See and Read: Detecting Depression Symptoms in Higher Education Students Using Multimodal Social Media Data

Paulo Mann,[1] **Aline Paes,**[1] **Elton H. Matsushima**[2]

[1]Institute of Computing, Universidade Federal Fluminense, Brazil
[2]Department of Psychology, Universidade Federal Fluminense, Brazil
{paulomann@id, alinepaes@ic}.uff.br, eh.matsushima@gmail.com

Abstract

Mental disorders such as depression and anxiety have been increasing at alarming rates in the worldwide population. Notably, the major depressive disorder has become a common problem among higher education students, aggravated, and maybe even occasioned, by the academic pressures they must face. While the reasons for this alarming situation remain unclear (although widely investigated), the student already facing this problem must receive treatment. To that, it is first necessary to screen the symptoms. The traditional way for that is relying on clinical consultations or answering questionnaires. However, nowadays, the data shared at social media is a ubiquitous source that can be used to detect the depression symptoms even when the student is not able to afford or search for professional care. Previous works have already relied on social media data to detect depression on the general population, usually focusing on either posted images or texts or relying on metadata. In this work, we focus on detecting the severity of the depression symptoms in higher education students, by comparing deep learning to feature engineering models induced from both the pictures and their captions posted on Instagram. The experimental results show that students presenting a BDI score higher or equal than 20 can be detected with 0.92 of recall and 0.69 of precision in the best case, reached by a fusion model. Our findings show the potential of large-scale depression screening, which could shed light upon students at-risk.

Introduction

Mental disorders have been alarmingly increasing in the worldwide population (WHO 2017). Individuals suffering from these problems may present a combination of abnormal thoughts, perceptions, emotions, and behavior. One of the most common mental disorder is depression, globally estimated as more than 300 million cases (WHO 2017). Particularly, Brazil has the highest prevalence of Major Depressive Disorder (MDD)[1] among South American countries, with nearly 5,8% (WHO 2017). These cases are not only valid to the general population but have also been increasingly observed in the academic environment, where students face many challenges and stressful events endorsed by academic-related situations. Reports show that graduate students are

---

[1]In this work, we use MDD and depression interchangeably.

more than six times likely to experience depression and anxiety, compared to the general population (Evans et al. 2018). Furthermore, a previous study has shown a higher prevalence of MDD in undergraduate courses, with up to 28,2% of prevalence in one of the investigated courses (de Melo Cavestro and Rocha 2006).

However, naturally, before an individual with depression receives treatment, this disorder must be detected. Many patients do not receive an earlier depression diagnosis in consultation with general practitioners, with roughly 50% of the cases detected (Kessler et al. 2002; Mitchell, Vaze, and Rao 2009); even worse, individuals might not have the money, knowledge, or they may have even fear of social stigma to look out for help (Andrade et al. 2014; Roness, Mykletun, and Dahl 2005). Because of that, the disorder may remain undiagnosed, unrecognized, and, therefore, untreated, which may further aggravate its symptoms. Thus, although the most reliable way to screen for depression is the clinical diagnosis with psychological and psychiatry doctors, it is crucial to enhance other detection options beyond the consultation-based ones that usually follows the *Diagnostic and Statistical Manual of Mental Disorders* (DSM) criteria.

Another common way of detecting MDD is relying on questionnaires, such as the *Beck's Depression Inventory* (BDI) and the *Center for Epidemiological Studies Depression Scale* (CES-D) (Beck, Steer, and Brown 1996; Radloff 1977). They evaluate the severity of depression through a final score obtained from the answers given to the questionnaire. There are at least two problems related to such methods. First, these questionnaires should also be handled by professionals, and the individual with MDD may not always have access to them. Second, these criteria have been defined years ago. As the world develops and evolves, the criteria to detect MDD should also change to go along with the new technologies that impact everyday routine and behavior.

Thus, the question that arises is if we could use regularly individual-generated data to detect depression. Notably, we want to investigate online environments such as social media, where the individual may express depression symptoms in a way different from the established DSM criteria. Several previous studies have already investigated social me-

dia features that characterize a user with depressive behavior (Shen et al. 2017; Ernala et al. 2018; Naslund et al. 2019; Jeri-Yabar et al. 2019). Related to that, there is also a great interest in using machine learning to automatically distinguish between depressive and non-depressive users using their own generated data in the social media environment, or leveraging such sites to automatically gather features inspired by the DSM and questionnaires criteria (De Choudhury et al. 2013; Tsugawa et al. 2015; Reece and Danforth 2017; Shen et al. 2017) (we expose some of them in Section §2). Screening depression symptoms from social media is related to the recently proposed concept of high-performance medicine (Topol 2019). In contrast with the traditional active diagnosis, when the individual seeks help after observing specific symptoms, the passive diagnosis systems inform individuals of possible disorders based on constant monitoring of their health, possibly through Machine Learning, for example.

The data shared by social media users, such as social networks, microblogs, and community networks consist mainly of texts and images. However, only a few recent works have focused on assessing depressive symptoms from multimodal sources of data (Morales, Scherer, and Levitan 2018). We believe that leveraging from both texts and images, which are the most common types of user-generated data, may help to distinguish different depressive groups, as depression symptoms may manifest through both verbal and nonverbal communication (Morales, Scherer, and Levitan 2017). We briefly explain multimodal learning techniques in section §3.

Thus, in this work, we gather data shared by higher education students from one of the largest Brazilian Universities in a broadly used picture-oriented with captions social media, namely Instagram. Next, we adopt such data and machine learning methods to classify the severity of depression symptoms directly from the verbal and nonverbal user-provided content. Choosing Instagram is based on the following reason: we are mainly motivated by the need of investigating the increasing number of mental disorders cases within the academic environment; accordingly, several previous works have pointed it out as one of the most trustful and used social platform by young adults (Shane-Simpson et al. 2018; Huang and Su 2018).

As ground truth, we use the results of the Portuguese translation of *Beck's Depression Inventory* (BDI) collected from an online, voluntarily answered, questionnaire[2]. Our primary research question is whether we can induce Machine Learning models from *a set of Instagram posts* that can distinguish students with moderate or severe depression symptoms from the others. Additionally, we would like to investigate if a model built from both images and texts performs better than using either only images or texts. We also want to assess whether we can achieve better results by learning the features and the classifier *directly* from the shared data with representation learning models, to avoid

the burden of inventing, engineering, and selecting specific metadata. Finally, to alleviate the negative black-box aspect of using representation learning methods, we also analyse the coefficients of a linear SVM over the induced features.

Our main contributions are as follows: (1) we create a methodology based on local search to generate a stratified oriented-to-the-individual dataset, with each example composed of a set of posts of a single individual (section *Dataset Generation*) so that our inferences do not consider only snapshots of posts but the target student instead; (2) we induce and compare the performance of several models that learn from *representation learning* (LeCun, Bengio, and Hinton 2015) techniques (section *Deep Learning Models*) and compare them with classifiers based on *metadata features* (section *Feature Engineering Models*), both from textual and visual data; (3) we propose an early fusion neural network-based architecture to handle together the textual and visual features from posts (section *Multimodal Classification*). All code is available at our GitHub repository[3].

