Chapter 6

Applications of Artificial Intelligence to Support the Diagnosis, Treatment, and Prognosis of Mental Disorders

Paulo Mann (UERJ), Elton H. Matsushima (UFF), Aline Paes (UFF)

Abstract

This theoretical short course aims to present the main challenges and trends in Artificial Intelligence applications to support the diagnosis, treatment, and prognosis of mental disorders. First, we approach the fundamental concepts of mental disorders that affect the general population the most — depressive disorders and anxiety disorders, with particular emphasis on the former. With this, we hope to provide greater visibility and knowledge to computing professionals about mental disorders' psychiatric and psychological aspects. Additionally, we will detail specific AI applications and techniques that support these disorders' diagnosis, treatment, and prognosis. Finally, we will discuss the main research challenges in this area, such as LGPD regulation and the ethical implications of automated systems that handle sensitive data. By the end of the short course, participants are expected to (i) understand the fundamentals of depressive disorders and anxiety disorders, (ii) know the main AI techniques and model architectures used by applications, (iii) understand the main methodologies of AI applications to deal with both disorders at different stages, (iv) be aware of the research trends in this area, and (v) comprehend the main ethical and legislative challenges that permeate the research and application of automated models to support the mental health field.

Resumo

O objetivo deste minicurso teórico é apresentar os principais desafios e tendências das aplicações de Inteligência Artificial para apoiar o diagnóstico, tratamento, e o prognóstico dos transtornos mentais. Em primeiro lugar, serão abordados os principais conceitos fundamentais dos transtornos mentais que mais afetam a população - os transtornos depressivos e os transtornos de ansiedade, com particular ênfase para o primeiro. Com isso, esperamos dar maior visibilidade e conhecimento para profissionais de computação

sobre aspectos psiquiátricos e psicológicos dos transtornos mentais. Ademais, iremos detalhar aplicações e técnicas específicas de IA que atuam no diagnóstico, tratamento e prognóstico destes transtornos. Por fim, trataremos dos principais desafios de pesquisa nessa área, como a regulamentação da LGPD e das implicações éticas de sistemas automatizados que lidam com dados sensíveis. Espera-se que ao final do minicurso os participantes sejam capazes de (i) conhecer os fundamentos sobre transtornos depressivos e transtornos de ansiedade, (ii) conhecer as principais técnicas e arquiteturas de modelos de IA mais utilizados pelas aplicações, (iii) conhecer as principais metodologias de aplicações de IA para lidar com ambos os transtornos em diferentes fases, (iv) conhecer as tendências de pesquisa nesta área, e (v) compreender os principais desafios éticos e legislativos que permeiam a pesquisa e aplicação de modelos automatizados para apoiar a área de saúde mental.

6.1. Introduction

Major Depressive Disorder (MDD) is the leading mental health disorder worldwide [Kupferberg et al., 2016; WHO, 2017], contributing significantly to global disability by driving dysfunctional behaviors that impair both social and professional functioning [Greenberg et al., 2021]. The socioeconomic impact of MDD is substantial, with costs in the United States alone exceeding $300 billion, primarily due to workplace-related expenses [Greenberg et al., 2021]. Despite the stability in the number of US adults receiving treatment over the past decade, the prevalence of depression has been on the rise, indicating that many individuals remain untreated and continue to suffer from depressive symptoms [Greenberg et al., 2021].

Several barriers prevent individuals from seeking treatment for MDD. These include fear of social stigma, limited knowledge about mental health, and financial constraints. While educational initiatives can address the first two issues, and improved public health services can alleviate the third, a significant number of individuals remain undiagnosed and untreated. To address this, there are a few possible solutions: (1) implementing effective screening mechanisms; (2) creating effective methods for supporting the treatment; (3) providing supportive interventions, such as educational programs and access to psychological and psychiatric services. Effective screening can identify a maximum number of individuals with depression and guide them toward appropriate assistance, while effective methods for supporting treatment will help through episodes of mental disorder. Supporting educational programs equates to providing a better prognosis both in terms of the individual and communities as a whole.

In addition to depression, anxiety disorders are another major category of mental health issues that often co-occur with depression [Aina and Susman, 2006], exacerbating the overall burden on individuals and society. Anxiety disorders, like MDD, contribute to significant functional impairment and can further complicate the course and treatment of depression. Therefore, any comprehensive mental health initiative must also account for the prevalence and impact of anxiety disorders.

