 5824–5836, 2020.

ZIRIKLY, A. et al. Clpsych 2019 shared task: Predicting the degree of suicide risk in reddit posts. In: **Proceedings of the sixth workshop on computational linguistics and clinical psychology**. [S.l.: s.n.], 2019. p. 24–33.