The obtained results point out that the deep multimodal classifier reaches precision and recall values good enough to be useful in the task of screening depression using Instagram. The feature engineering models are competitive in terms of F1 score compared to the deep learning models. However, deep learning systems naturally lead to transfer the trained weights to other related domains or tasks. Furthermore, they avoid the effort of investigating and engineering the metadata to solve the task. Novel methods can provide further interpretability of black-box deep learning models.

Detecting Depression (Symptoms) from Social Media

Guntuku *et al.* survey the two main ways of assessing depression from social media, namely (1) using answers of psychological tests as attributes to fed a supervised machine learning task; (2) extracting public social media data shared by individuals that have declared themselves as suffering from depression (Guntuku et al. 2017). In the present work, we follow a hybrid approach: we rely on the BDI psychological test to obtain the class attribute, but the features come from the user-provided content. In this way, we have a more reliable class than the auto-declaration and, at the same time, more intrinsic and general features than the ones observed in the tests, aiming at fulfilling our goal: to investigate if there are underlying patterns from the user-provided content that may point out some depression tendency.

Previous works have also followed such a hybrid approach to investigate the predictive characteristics of depression reflected in the content of social media. In (De Choudhury et al. 2013), for example, tweets from individuals that answered the CES-D test were the content source. They created a binary supervised classification test according to a threshold of 22 in the value of the CES-D test. However, different from us that want to assess whether it is possible to avoid the effort of creating metadata by learning directly from the data, they rely only on feature engineering

---

to extract attributes encompassing depressive language, linguistic style, emotion words, among others. In (Tsugawa et al. 2015), the methodology was the same as the previous work but targeting Japanese individuals recruited from an advertisement posted on Twitter. A surprising aspect observed from these both studies is that the former results have pointed out that the posting time and the numbers of followers and following are crucial attributes to distinguish between depressive individuals and the others. However, in the later, this difference was not observed, suggesting that cultural aspects, or merely the observed sample of individuals, may interfere in the detected patterns of depression.

In (Shen et al. 2017), the authors focus on classifying people from the general population as depressed or not based on their tweets. The positive examples were the ones satisfying the pattern "(I'm/ I was/ I am/ I've been) diagnosed depression", or the ones that loosely mention "depress". They build the machine learning models using features extracted from the tweets, computed from the users behavior in the social media and their profile. They create a multimodal dictionary to handle the features represented by different types (numeric, vector, *etc.*). That work was later extended in (Shen et al. 2018) to transfer a model learned from one social site to another one, aiming at avoiding labeling new data. All those features are enlightening and grounded in psychological theories, but here we would like to mainly investigate how deep learning classifiers performs when trained directly from the data, avoiding the efforts invested in engineering metadata.

A similar motivation inspired the work presented in (Trotzek, Koitka, and Friedrich 2018), where convolutional neural networks are trained from linguistic metadata (gathered with *Linguistic Inquiry and Word Count* (LIWC) tool and others) and from embeddings of textual content. Several different embeddings techniques were also used in (Orabi et al. 2018) to detect depression from tweets. Different from the two later and the two previously mentioned works, we investigate the data from Instagram, which is picture-oriented, making the users express their feelings and state-of-mind using both nonverbal and verbal communication (Morales, Scherer, and Levitan 2017). We build a fusion model to consider these types of data.

Regarding the nonverbal communication, in (Reece and Danforth 2017), the authors aim at distinguishing posts of individuals with depression from the rest of the users using metadata and measures related to the published images (for example, the number of likes, number of comments, number of faces in the images, *etc.*). They investigated the color patterns of the images, based on studies pointing out that individuals with depression tend to see the world more in tones of gray. We, on the other hand, also benefit from the captions of the pictures and from visual features learned directly from the pictures.

Previous works have also demonstrated that the pattern of social media usage is different among depressed and non-depressed users on both Twitter and Facebook (Park, McDonald, and Cha 2013; Park et al. 2013). In this work, however, we assess whether this pattern exists — or not — by leveraging Machine Learning models capable of distinguishing depressed and non-depressed behavior automatically.

Some of the previous works classify the posts in social media instead of the individuals. However, they are only short-content snapshots, due to the online communication nature, and probably do not have enough information to classify depression symptoms. For us, one example in the dataset is composed of a set of posts collected during a a certain period, in this way, we make the classification robust, and less error-prone.

A Brief on Multimodal (Fusion) Learning

Multimodal learning techniques induce a model by combining more than one modality of data, such as text, images, audio, video, *etc.*, to solve applications ranging from the alignment of multiple data to classification from distinct sources (Ngiam et al. 2011). Recently, multimodal learning has increasingly gained attention due to the possibility of extracting latent features represented in a low-dimensional vector space with Deep-Representation learning (Ramachandram and Taylor 2017). Furthermore, this way of tackling data is particularly useful for the social media environment, where the users may express their feelings and thoughts using text, pictures, and even short videos (Duong, Lebret, and Aberer 2017).

To leverage those different data sources to induce a single, unified model, one can either fuse the data following a feature-based approach (early-fusion) or a decision-based approach (late-fusion) (Baltrušaitis, Ahuja, and Morency 2018). In the first case, one may extract the features for each modality separately, followed by merging the features to feed a classifier. When using Deep Learning, commonly, the feature extraction process is to collect the weights matrix of a layer in the network (Ramachandram and Taylor 2017). The other possibility, still in the feature-based approach, is to extract the features in a shared space, by jointly creating them from the multiple sources of data. In the decision-based approach, the final answer is based on the decisions taken from each modality by combining them using, for example, a voting process. The type of modality faced by Instagram data is particularly challenging as they are characterized by *meaning multiplication* (Bateman 2014): the caption and the pictures in the same post may refer to distinct contexts, but both modalities are essential to creating a new meaning that diverges from merely making a decision separately from the unimodal meanings. To tackle that, in this work, we contribute with a model that induces a classifier from concatenated textual and visual features.

Previous works have also focused on multimodal social media data sources to detect disorders, for example, the relationship between eating disorders and the removal of posts from Instagram (Chancellor, Lin, and De Choudhury 2016). Focusing on depression, the work presented in (Victor et al. 2019) considers visual and verbal communication features in their dataset. The data was produced specifically to conduct the research, and not on a regular-basis data added in social media. Here, we are particularly interested in laying the foundations of a passive diagnosis from social media instead. Audiovisual features are also combined to detect depression symptoms in (Scherer et al. 2014), using a

dataset created from dyadic interactions between an interviewer and paid participants. In (Morales, Scherer, and Levitan 2018), several fusion approaches are built from features extracted from video, audio, and transcripts. The dataset is made through interviews conducted by an animated virtual interviewer controlled by a human in another room. In this work, we also investigate the benefits of a fusion architecture, but, different from there, from data extracted from a social media.