One way to support the diagnosis is to automatically screen individuals with depression and anxiety to raise awareness and knowledge about these disorders. Automated screening can help with early identification and intervention, improving mental health

literacy. Mental health literacy encompasses understanding how to achieve and maintain positive mental health, recognizing mental disorders and their treatments, reducing stigma, and enhancing help-seeking efficacy [Kutcher et al., 2016]. According to the World Health Organization (WHO), health literacy, defined as "the ability to gain access to, understand, and use information in ways that promote and maintain good health," is a crucial predictor of health quality [Jorm et al., 1997].

In line with this rationale, we focus on Artificial Intelligence (AI) applications that help not only in screening individuals with depression or anxiety but also support the therapeutic and prognostic phases as well. As AI models rely on data to be trained, we give particular emphasis to social media data from Social Media Platforms (SMPs) like Twitter, Reddit, Weibo, and Instagram. This approach allows for non-intrusive screening that does not disrupt individuals' daily lives.

Furthermore, the integration of supportive interventions following the screening process is essential. These interventions include educational programs designed to increase mental health literacy, reduce stigma, and provide information on where and how to seek help. Access to psychological and psychiatric services is another critical component, ensuring that individuals identified through screening receive the necessary professional care.

With that, we expect to show applications and trends that aim to identify individuals with MDD and anxiety disorders but also strive to foster a broader understanding and acceptance of mental health issues. By leveraging social media data and advanced ML techniques, we show, through examples, that it is possible to create a robust system that addresses both the screening and support aspects of mental health care. The diverse array of applications has the potential to reach a broad audience, offering a proactive solution to the growing mental health crisis.

This chapter is organized as follows. In Section 6.2, we present the formal definition of Major Depressive Disorder and Anxiety Disorder. In Section 6.3, we demonstrate the main applications for screening mental health disorders using social media data. In Section 6.4, we explore the main ethical implications of applications that automatically screen for mental health disorders; to do that, we create an analytical framework. Finally, we conclude in Section 6.5.

6.2. Major Depressive Disorder and Anxiety Disorder

Major Depressive Disorder (MDD) is a Mental Disorder defined as a persistent feeling of sadness and loss of interest, which negatively affects how you feel, the way you think and how you act, leading to a several emotional and physical problems that disrupt your ability to function in almost every context [Association et al., 2013]. To be diagnosed with a mental disorder, the behavior should reflect a severe dysfunction in the individual's cognition, emotions, and functioning, which ultimately causes one to suffer.

Specifically, MDD has nine associated symptoms: (1) depressed mood; (2) loss of interest or pleasure; (3) significant weight loss or gain; (4) insomnia or hypersomnia; (5) psychomotor agitation or retardation; (6) fatigue or loss of energy; (7) feelings of worthlessness; (8) impaired concentration, indecisiveness; (9) recurring thoughts of death

or suicide [Association et al., 2013]. To say that an individual is sufffering from MDD, five or more of these symptoms have to be present nearly every day during a 2-week period and represent a change from previous functioning. Additionally, one of the five symptoms has to be at least the symptom (1) or (2) [Association et al., 2013]. The symptoms cause clinically significant distress or impairment in social, occupational, or other important areas of functioning and are not attributable to the physiological effects of a substance or another medical condition [Association et al., 2013]. These criteria represent a Major Depressive Episode (MDE) with varying degrees of severity, ranging from one episode to a recurrent number of episodes and remission status that provide the diagnostic for the MDD[1].

On the other hand, Anxiety Disorders[2] are a group of mental disorders characterized by significant and excessive fear or anxiety that interferes with daily activities [Association et al., 2013]. While fear is related to perceived imminent threat, anxiety is related to the anticipation of future threat [Association et al., 2013]. These disorders are marked by persistent, intense, and often irrational worry that affects how individuals feel, think, and behave. It leads to various emotional and physical symptoms that can severely disrupt their ability to function in different contexts [Association et al., 2013]. According to the DSM-5, an anxiety disorder diagnosis requires that the behavior reflects severe dysfunction in cognition, emotions, and overall functioning, causing significant distress and impairment.

Specifically, Generalized Anxiety Disorder (GAD), one of the most common anxiety disorders, has several associated symptoms: (1) excessive anxiety and worry occurring more days than not for at least six months about several events or activities; (2) difficulty controlling the worry; (3) anxiety and worry associated with three or more of the following symptoms: restlessness or feeling keyed up or on edge, being easily fatigued, difficulty concentrating or mind going blank, irritability, muscle tension, and sleep disturbance (difficulty falling or staying asleep, or restless, unsatisfying sleep) [Association et al., 2013]. To diagnose GAD, anxiety and worry should be present for more days than not for at least six months and should be about several events or activities.