Methods

In order to induce the machine learning models, both the proposed models that learn directly from social media data and the ones based on metadata, it is first necessary to create the datasets. In the next subsections, we describe how we perform these major tasks, namely the data collection, the dataset generation, and the induction of ML models.

Data Collection

To collect the Instagram data published by the students, we first created a Google forms questionnaire composed of (1) a number of demographic questions, such as the time spent on Facebook, Twitter, and Instagram; if they were diagnosed with depression; if they work; monthly pay income; Instagram username, *etc.*, and (2) the already mentioned psychometric test, BDI. Then, we published a call for participation in various Facebook groups, and also asked the *Universidade Federal Fluminense* (UFF) to publish the call through the official email lists. The volunteers were presented with a written explanation of the overall goals of the project, the information that would be gathered, and how their information would be used. To answer the questionnaire, they needed to be regularly enrolled in any course of the University and be at least 18 years old; to ensure data integrity, we used the transparency portal that the University provides[4] to validate the students registration number and their enrollment status. We did not have any personal contact with the students as the whole process was performed online.

We relied on BDI as a primary tool to assess the severity of the depressive symptoms in a student and to annotate the examples. BDI is a questionnaire comprised of 21 self-reported questions about the mental and psychological state of the individual, wherein each question has a score from zero to three points to determine the level of that specific symptom severity, where higher scores mean higher levels of that symptom. The final score is the sum of all the 21 questions scores. It can be interpreted as follows: 0–13, minimal; 14–19, mild; 20–28, moderate; and 29–63, severe (Gorenstein et al. 2011). We first organize the data following these four intervals of depression intensity, yielding 37% of the sample marked as severe; 23% as moderate; 14% as mild; and 26% as minimal. However, as done in previous work (De Choudhury et al. 2013; Shen et al. 2017) we separated the individuals into two classes: one comprising the students with non-intense depression symptoms (the ones scored in the minimal and mild

classes) and the other one comprising the students with intense depression symptoms (the ones scored in the moderate and severe categories). In a real-world follow-up application of our method, the individuals classified in this last case would be the ones indicated to psychological treatment.

We gathered the Instagram data that were posted *prior to the day the survey had been taken* for each student, considering three different observation periods, namely 60 days, 212 days, and 365 days. For example, if a student answered the online questionnaire on October 15, considering the observation period of 60 days, we would collect all the student data ranging between August 16 to October 15. In this way, we prevent post introduced with the sole purpose of influencing the study. We choose 60 days because it was found to be the optimum period in (Tsugawa et al. 2015), whereas 365 was investigated in (De Choudhury et al. 2013), and 212 is the mean between these two values.

Dataset Generation

Our target is the student classification, and not a single post, which is a snapshot of the student behavior in time. Thus, we formalized the problem as a *Multiple Instance Learning* task (Carbonneau et al. 2018), where the training instances are arranged in bags, and the label is provided to the entire bag. Here, the bag is the entire set of pictures or texts (or both) of each student, and the class (non-grave or grave depression symptoms) is given to the bag. In other words, the set of examples E is composed of a set of bags, *i.e.,* $E = \{S_1, S_2, \ldots, S_m\}$, where $S_i = \{post_1, post_2, \ldots, post_n\} \in E$ is the bag related to a single student i, and $post_k \in S_i$ is either (1) a tuple $post_k = (p_k, c_k)$ where $post_k$ is an individual post of the student, p_k is a picture and c_k is its caption, or (2) $post_k = p_k$, when either the post contains only a picture or when we use the examples only for image classification, or, still, (3) $post_k = c_k$, when the post is used only for text classification. Note that the size of S_i may vary from student to student since we do not oblige a maximum number of collected posts. As we still need a class for each element in the bag, we make each $post_k \in S_i$ to have the same label y_i of S_i.

To acquire the training, validation, and test sets, we must require that a bag S_i is not split into those different sets, as this would make the same student appearing in different phases of the learning and test process. It is also crucial to make the distribution of those sets to resemble the original distribution of the dataset. However, it is not trivial to attend all these conditions when considering both the number of bags *and* the size of each bag. In this way, to generate training, validation, and test sets, we implemented a local search method (Gendreau, Potvin, and others 2010) to find the optimal solution in the space of candidate solutions. We start at an initial solution with three random sets V_1, V_2 and V_3, each one containing examples $S_i \in E$ selected at random. Next, we generate the space of candidate solutions by composing : (1) half of the solutions chosen at random; (2) for the other half, we select, at random, two bags from two distinct sets, namely, $S_j \in V_w$ and $S_k \in V_p$, and switch them making $S_k \in V_w$ and $S_j \in V_p$. The evaluation function of the local search checks if these newly generated solutions are better
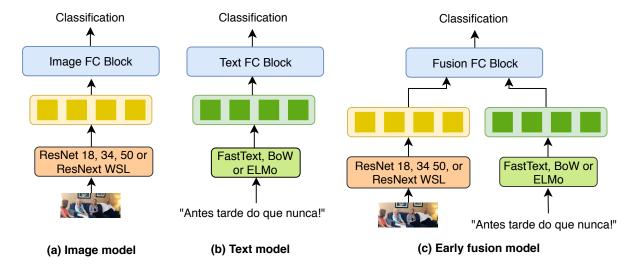
Figure 1: Deep learning architectures we have used to predict the intensity of depressive symptoms. Image, text, and fusion Fully Connected (FC) blocks are neural network classifiers designed especially for their particular modality.

than the existing ones, according to the sum of the differences between the distributions of the new solutions and the original data distribution; if the new solution has a better distribution than the previous best one, then the new solution becomes the selected one. The stop criteria is either the runtime (5 minutes), or when the newly generated solution has a very similar distribution to the original distribution for the binary BDI (low intensity: 40.27%, high intensity: 59.73%) and to the defined dataset proportion: 60% of the examples for the training set, 20% for validation, and 20% for test. After this process, we end up with ten different datasets for each observation period.

Deep Learning Models

Our central hypothesis is that we can build the depression classifiers directly from the data shared in the social media, avoiding the effort of building and investigating metadata. Furthermore, we argue that the meaning multiplication of multimodal data has more to add than relying only on unimodal data. To assess these assumptions, we first focus on classifiers that take the students' pictures and written posts separately. Then, we investigate how these two types of data cope together to make the final decision.

The Figure 1 illustrates the three types of models examined here: (a) models created from the individual images of the students (b) models created from the individual captions; (c) a fusion model that puts together the latent features extracted from the two previous types of models. As our target is the student, we combine the individual results for each post by calculating the average of all students' posts predictions to the positive class. Thus, given a student i set of posts $S_i = \{post_1, post_2, ..., post_n\}$, and their respective probabilities of being in the positive class determined by the softmax function $probas_i = \{p_1, p_2, \ldots, p_n\}$, we take the average of $probas_i$ to compute the student probability of being in the positive class.