Additionally, the symptoms must cause clinically significant distress or impairment in social, occupational, or other important areas of functioning. They must not be attributable to the physiological effects of a substance (e.g., a drug of abuse, a medication) or another medical condition [Association et al., 2013]. These criteria define Generalized Anxiety Disorder, but similar criteria are used for diagnosing other anxiety disorders, such as Panic Disorder, Social Anxiety Disorder, and Specific Phobias, each with its specific symptom profile and duration requirements.

Anxiety disorders are often chronic and can fluctuate in severity, with periods of increased symptoms and times of relative calm. The persistence and recurrence of anxiety can severely impact an individual's quality of life, requiring comprehensive treatment

[1]Depression is an overloaded term that is often used for short periods of distress or mourning. For this paper, we interchangeably use the terms "depression", and "MDD" to refer to the Major Depressive Disorder.

[2]We refer simply as "anxiety" to the whole group of disorders characterized as an Anxiety Disorder, such as Generalized Anxiety Disorder (GAD), Agoraphobia, Panic Disorder, Social Anxiety Disorder, Specific Phobia (the most common Anxiety Disorder in the U.S), etc.

approaches, including psychotherapy, medication, and lifestyle modifications to manage symptoms effectively.

However, the symptoms of Anxiety and Depressive disorders may vary under different sociocultural norms, as one behavior is acceptable or encouraged by one society. Albeit there are well-defined criteria for diagnosing a mental disorder, the spectrum of sociocultural norms should always be considered for a proper diagnosis. For example, a sample of Japanese individuals who are not depressed might behave similarly to a sample of depressed Brazilian individuals. As culture evolves, clinicians should be prepared to differentiate normative behavior from an impairing symptom criterion for diagnosis. Significantly, the cultural etiquette formed in online social media is an organism that evolves even faster.

Moreover, screening depression and anxiety using social media cues becomes challenging as online language and identity are fluid over time. Social media constantly adopts new memes, terms, inside jokes, and new hashtags that provide a new way to interact, but at the same time, it is also a social space used by users with a mental disorder — or not. Individuals with depression and anxiety disorders use tools that do change over time — and frequently — which inevitably will make the manifestation of their symptoms take different forms: a new hashtag, a newly coined term; although different, it is still a manifestation of a depressive or an anxiety symptom.

Accordingly, it has become increasingly difficult for any human to keep up with the constant flow of new data, especially with social media; this is no different for clinicians. To cope with this fast pace and vast amounts of data, machine learning-aided diagnosis, usually referred to as high-performance medicine [Topol, 2019], is a tool that can help healthcare professionals in the diagnosis. In this way, automated methods can help analyze from a single piece of information, to the general holistic view. Thus, ML trained models on a specific sample can be swiftly adapted to new individuals and new social media terms.

Finally, the traditional way to screen depressed individuals is through a clinical interview. However, this is often costly and error-prone, and it requires an active role of the individual to look for help. To help general practitioners, psychometric tests can provide a second opinion on the intensity of depressive or anxiety symptoms. There are two most widely used psychometric tests in the literature for identifying the intensity of depressive symptoms: Beck's Depression Inventory (BDI) [Beck et al., 1996] and the Center for Epidemiologic Studies Depression Scale (CES-D) [Radloff, 1977].

The Beck Depression Inventory-II (BDI-II) is a self-report questionnaire that measures the severity of depression by evaluating all of its symptoms. It has 21 items responded in a four-alternative scale; each answer yields a score between 0–3, which indicates the severity of the symptom the question is evaluating. The final BDI-II score is the sum of all 21 questions answers, which might yield a score in the range 0–63. The final score is classified into four distinct categories: 0–13 is minimal; 14–19 is mild; 20–28 is moderate; and 29–63 is severe depression [Gorenstein et al., 2011].

The Generalized Anxiety Disorder 7-item (GAD-7) scale is a self-report questionnaire designed to measure the severity of generalized anxiety disorder (GAD) by as-

sessing its symptoms over the past 2 weeks. It consists of 7 items, each describing a symptom associated with GAD. Respondents rate how often they have been bothered by each symptom over the past two weeks on a 4-point scale, where 0 means "not at all," 1 means "several days," 2 means "more than half the days," and 3 means "nearly every day." The total GAD-7 score is calculated by summing the scores of all 7 items, resulting in a range of 0 to 21. The final score is classified into four distinct categories to indicate the severity of anxiety, where 0–4 is minimal anxiety, 5–9 is mild anxiety, 10–14 is moderate anxiety, and 15–21 is severe anxiety.