Image classification To create the pictures classifier, we selected the ResNet (He et al. 2016) deep network as the representation learner, since it is widely used, easy to access in public frameworks, and won the ILSVRC 2015[5] competition with the ImageNet dataset. We also used the ResNeXt (Xie et al. 2017) network, pretrained with Instagram images, and fine-tuned on ImageNet1k (Mahajan et al. 2018), available at PyTorch Hub[6]. We selected this network because it was pretrained on 940 million public Instagram images, and we hypothesize that it could further help the image-based predictions. The bag associated with a single student in this case is $S_i = \{post_1, post_2, \ldots, post_n\}$ and $post_k = p_k$.

We trained four distinct-size architectures with the PyTorch framework (Paszke et al. 2017), namely ResNet-18, ResNet-34, ResNet-50, and ResNeXt-101 32x8d, all of them starting with the pretrained weights mentioned before. To extract the latent features, we partially freeze the pretrained weights (70%) and change the fully connected layer (FC) with the image FC block, which is a dropout layer ($p = 0.5$) followed by a linear layer. We induced a total of 12 image classifiers, considering the datasets created from the three observation periods (60, 212, 365), each ResNet (18, 34, 50) and ResNeXt architectures. We selected the model that reaches the best accuracy in the validation set.

We resized the pictures to 224×224 of height and width since this is the input that both ResNet and ResNeXt implementations requires. We also standardize the pictures using the original ImageNet training mean and standard deviation.

Text classification We use the classical Bag of Words (BOW), FastText (Bojanowski et al. 2017), and ELMo (Peters et al. 2018) techniques to extract the textual feature representations. BOW is computed with SciKit Learn (Pedregosa et al. 2011), FastText with the Gensim implemen-

---

tation (Řehůřek and Sojka 2010), and ELMo with the AllenNLP platform (Gardner et al. 2018). In all of these cases, the examples are the captions captured from the Instagram posts, such that $S_i = \{post_1, post_2, \ldots, post_n\}$ and $post_k = c_k$. If the $post_k$ has no caption, we use an empty string (c_k = ""). After extracting the textual features with each technique, we use a text FC block, which is a linear layer, followed by a batch normalization layer, a ReLU non-linearity, and a final linear classification layer. This architecture was chosen after achieving better convergence speed in the development set.

The Bag of Words (BOW) model works by computing a value for each distinct word in a corpus. Here, our final matrix of examples when using BOW has the dimension $\sum_{i=1}^{|E|} |S_i| \times |V|$, where $|V|$ is the vocabulary size. We used the *Term frequency-inverse document frequency* (tf-idf) metric to compute the value associated with each word within the example to balance the importance between frequent and uncommon terms.

Different from the BOW approach, word embeddings has a crucial role in deep learning techniques. To that end, Word2Vec (Mikolov et al. 2013) was one of the pioneer techniques to achieve improvements in several NLP tasks by allowing words to capture multiples degrees of meaning through their low-dimensional latent representation. However, this technique has a few limitations that the other recent ones, used in this work, does not have. First, it can not represent polysemy because of the same vector representation for the word regardless of context. Second, all embeddings are trained to an entire corpus, which means that words not seen during training are not represented at test time. Third, it does not consider hierarchical representation for words, impairing the representation of syntax and semantics aspects.

The techniques used in this work, namely, FastText and ELMo, partially or integrally solve those limitations. FastText is similar to Word2Vec, but it is robust to noisy data, as it considers subword information, which means that it can derive representations of words from morphemes, and retrieve good representations even for a small dataset (Bojanowski et al. 2017). Furthermore, it is even capable of representing some of out-of-vocabulary (OOV) words — if their morphemes are available in training time.

ELMo, on the other hand, is a Language Model (LM), different from Word2Vec and FastText. ELMo can model polysemy, subword information with character convolutions in the first layer, and hierarchical representation with two bidirectional LSTM layers on the top. The first LSTM layer usually models aspects of syntax, while the second LSTM layer retrieves aspects of contextual meaning (Peters et al. 2018). The final ELMo representation layer ($ELMo_k^{task}$) is generated by a linear combination of all these layers, which are softmax-normalized. By relying on ELMo, we allow for the implicit capture of syntax and context-dependence aspects, leaving to the model to decide which one is the most important to the task of screening depression.

Given that we were only able to collect a small dataset, we used pretrained Portuguese weights for both models: Fast-Text weights as provided by Facebook[7], and ELMo weights by AllenNLP[8], both pretrained on a dump of the Portuguese Wikipedia. Moreover, since ELMo and FastText retrieve word embeddings, we take the arithmetic mean of the word embeddings for the caption representation.

We normalize all captions by removing punctuations, emojis and hashtags. We also changed irregular entities to a specific label: we convert numbers to "0", any URL to "url", @username to "username" (since it is not a Portuguese word), and email to "email" labels. The general architecture of the text classification model can be seen in Figure 1b.

Multimodal classification To classify the severity of depression symptoms using both the pictures and captions from users' posts, we define $post_k = (p_k, c_k)$, and, as in the text classification, we use an empty string if the picture p_k has no caption. To obtain the multimodal features, we first retrieve the textual and the visual features according to the previous explained models. Inspired by the concept of meaning multiplication, where both picture and caption can create a new complex meaning, we concatenate the features from both modalities, and then we perform the final classification with the fusion FC Block, which is a dropout layer ($p = 0.5$) followed by a final linear layer. We only optimize the fusion FC block.

Feature Engineering Models

To compare our findings with baseline classifiers based on metadata, we also performed a feature engineering task from both modalities. We trained the machine learning models with the same three observation periods, and text pre-processing as used in the deep learning methods.

For textual features, we use the *Linguistic Inquiry Word Count* (LIWC) (Pennebaker, Francis, and Booth 2001) Portuguese translation (Balage Filho, Pardo, and Aluísio 2013), that was extensively investigated as useful to the task of detecting depression (Morales, Scherer, and Levitan 2018; De Choudhury et al. 2013; Resnik et al. 2015). LIWC is a text analysis program that counts words in psychologically meaningful categories (Tausczik and Pennebaker 2010). Its words categories range from, for example, linguistic style usage, as the number of used pronouns, verbs, and adverbs; and other emotional categories such as positive and negative affect words. To obtain the user-level features, we aggregate the features over all posts by taking the arithmetic mean, standard deviation, and total sum, resulting in 64 features.

As the color of images is one of the most notable features to the human eye, we extract HSV — hue, saturation and value (or brightness) — features by taking the average of the pixels in the image. Furthermore, other studies found the HSV values to be correlated with the severity of depression (Reece and Danforth 2017). We also capture the number of faces for each image using a deep-learning-based face detection model[9]. The user-level visual features are also aggregated in the same way as the textual features, resulting in 12 features.