The GAD-7 scale is widely used in both clinical and research settings due to its brevity and effectiveness in screening for generalized anxiety disorder. It helps healthcare providers quickly identify the level of anxiety and make informed decisions about further evaluation and treatment options. The scale is also helpful in monitoring changes in anxiety symptoms over time, making it a valuable tool for both initial assessment and ongoing management of anxiety disorders.

6.3. Artificial Intelligence for Mental Disorders

In this section, we present a systematic literature review to summarize the main applications of AI for mental health disorders. The work we present here has been collected from the following digital libraries: Web of Science, IEEE, ACM Digital Library, and SCOPUS. When applicable, if the search yielded more than 300 studies, we get the 300 most relevant works. We sort the result by the number of citations, the relevance of the publishing medium, and the publication date — recent publications with higher priority.

Finally, the study selection process occurs as follows: (1) we first submit the search string to the mentioned repositories, which yielded 635 items across all repositories, including publications from journals and conferences. We saved and managed the references using the Mendeley Reference Manager[3] tool; (2) next, the authors read all paper's titles, keywords, and abstracts, removing the studies that met at least one of the exclusion criteria. This phase resulted in 28 articles; (3) next, the author read the introduction and conclusion of all remaining papers and removed those that meet the exclusion criteria, thus resulting in 15 articles, of which 4 are surveys. The remaining papers were completely read and did not fit any exclusion criteria.

Furthermore, we also applied the snowballing technique [Jalali and Wohlin, 2012] to find relevant references of a study (backward snowballing) or relevant works that mention the final selection of 23 articles (forward snowballing). We do not include the four surveys [Skaik and Inkpen, 2020; Ríssola et al., 2021; Dhelim et al., 2023; Mathur et al., 2023] in the list of selected works, demonstrated in Table 6.1; they are, instead, used as seeds for the forward and backward snowballing method, and as a reference to structure this section.

This section is organized as follows: we begin with Section 6.3.1 by exploring the aspects of data, such as which social media the data was collected and how digital data footprints can leave signs of mental health expressions. In Section 6.3.2, we explore different methods for obtaining data, such as asking users for permission — explicitly or

[3]http://mendeley.com

implicitly — to collect their data. In Section 6.3.3, we show different methods for extracting the features from different modalities of data and the main methods for classification.

6.3.1. Data

In the broad field of psychotherapy, understanding the nuances of how different methodologies process various input data can illuminate their practical applications in a therapeutic context. Different approaches to psychotherapy process the input data differently — for input data, consider the voice tone, for example. The Cognitive Behavioral Therapy (CBT) often relies on the theory of behavior and cognition to find and intervene in thought processes, behavior, and habits that originate dysfunctional ways of living [Rice, 2015]. For example, journaling is one of the cognitive techniques used in CBT. Taking notes in a stressful situation might reveal reinforcement mechanisms of dysfunctional outcomes [Rice, 2015], which helps both the individual suffering and the therapist with enhanced information. Cognitive restructuring is another technique to help individuals identify, evaluate and modify the faulty thoughts responsible for their psychological disturbance [Clark, 2013]. In that sense, the analyst collaborates with the patient to improve their quality of life. The input received by the analyst is processed in methodological ways to produce the desired outcome.

Psychoanalysis, on the other hand, was initially focused on the unconscious and free association. The psychoanalyst encourages patients to talk freely about anything. In the famous case of obsessional neurosis as written by Sigmund Freud, the "Rat man" said "[...] I used to have a morbid idea that my parents knew my thoughts [...] There were certain people, girls, who pleased me very much, and I had a very strong wish to see them naked. But wishing this I had an uncanny feeling, as though something must happen if I thought such things [...] that my father might die." [Freud, 1909]. After conducting several sessions, Freud concluded that the "Rat Man", as a child, was scolded by his father for some situation related to masturbation. Since then, he has created not only trauma from masturbation but also a terrible grudge against his father, which has motivated his thoughts of death against his father. Freud freely associated the "Rat Man"'s discourse with his creativity and the theory of unconsciousness to conclude that certain events in his childhood led to developing his obsession with his father's death. Although one might argue that this methodology seems vague and opaque, it illustrates the inherently different methods of different psychotherapies and how they process input data.

Not too distant, nonetheless, is the data generated through social media. While the "environment" and objective of talk therapy are inherently different from social media, both share a common aspect: behavior footprint. While in talk therapy, patients, intentionally or not, let several clues about their mental state through the spoken content or facial expressions and posture. This behavior footprint is the raw data for therapy. While on social media, users leave digital footprints behind. Although there is room for fictional online identities, a body of evidence shows how the language and behavior of individuals with mental health issues using social media differ from control groups [De Choudhury et al., 2013; Nguyen et al., 2014; Pan et al., 2020; Chancellor and De Choudhury, 2020; Kelley and Gillan, 2022]. Furthermore, social media is available most of the time, users can post virtually any time. On the other hand, the interaction with the health care system, be it for a general practitioner or a regular psychotherapist, is usually infeasible or not fre-

Table 6.1. List of related works and their main characteristics. The features column shows techniques for extracting features from textual, visual, or metadata content, FE: feature engineering. The user-level column refers to studies that train models for user-level classification by aggregating the post-level features in any way (be it an LSTM or simply taking the average of the post embeddings). "Data" shows the dataset's source, be it manually collected by the researchers from Twitter, Reddit, Instagram, Facebook, Weibo, or Dcard, or using a dataset created for a shared task, such as eRisk [Parapar et al., 2021], CLPsych [Coppersmith et al., 2015] and SMHD [Cohan et al., 2018]. The field column is related to a mental disorder or a specific symptom of a mental disorder. MDD: Major Depressive Disorder, AN: Anxiety, SD: Suicide Ideation; "Mental health" refers to general mental health expressions, whether the user disclosed suffering from mental health symptoms.

Reference	Features	Classifier	Label	User-level	Post-level	Data	Modality	Field	Data Gather Method
[Tsugawa et al., 2015]	FE, LDA, BoW	SVM	CES-D	Yes	No	Twitter	Text	MDD	Explicit
[Bagroy et al., 2017]	N-grams	LR	Inductive Transfer Learning	No	Yes	Reddit	Text	Mental Health	Implicit
[Shen et al., 2018]	FE	NB, MSNL, WDL, MDL	Self-report	Yes	No	Twitter	Various	MDD	Implicit
[Reece and Danforth, 2017]	FE	RF	CES-D	No	No	Instagram	Image	MDD	Explicit
[Cheng et al., 2017]	SC-LIWC	SVM	SPS, DASS-21	Yes	No	Weibo	Text	MDD, AN, SD	Explicit
[Orabi et al., 2018a]	word2vec	Bi-LSTM, CNN	Self-report	Yes	No	CLPsych 2015	Text	MDD	Implicit
[Coppersmith et al., 2018]	GloVe	LSTM	Self-report	No	Yes	Twitter	Text	SD	Mixed
[Ricard et al., 2018]	FE	LR	PHQ-8	Yes	No	Instagram	Text	MDD	Explicit
[Trotzek et al., 2018a]	FE, GloVe, fastText	CNN, LR	Self-report	Yes	Yes	eRisk 2017	Text	MDD	Implicit
[Wongkoblap et al., 2018]	FE	SVM, LR, DT, NB	CES-D	Yes	No	Facebook	Text	MDD	Explicit
[Aragón et al., 2019]	BoSE	SVM	Self-report	Yes	No	eRisk 2017, eRisk 2018	Text	MDD	Implicit
[Liu et al., 2019]	N-grams, FE	SVM, DT, RF, LR	Expert	No	Yes	Weibo	Text	SD	Implicit
[dos Santos et al., 2020]	LIWC, TF-IDF, word2vec	MLP, LR	Self-report	Yes	No	Twitter	Text	MDD	Implicit
[Fu et al., 2021]	BERT, FE	MLP	Expert, Knowledge Graph	No	Yes	Weibo	Text	SD	Implicit
[de Souza et al., 2022]	GloVe, Word2Vec	LSTM, CNN	Self-report	Yes	No	SMHD	Text	MDD, AN	Implicit
[Cha et al., 2022]	word2vec, BERT	Bi-LSTM, CNN, MLP	Lexicon	No	Yes	Twitter	Text	MDD	Implicit
[Mukta et al., 2022]	MPNet	KNN, RF, AdaBoost, LGBM	MDI, PSS, TIPI, GSE	Yes	No	Facebook	Text	MDD	Explicit
[Bucur et al., 2023]	CLIP, EmoBERTa	Transformer, LSTM, GRU	Self-report	Yes	No	Reddit, Twitter	Text, Image	MDD	Implicit
[Wu et al., 2023]	BERT	LSTM	Expert	No	Yes	Dcard, open data	Text	SD	Implicit

quent enough for most individuals — e.g., either because of professional unavailability or the associated cost.