Table 1: The ten most commonly used hashtags by different groups of BDI, from the most (top) to less frequent (bottom). *Nikiti is a nickname for the city of Niteri.

minimal	mild	moderate	severe
#art	#destinyrj	#rj	#love
#photooftheday	#womansolar	#erasmusstudent	#rj
#photography	#inktober	#uffabroads	#tbt
#tbt	#inktober2018	#eurotrip	#smile
#artsy	#tbt	#instadesign	#summer
#drawing	#photooftheday	#erasmus	#nature
#vsco	#pictureoftheday	#europe	#friends
#painting	#homesweetocean	#lisbon	#nikiti*
#artistoninstagram	#guidetoniteri	#city	#photography
#blackandwhite	#proudtobeofniteri	#life	#mumbling

Table 2: Instagram data distribution (percentage of posts) for each observation period, and for each level of depression as obtained by the BDI.

Period\BDI	Minimal	Mild	Moderate	Severe
60 days	26.62%	13.66%	18.02%	41.70%
212 days	25.43%	14.96%	16.44%	43.17%
365 days	26.05%	14.80%	15.35%	43.80%

Table 3: Mean and standard deviation of posts for each observation period considered in the study.

	Mean	Std
Posts per person (60 days)	16.73	24.67
Posts per person (212 days)	26.27	34.85
Posts per person (365 days)	37.04	46.61

To evaluate the hypothesis of meaning multiplication, we also investigate the multimodality vs. unimodality by simply concatenating the above features. Different from the deep learning models, here we already obtain user-level features by aggregating each post features values. For the classification, we used the same neural network architecture as in the text FC block.

Results

In this section, we present the experimental results obtained from the deep learning and feature engineering models, as explained before. We start by presenting the statistics related to the student sample we gathered, followed by the results considering the demographic data, and the engineered features. To that, we inspect the coefficients weights of a linear SVM model. Next, we evaluate the classifiers on the task of screening depressed individuals using text only, image only, and both types of media. The experiments were conducted on an NVIDIA DGX-1.

Data Statistics

We received a total of 416 answers between October 12 and December 2, 2018, and 2–9 April, 2019. We removed six answers that were not from currently enrolled students, and 221 students agreed to provide access to their Instagram data. Thus, we have collected these 221 students data using an Instagram scraper API[10] for Python.

Our final sample contains 136 females and 85 males with a median age of 23. For the education levels, we have 12 enrolled in Doctor's degree, 11 in Master's degree, and 198 in Bachelor's degree. For the BDI scores, we obtained a total of 82 students in the severe class, 50 in the moderate, 32 in the mild, and 57 in the minimal. We believe that the greater number of students in the severe group is because students with perceived depression might tend to participate more than their counterparts.

The Table 2 shows the distribution of posts according to each category in the BDI. As we can see, students in the severe category have almost half of the data (Instagram posts) collected for each observation period considered. We can also observe in the Table 3 the mean and standard deviations of posted pictures for each observation period.

We also investigated the most frequent hashtags that the sample of students use. As we can see in Table 1, the mild group uses hashtags that refer to the university's city (Niteri), and state (Rio de Janeiro — RJ), where the University (UFF) is placed. On the other side, the students in the moderate group — who could be considered as depressed — use more hashtags related to traveling abroad. For example, Erasmus stands for European Community Action Scheme for the Mobility of University Student[11] and is a European Union student exchange program. In this group, we also have mentions to "#eurotrip", "#lisbon", and "#europe". We found intriguing the presence of so many references for traveling abroad or going to a foreign University. They might indicate a hope of a better life in another place, different from the one they are immersed. The severe group, however, was surprising as it frequently contains hashtags related to nature, summer, smile, and love. We hypothesize that the severe group might use such hashtags as a defense mechanism to alleviate depression symptoms, using a positive thinking perspective. The minimal BDI group, unlike the moderate and severe groups, focus on photography and art in general, more similar to the mild group. However, all those hypothe-

---

[10]https://github.com/rarcega/instagram-scraper
[11]https://www.erasmusprogramme.com/

(a) Visual + textual features

(b) Visual features

(c) Textual features
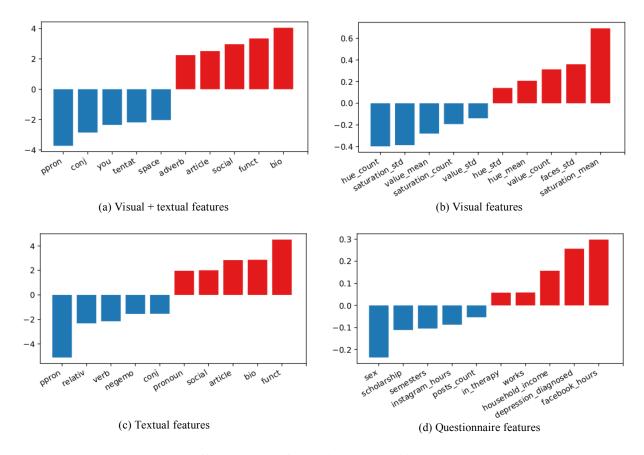

(d) Questionnaire features

Figure 2: Linear SVM's coefficient weights for predicting the positive (red) and negative (blue) classes.

ses require further investigation preferably conducted by a domain expert.

Predictive Results

We now focus on the predictive results obtained from the ML models, considering the students with the most severe symptoms as the positive class. When screening depression, it is particularly important to evaluate whether a person with high severity symptoms is incorrectly classified as possessing low severity symptoms (False Negative). Although the opposite is also important (False Positive), when screening individuals with depression, the false negative spectrum is alarming because a person with high severity symptoms, who should be detected for further treatment, is kept unknown. To that end, we choose precision, recall, and F1 metrics for model evaluation; in that way, we can have a precise measurement of how well our model is screening individuals at risk.

We perform a 10-fold cross-validation over all experiments, and report the average metrics across all the folds. We train all models with the SGD optimizer. Table 4 brings the other hyperparameters used for training. Next, we first show the most important features with the linear SVM coefficients; then, we show models' predictions results.

Analysis about the sample and elicited features To gain insights about the classification, we employ an analysis

based on linear SVM coefficients using the elicited features. We plot the top five most contributing features for the task of screening depression in Figure 2. The absolute size difference to each other can be used to determine the feature importance.

Table 4: Hyperparameters used in the learning process. *The number of MLP hidden units is always half of the input features when not used for classification.