Individuals use Social Media Platforms (SMP) for several reasons: to connect with distant relatives, to meet new people, to engage in sociopolitical activities, to talk to their friends, or to benefit from social relationships. Benefitting from social relationships can be seen as generating what is called social capital [Pan et al., 2020]. Although this concept is general enough for any social interaction, be it digital or not, the increasing popularity of SMPs, their convenience, and the paucity of time for real-life interactions make SMPs a natural gateway for social interactions in the digital world, especially for young adults. Although users might participate in SMP for several motivations, there is a crucial reason for screening mental health on social media: contributing an original post and responding to posts by other users, which might generate social capital [Pan et al., 2020]. Analyzing the relationship between social capital gains — or lack thereof — with mental health expressions can offer a unique opportunity to help individuals. Although users are not always looking for social support in social media, the word choices, the topic they are discussing, which posts they interact with, or what users they are friends with — all reveal e behavioral footprint that will inevitably depend on the structure and features provided by the specific social media platform. Next, we will discuss how different applications capitalize on the digital footprint.

6.3.2. Data Collection

How do researchers obtain the data? In order to obtain data, they either (1) asked volunteers to participate through formal questionnaires and with explicit terms and conditions to guarantee participant's privacy and knowledge; (2) scraped publicly available content without asking for the owner's direct permission; (3) or used one dataset already collected using method 1 or 2, generally known as datasets for shared tasks, such as the eRisk, CLPsych, and SMHD datasets in Table 6.1.

For the first method, the study generally applied for the local Institute Review Board (IRB) and often stored the data anonymously and securely. We refer to this method as explicitly asking for the participant's consent — as demonstrated in the Data Gather Method column in Table 6.1. Previous research asked for participants' permission to scrape their social media content and answer psychometric tests. However, even when users thoroughly answered the psychometric test, several participants refused to share their social media data. For example, in Reece et al. (2017) [Reece and Danforth, 2017], 43% of individuals who completed the survey refused to share the Instagram data. In Wongkoblap et al. (2018) [Wongkoblap et al., 2018], for one dataset, only 18% provided access to their Facebook data (931 individuals). Perhaps, a consequence of survey questionnaires is the demonstration that privacy often matters for individuals, and they have more concerns about their data being analyzed, especially for mental health footprint.

Informed consent is often the desirable method for ethical guidelines in research, and a few studies compensated the participants with monetary payment to boost response rate [Reece and Danforth, 2017; Cheng et al., 2017; Ricard et al., 2018]. It also has the benefit of asking individuals to answer psychometric tests to serve as a gold label and to obtain sociodemographic statistics. However, asking for informed consent also has

downsides, such as selection bias in online surveys, when a particular sociodemographic class is more prone to answer and help than others. For example, in Wongkoblap et al. (2018) [Wongkoblap et al., 2018] 59% of participants are female. On the other hand, in Tsugawa et al. (2015) [Tsugawa et al., 2015], for a sample of Japanese individuals, 58% are male. While sociodemographic factors may influence, there is also the possibility of bias related to the communication channels used to disseminate the research invitation. Additionally, as the online survey invitation often already reveals the research focus, i.e., mental health, individuals who believe they are experiencing mental health issues could be more inclined to participate.

For this last issue, consider the prevalence of depression among the general population: 5,8% for Brazil and 5,9% for the US [WHO, 2017]. Next, consider the prevalence of depression for each explicitly collected dataset. The Instagram dataset [Mann et al., 2020] contains 60% of depressed individuals as measured by the BDI-II; in Wongkoblap et al. (2018) [Wongkoblap et al., 2018], they reported 76% of depressed individuals as measured by Center for Epidemiologic Studies Depression Scale (CES-D); Tsugawa et al. (2015) [Tsugawa et al., 2015] reported 39% of depressed individuals as measured by CES-D; Reece et al. (2017) [Reece and Danforth, 2017] reported 43% also measured by CES-D. Note the difference between the prevalence of depression in each sample and the prevalence for the general population.

Furthermore, in all the research mentioned earlier, participants are often young adults, which explicitly excludes a big part of society — such as older individuals or individuals without access to the internet and SMPs. Machine learning models trained on biased data often underperform when faced with a data distribution not seen during training (more on this subject in Section 6.4). The high amount of young adults with depression in all those datasets demonstrates the difficulty of accessing a balanced sample according to sociodemographic markers.

The most preferred method for gathering data is scraping the data publicly without explicit consent from owners. The consent is implicit because SMP users must agree to the social media data policy and terms of service, which usually inform users that the SMP has broad rights to use and distribute the data that users create and share. Collecting data implicitly is more straightforward, less time-intensive, and cheaper. Thus, 64% of the related works relied on data obtained implicitly, as demonstrated in Table 6.1.