Name	Value	Name	Value
Epochs	30	# MLP h units	$\frac{size(input)}{2}$*
Learning rate	0.001	Batch size	32
LR decay gamma	0.85	Nest. moment.	0.9
LR decay epochs	7	Optimizer	SGD

As we can see from Figure 2c, among the most important features for classifying depression, the number of pronouns, social words — about family, and friends —, and bio (biological processes: eat, blood, pain) were amongst the top five correlated features for the depressed class. On the other hand, the least depressed group was correlated with the usage of personal pronouns (ppron). Although different from previous studies that found correlated signals between personal pronouns usage and depression (Rude, Gortner, and Pennebaker 2004; Morales, Scherer, and Levitan 2018; De Choudhury et al. 2013), our sample may use lan-
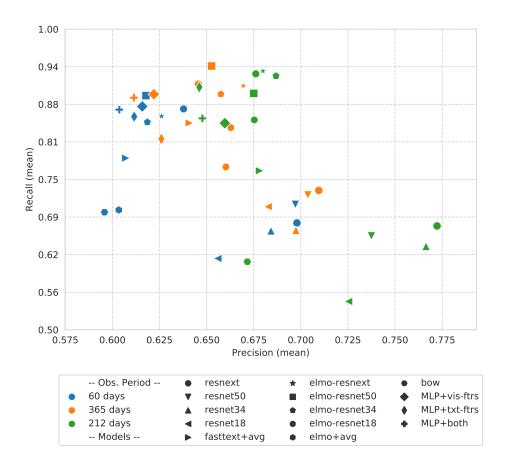
Figure 3: Predictive results of the positive class using various models with different observation periods. All results are for students predictions, not posts, over 10 different datasets.

guage differently. Particularly because in Portuguese it is not mandatory to use personal pronouns (for example, it is correct, although colloquial, to say *"going to somewhere"* instead of *"I'm going to somewhere"*). This simple example reinforces that the origin of our data may differ significantly from the previous studies, and the use of language can change across different domains.

For the visual features (Figure 2b), we found that the standard deviation of the number of faces ("faces_std"), and saturation were the most correlated features with the depressed class. We hypothesize that the standard deviation of the number of faces can be correlated with depression in the sense that more depressed people post pictures, sporadically, with a higher number of friends, but not frequently. For example, they might regularly post "selfies," or photographs of landscapes, and only a few pictures with a group of friends.

We also found that sex, and possessing a scholarship are correlated with the less depressed class (Figure 2d). On the other side, the time spent using facebook, total monthly income ("household_income"), and whether the person was diagnosed with depression are all strongly correlated with the depressed class.

Surprisingly, when putting together both visual and textual features (Figure 2a), the results are almost the same as

when using only textual features. This finding also supports previous research (Morales, Scherer, and Levitan 2018; Shen et al. 2017) that merely concatenating the values of the features do not work very well when detecting depression.

Models Predictions We exhibit the results with a scatter plot in the Figure 3. As one can see, results using an observation period of 60 days generally yields lower precision, along with higher recall scores. In this period, the model needs to give a "diagnosis" using data from 60 days only. For comparison, in a clinical setting, psychologists are encouraged to make a longitudinal evaluation, and a few sessions are not sufficient to make a final judgment, even in the presence of more evidence to support their hypotheses — like facial expressions, hand gestures, and general body language. Thus, when we train the model with an observation period of 60 days, higher recall scores suggests that the model has sufficient information to not classify a positive as a negative example comparable with higher observation periods. We expect this behavior since the BDI questionnaire asks how respondents have been feeling during the past two weeks onset of answering the questionnaire. By this means, the model supports finding individuals at higher risk as according to the BDI, even when using less data.

On the other hand, lower values of precision suggest that

the model is more susceptible to classify negatives examples as positives, which might happen due to the small number of examples for training. When we feed more data to the model, it becomes clear that there is a tendency for achieving better precision scores — keeping, or even increasing the recall. However, there is one exception: visual-oriented deep learning models tend to have higher precision scores, even when facing only 60 days of data. This might happen because Instagram is a picture-oriented social media, and it can be easier to classify examples as true negatives using image embeddings.

For the textual representations, Bag of Words performed poorly in all settings. We hypothesize that the frequency of words, although important, is not the single most relevant feature to the task of screening depression. Previous studies have pointed out the relationship between depression and syntax, or semantics (Morales, Scherer, and Levitan 2018; De Choudhury et al. 2013), where ELMo has been demonstrated to leverage these features (Peters et al. 2018). By regarding these aspects, ELMo achieves better results compared to all the textual techniques used in this study, with nearly 0.0256 of F1 improvement over the best FastText result. However, it is important to note that FastText is considerably more straightforward, and it is fast to train with few resources compared to ELMo.

Table 5: Best F1 results for each modality. All results are for the observation period of 212 days.

Model	Precision	Recall	F1	Architecture
Multimodal	0.69	0.92	0.79	ELMo+RN34
Text	0.68	0.85	0.75	ELMo
Image	0.77	0.67	0.72	ResNeXt
Feature Eng.	0.65	0.90	0.75	Txt features

Textual models usually performed better than visual models in terms of F1 score. For example, the best textual and visual models are, respectively, ELMo with 0.75, and ResNeXt with 0.72 of F1 scores for 212 days, as can be seen in Table 5. The best visual result from ResNeXt is not surprising as the pretrained weights were trained with 940 million Instagram images. However, visual models, as previously discussed, usually provided better precision scores, while textual models had higher recall scores.

For the feature engineering dataset, we had surprisingly good results. Isolated textual and visual features achieved, respectively, 0.75 and 0.73 of F1 score. This result is equivalent to their deep learning counterparts, but much more straightforward and naturally explain the classification, as we previously demonstrated with the linear SVM coefficients, which further supports the importance of syntax features for screening depression.

Considering the fused visual and textual features — for the deep learning models —, we achieve almost equivalent scores using ELMo concatenated with any ResNet, and ResNeXt architectures, where the best F1 score (0.79) was achieved with ELMo + ResNet-34. For the feature engineering dataset, however, the F1 score was not improved as expected when using fused features, resulting in a worse F1

score (0.73, for 212 days). This result can be related to the difference of features when concatenating both modalities, since we have 64 textual features and 12 visual features disposed into different representational spaces. It also indicates the necessity of more investigation on how to fuse modalities when using feature engineering, as previously explored in other studies (Morales, Scherer, and Levitan 2018). Finally, we also plot ROC and precision-recall curves in Figure 4 for a single dataset for the best results in each modality, as in Table 5.

Discussion

In general, using a deep multimodal classifier is beneficial for the task of screening depression. The feature engineering models (our baseline), on the other hand, yields competitive results when considering text or image separately; however, when using concatenated features, the results are worse. Previous studies have pointed the same direction for the screening depression task: simply concatenating engineered features makes the model focus on unimodal features instead of paying attention to both, that is why it is necessary to develop techniques for better multimodality representation, using, for example, *informed fusion* (Morales, Scherer, and Levitan 2018). Our results also support this finding, for the feature engineering models, that concatenating visual and textual features do not improve model accuracy, as previously demonstrated by the SVM coefficients in Figure 2a, relying only on textual features. One possible reason is the difference in the representational space, where we have 64 features for text, and only 12 features for images. Some alignment might be necessary in order to appropriately take advantage of both modalities in this scenario.

Instagram is a picture-oriented social media platform. Intuitively, as one might expect, detecting depression using image features should lead to improved results compared to textual features. However, our findings suggest that — with both deep learning and feature engineering — textual features perform better than using image features only. We hypothesize that this is because people express their feelings more explicitly through written texts, making the problem easier for the ML models. However, this argument needs further investigation from the psychological literature.

As we can see from the results, the feature engineering models yield competitive performance compared to the deep learning methods. However, we lose interpretability when using deep learning, which is important for trusting issues in AI-based systems. Nevertheless, deep learning naturally leads to transfer learning the trained weights, which in turn might be beneficial for detecting depression, as the acquired reliably-annotated datasets are usually quite small. Additionally, when doing feature engineering, one may find other features more relevant and change them across domains, which implicates on the need of retraining the entire model from scratch. Furthermore, social media usually implements the same paradigm: posts contain media, and media can be textual or visual. This paradigm simplifies the deployment of the same model across different social media platforms, leveraging previously acquired knowledge.
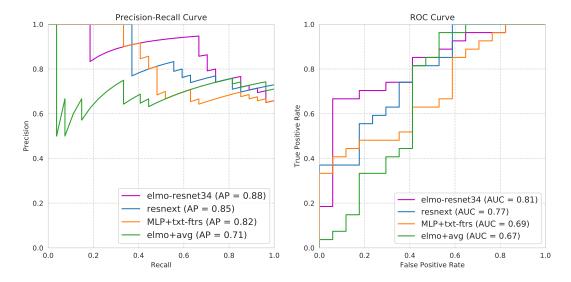
Figure 4: Precision-Recall and ROC curves for the best image classifier (ResNeXt), text classifier (ELMo) and fusion classifier (ELMo + ResNet-34) for the observation period of 212 days.

Conclusions and Future Work

The ability to distinguish between different levels of depressive symptoms from social media is a promising path for passive diagnosis of individuals at risk. To contribute in this direction, we leverage six different groups of ML architectures to distinguish students with intense depression symptoms from healthy students, relying on Instagram posts (containing both pictures and their captions). We create three deep learning models, and three feature engineering models, each based on the following media types: text-only, image-only, and the fusion of text+image. Among all the classifiers, we obtain the best predictive results with the deep multimodal classifier using ELMo and ResNet-34 concatenated features with 0.69 of precision, and 0.92 of recall scores. This finding suggests that a deep multimodal classifier is helpful in the task of screening depression using Instagram. Feature engineering-based models also achieve competitive results, with the advantage of more easily providing insights about the model prediction. Deep learning, on the other hand, allows for natural transfer learning across different domains, which may help when the sample is small.

As future directions, we first envision to investigate the possibility of transfer our learned models to evaluate students in other universities. We intend to address explainable deep multimodal learning by employing novel methods such as attention (Vaswani et al. 2017). We also expect to refine our model by interviewing the individuals and obtaining a ground truth defined by the experts. Finally, we plan to include data from other social media sites, such as Twitter and further investigate the multimodal learning possibilities.

To conclude, we believe that our contributions show a potential of help on passive diagnosis of depression, to shed light upon students at-risk and guide them to receive adequate treatment.

References

[Andrade et al. 2014] Andrade, L. H.; Alonso, J.; Mneimneh, Z.; Wells, J.; Al-Hamzawi, A.; Borges, G.; Bromet, E.; Bruffaerts, R.; De Girolamo, G.; et al. 2014. Barriers to mental health treatment: results from the who world mental health surveys. *Psychological medicine* 44(6):1303–1317.

[Balage Filho, Pardo, and Aluísio 2013] Balage Filho, P. P.; Pardo, T. A. S.; and Aluísio, S. M. 2013. An evaluation of the brazilian portuguese liwc dictionary for sentiment analysis. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*.

[Baltrušaitis, Ahuja, and Morency 2018] Baltrušaitis, T.; Ahuja, C.; and Morency, L.-P. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[Bateman 2014] Bateman, J. 2014. *Text and image: A critical introduction to the visual/verbal divide*. Routledge.

[Beck, Steer, and Brown 1996] Beck, A. T.; Steer, R. A.; and Brown, G. K. 1996. Beck depression inventory-ii. *San Antonio* 78(2):490–8.

[Bojanowski et al. 2017] Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2017. Enriching word vectors with subword information. *Trans. of the Association for Computational Linguistics* 5:135–146.

[Carbonneau et al. 2018] Carbonneau, M.-A.; Cheplygina, V.; Granger, E.; and Gagnon, G. 2018. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition* 77:329–353.

[Chancellor, Lin, and De Choudhury 2016] Chancellor, S.; Lin, Z. J.; and De Choudhury, M. 2016. This post will just get taken down: characterizing removed pro-eating disorder social media content. In *Proc. of the 2016 CHI Conference on Human Factors in Computing Systems*, 1157–1162. ACM.

[De Choudhury et al. 2013] De Choudhury, M.; Gamon, M.; Counts, S.; and Horvitz, E. 2013. Predicting depression via social media. *ICWSM* 13:1–10.

[de Melo Cavestro and Rocha 2006] de Melo Cavestro, J., and Rocha, F. L. 2006. Prevalência de depressão entre estudantes universitários. *J Bras Psiquiatr* 55(4):264–267.

[Duong, Lebret, and Aberer 2017] Duong, C. T.; Lebret, R.; and Aberer, K. 2017. Multimodal classification for analysing social media. *CoRR* abs/1708.02099.

[Ernala et al. 2018] Ernala, S. K.; Labetoulle, T.; Bane, F.; Birnbaum, M. L.; Rizvi, A. F.; Kane, J. M.; and Choudhury, M. D. 2018. Characterizing audience engagement and assessing its impact on social media disclosures of mental illnesses. In *Proc. of the 12th Int. Conf. on Web and Social Media, ICWSM*, 62–71.

[Evans et al. 2018] Evans, T. M.; Bira, L.; Gastelum, J. B.; Weiss, L. T.; and Vanderford, N. L. 2018. Evidence for a mental health crisis in graduate education. *Nature biotechnology* 36(3):283.

[Gardner et al. 2018] Gardner, M.; Grus, J.; Neumann, M.; Tafjord, O.; Dasigi, P.; Liu, N. F.; Peters, M.; Schmitz, M.; and Zettlemoyer, L. 2018. Allennlp: A deep semantic natural language processing platform. In *Proc. of Workshop for NLP Open Source Software (NLP-OSS)*, 1–6.

[Gendreau, Potvin, and others 2010] Gendreau, M.; Potvin, J.-Y.; et al. 2010. *Handbook of metaheuristics*, volume 2. Springer.

[Gorenstein et al. 2011] Gorenstein, C.; Pang, W.; Argimon, I.; and Werlang, B. 2011. Manual do inventário de depressão de beck–bdi-ii. *São Paulo: Editora Casa do Psicólogo*.

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

[He et al. 2016] He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proc. of the IEEE conference on computer vision and pattern recognition*, 770–778.

[Huang and Su 2018] Huang, Y.-T., and Su, S.-F. 2018. Motives for instagram use and topics of interest among young adults. *Future Internet* 10(8):77.