However, data obtained implicitly often lack several benefits from explicitly asking permission. First, social media users are often anonymous, so it is hard to scrape sociodemographic statistics. Second, to label posts (or users) for mental health disorders, researchers usually rely on posts containing the self-report of a specific mental disorder. For example, users who wrote posts that match the pattern "(I'm/ I was/ I am/ I've been) diagnosed depression", or the ones that loosely mention "depress", or even through regular expressions, would be labeled as "depressed", among other mental health disorders or symptoms [Shen et al., 2018; Trotzek et al., 2018a; Coppersmith et al., 2018; Orabi et al., 2018a; Aragón et al., 2019; dos Santos et al., 2020]. There is also a concern about the validity of the collected data and whether the self-report text is accurate, for example.

On the other hand, the benefit of scraping data without asking for explicit permission is the possibility of scraping much more data in a large-scale scenario. When

gathering public data, researchers can collect up to 152,834 unique Reddit users with 446,897 posts [Bagroy et al., 2017], or 36,993 depression-candidate Twitter users with over 35 million tweets [Shen et al., 2018]. In comparison, asking for permission yields a dataset with no more than 1000 individuals [Tsugawa et al., 2015; Reece and Danforth, 2017; Cheng et al., 2017; Coppersmith et al., 2018; Ricard et al., 2018; Wongkoblap et al., 2018; Mukta et al., 2022], where studies with bigger datasets, in general, provided monetary compensation for participation [Cheng et al., 2017; Ricard et al., 2018]. Relying on individuals' self-reporting rather than using a psychometric instrument comes with the benefit of a larger dataset. With modern deep learning techniques relying less and less on specific architectural innovations and more on large datasets, especially with relaxed inductive biases on Transformer architectures, gathering more data could lead to improved prediction scores [Bucur et al., 2023]. Considering the listed related works in Table 6.1, there is still work to be done to understand the impacts of different types of datasets both in performance and ethical guidelines.

The third method to obtain data is to ask for other researchers who either collected explicitly or implicitly or to use shared tasks datasets, often distributed for competition purposes. There are two widely used datasets in the literature: CLPsych and eRisk.

6.3.3. Feature Extraction and Classifiers

In this section, we will explore two essential aspects of automatic classification: (1) extracting features from multiple modalities of data; (2) and the common classification methods used by the related works.

6.3.3.1. Extracting Features

There is also a significant concern about how to generate the model's input. Several studies in the literature use textual features to detect mental health disorders, mainly based on the assumption that there are psychological traits in the text produced by individuals [Pennebaker et al., 2003]. Among them, using the taxonomy proposed by Dhelim et al. (2023) survey [Dhelim et al., 2023], there are three high-level categories of features: textual features, multimedia features, and behavioral features.

Textual Features

There are three main categories of textual features: linguistic, sentiment, and ideograms. One of the most common choices of features is obtaining the psychological categories of words. The Linguistic Inquiry Word Choice (LIWC) is one of the most widely used psychological dictionaries (under the linguistic category in the taxonomy). Previous studies used LIWC because of its simplicity, ease of use, and the expressivity of the features, such as obtaining meaningful categories for words: positive and negative emotions, anger, and personal pronoun usage, among others. Several studies use it as a baseline method [Cheng et al., 2017; Liu et al., 2019; Mukta et al., 2022; dos Santos et al., 2020; Shen et al., 2018; Trotzek et al., 2018a; Wongkoblap et al., 2018]. There are other psychological dictionaries, such as empath [Fast et al., 2016], a psychological dictionary based on deep learning techniques that produce categories based on a set of seed words, which one of the related works also use [Mukta et al., 2022].

On the other hand, topic modeling is used to compute a set of latent topics that indicates the general discussion or directions — or simply topics — from a set of textual contents. The set of textual content could be from one social media user or the entire dataset. Researchers apply supervised or unsupervised topic modeling in the entire dataset to understand the general topic of discussion among depressed individuals against control groups and use it as a feature vector to feed classification models [Tsugawa et al., 2015; Shen et al., 2018].

Another commonly used linguistic feature, especially for baseline methods, is the Bag of Words (BoW), sometimes used with the TF-IDF weighing scheme [Ricard et al., 2018; dos Santos et al., 2020]. It is a simple method based on the count of the frequency of words. The product of this technique is a matrix where the rows represent posts (in post-level classification) or users (in user-level classification), and the columns represent the vocabulary. However, the main issue with this representation is that the order of words in a post is not represented, and the temporal dimension of mental disorders will not be modeled as part of the problem. N-grams is a technique that considers the probability of occurrence of a word (w_i) given a history of words preceding the word to counterbalance this issue. For simplicity, the entire history is often not considered but only a small window; hence, that is why N-gram. If $N = 2$, we simplify the problem of determining the probability of occurrence of a word given its entire history by the probability of a word given its preceding word, also called the Markov assumption. The product of this process is often a matrix where the columns are the N-grams, and the rows are the posts (post-level classification) or users (user-level classification), where each cell contains the frequency or the likelihood ratio of the N-gram.