[Jeri-Yabar et al. 2019] Jeri-Yabar, A.; Sanchez-Carbonel, A.; Tito, K.; Ramirez-delCastillo, J.; Torres-Alcantara, A.; Denegri, D.; and Carreazo, Y. 2019. Association between social media use (twitter, instagram, facebook) and depressive symptoms: Are twitter users at higher risk? *International Journal of Social Psychiatry* 65(1):14–19.

[Kessler et al. 2002] Kessler, D.; Bennewith, O.; Lewis, G.; and Sharp, D. 2002. Detection of depression and anxiety in primary care: follow up study. *Bmj* 325(7371):1016–1017.

[LeCun, Bengio, and Hinton 2015] LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *Nature* 521(7553):436.

[Mahajan et al. 2018] Mahajan, D.; Girshick, R.; Ramanathan, V.; He, K.; Paluri, M.; Li, Y.; Bharambe, A.; and van der Maaten, L. 2018. Exploring the limits of weakly supervised pretraining. In *ECCV*, 181–196.

[Mikolov et al. 2013] Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

[Mitchell, Vaze, and Rao 2009] Mitchell, A. J.; Vaze, A.; and Rao, S. 2009. Clinical diagnosis of depression in primary care: a meta-analysis. *The Lancet* 374(9690):609–619.

[Morales, Scherer, and Levitan 2017] Morales, M.; Scherer, S.; and Levitan, R. 2017. A cross-modal review of indicators for depression detection systems. In *Proc. of the 4th Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, 1–12. ACL.

[Morales, Scherer, and Levitan 2018] Morales, M.; Scherer, S.; and Levitan, R. 2018. A linguistically-informed fusion approach for multimodal depression detection. In *Proc. of the 5th Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, 13–24.

[Naslund et al. 2019] Naslund, J. A.; Aschbrenner, K. A.; McHugo, G. J.; Unützer, J.; Marsch, L. A.; and Bartels, S. J. 2019. Exploring opportunities to support mental health care using social media: A survey of social media users with mental illness. *Early intervention in psychiatry* 13(3):405–413.

[Ngiam et al. 2011] Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; and Ng, A. Y. 2011. Multimodal deep learning. In *Proc. of the 28th Int. Conf. on machine learning (ICML-11)*, 689–696.

[Orabi et al. 2018] Orabi, A. H.; Buddhitha, P.; Orabi, M. H.; and Inkpen, D. 2018. Deep learning for depression detection of twitter users. In *Proc. of the 5th Workshop on Comp. Linguistics and Clinical Psychology: From Keyboard to Clinic*, 88–97.

[Park et al. 2013] Park, S.; Lee, S. W.; Kwak, J.; Cha, M.; and Jeong, B. 2013. Activities on facebook reveal the depressive state of users. *Journal of medical Internet research* 15(10):e217.

[Park, McDonald, and Cha 2013] Park, M.; McDonald, D. W.; and Cha, M. 2013. Perception differences between the depressed and non-depressed users in twitter. In *Proc. of the 7th ICWSM*.

[Paszke et al. 2017] Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch.

[Pedregosa et al. 2011] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

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

[Peters et al. 2018] Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

[Radloff 1977] Radloff, L. S. 1977. The ces-d scale: A self-report depression scale for research in the general population. *Applied psychological measurement* 1(3):385–401.

[Ramachandram and Taylor 2017] Ramachandram, D., and Taylor, G. W. 2017. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine* 34(6):96–108.

[Reece and Danforth 2017] Reece, A. G., and Danforth, C. M. 2017. Instagram photos reveal predictive markers of depression. *EPJ Data Science* 6(1):15.

[Řehůřek and Sojka 2010] Řehůřek, R., and Sojka, P. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proc. of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. ELRA.

[Resnik et al. 2015] Resnik, P.; Armstrong, W.; Claudino, L.; Nguyen, T.; Nguyen, V.-A.; and Boyd-Graber, J. 2015. Beyond lda: exploring supervised topic modeling for depression-related language in twitter. In *Proc. of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 99–107.

[Roness, Mykletun, and Dahl 2005] Roness, A.; Mykletun, A.; and Dahl, A. 2005. Help-seeking behaviour in patients with anxiety disorder and depression. *Acta Psychiatrica Scandinavica* 111(1):51–58.

[Rude, Gortner, and Pennebaker 2004] Rude, S.; Gortner, E.-M.; and Pennebaker, J. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion* 18(8):1121–1133.

[Scherer et al. 2014] Scherer, S.; Stratou, G.; Lucas, G.; Mahmoud, M.; Boberg, J.; Gratch, J.; Morency, L.-P.; et al. 2014. Automatic audiovisual behavior descriptors for psychological disorder analysis. *Image and Vision Computing* 32(10):648–658.

[Shane-Simpson et al. 2018] Shane-Simpson, C.; Manago, A.; Gaggi, N.; and Gillespie-Lynch, K. 2018. Why do college students prefer Facebook, Twitter, or Instagram? site affordances, tensions between privacy and self-expression, and implications for social capital. *Computers in Human Behavior* 86:276–288.

[Shen et al. 2017] Shen, G.; Jia, J.; Nie, L.; Feng, F.; Zhang, C.; Hu, T.; Chua, T.-S.; and Zhu, W. 2017. Depression detection via harvesting social media: A multimodal dictionary learning solution. In *Proc. of the 26th Int. Joint Conf. on Artificial Intelligence (IJCAI-17)*, 3838–3844.

[Shen et al. 2018] Shen, T.; Jia, J.; Shen, G.; Feng, F.; He, X.; Luan, H.; Tang, J.; Tiropanis, T.; Chua, T.; and Hall, W. 2018. Cross-domain depression detection via harvesting social media. In *Proc. of the 27th Int. Joint Conf. on Artificial Intelligence (IJCAI-2018)*, 1611–1617.

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

[Topol 2019] Topol, E. J. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine* 25(1):44.

[Trotzek, Koitka, and Friedrich 2018] Trotzek, M.; Koitka, S.; and Friedrich, C. M. 2018. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Trans. on Knowledge and Data Engineering*.

[Tsugawa et al. 2015] Tsugawa, S.; Kikuchi, Y.; Kishino, F.; Nakajima, K.; Itoh, Y.; and Ohsaki, H. 2015. Recognizing depression from twitter activity. In *Proc. of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 3187–3196. ACM.

[Vaswani et al. 2017] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

[Victor et al. 2019] Victor, E.; Aghajan, Z. M.; Sewart, A. R.; and Christian, R. 2019. Detecting depression using a framework combining deep multimodal neural networks with a purpose-built automated evaluation. *Psychological assessment*.

[WHO 2017] WHO. 2017. Depression and other common mental disorders: global health estimates.

[Xie et al. 2017] Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; and He, K. 2017. Aggregated residual transformations for deep neural networks. In *Proc. of the IEEE Conf. on computer vision and pattern recognition*, 1492–1500.