Unlike BoW and the N-grams approach, word embeddings are crucial in deep learning techniques. To that end, word2Vec [Mikolov et al., 2013] was one of the pioneer techniques to improve several NLP tasks by allowing words to capture multiple degrees of meaning through their low-dimensional latent representation. However, this technique has a few limitations that the other recent ones do 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. A few works use either word2vec, or similar variants with similar limitations [Orabi et al., 2018a; Coppersmith et al., 2018; Trotzek et al., 2018a; dos Santos et al., 2020; Cha et al., 2022].

Recent word embedding techniques were improved to represent polysemous words, richer representations with improved methods, and a more extensive training corpus with a better tokenization strategy. Among several variations, BERT, CLIP, and RoBERTa are often used for representing words, or sentences, as vectors [Fu et al., 2021; Cha et al., 2022; Wu et al., 2023; Bucur et al., 2023].

Sentiments are primarily obtained from LIWC but can also be obtained through sentiment analysis algorithms. A few works rely on specific emotion lexicons, such as the Affective Norms for English Words (ANEW), VADER, NRC, and the Opinion Lexicon [Trotzek et al., 2018a]. Other works elaborated their lexicon based on experts [Cha

et al., 2022]. One naive approach is to consider posts with words such as “pain”, “depression”, and “unhappy” from individuals with possible signs of depression. Although this is simple enough for text matching, it is difficult to know whether the word, in context, does not have an entirely different meaning, such as irony.

Regarding ideograms, a few works removed emoji from the text as they were considered “incompatible with many text processing algorithms” [Shen et al., 2018; Fu et al., 2021]; other works included an emoji sentiment scale to map emojis to a happiness score [Ricard et al., 2018]. However, several works need to describe what they do with emojis: adding, modifying, or removing them entirely. One of the exciting things about the recent — contextual — word embedding techniques, such as BERT, is that they support the representations of emojis without relying on tricks to overcome the limitations of the previous static word embeddings, such as word2vec.

Multimedia Features

A few works relied on images for multimedia features, especially given the rise of photo-oriented platforms in recent years. They either manually extracted features, such as the hue, value, brightness, and number of faces [Reece and Danforth, 2017; Shen et al., 2018], or they relied on visual representation learning techniques [Bucur et al., 2023]. It has been demonstrated that using more than one modality improves the performance scores in several experiments [Shen et al., 2018; Mann et al., 2020, 2022; Bucur et al., 2023]. The type of modality faced by multimodal data is particularly challenging as they are characterized by meaning multiplication [Bateman, 2014]: the textual and visual contents 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. However, it still needs to be determined how much each modality contributes to the performance and exactly how. None of the selected related works investigate audio data. Convolutional Neural Network (CNN) and CLIP is the preferable method for extracting visual representations [Bucur et al., 2023]. Although a few works use CNN, they use it for sentence classification instead of obtaining visual features.

Furthermore, manually extracted features bring awareness to the bias problem, where the choice of features relies on the researcher’s knowledge and how their culture perceives the mental disorder. The way society perceives depression in one person is often dependent on the subjectivity in culture and environment, which ultimately is reflected upon the diagnosis criteria of psychiatric manuals [Association et al., 2013]. For example, while the posting time is a determinant for classifying depression in a sample of American individuals, it is not for Japanese individuals [Tsugawa et al., 2015]. By leveraging Representation Learning and Deep Learning [Bengio et al., 2013], the model automatically learns feature representations according to the task using generic priors, eliminating the need to extract features manually. Additionally, we can use the learned representations to transfer the knowledge with transfer learning [Pan and Yang, 2009] to other domain-related problems; such a procedure is not straightforward to replicate for handcrafted features.

Behavioral Features

Social Media Platforms offer various options for users to connect or interact with

the platform itself or other users. For behavioral features, there are several categories of behavior. For example, online activities involve interaction with other posts and users, such as liking, sharing, or commenting on a post, the number of followers, or how many users they follow. Another behavioral feature is associated with communities, where individuals can affiliate with subreddits (Reddit) or Facebook groups. The information on which groups and the kind of relationship the user has with the group is another rich source of behavioral information.

Several related works rely on behavioral features to feed their classification models [Tsugawa et al., 2015; Shen et al., 2018; Reece and Danforth, 2017; Ricard et al., 2018; Wongkoblap et al., 2018; Fu et al., 2021]. One possible reason for not relying on such data is that the behavioral activity is inherently attached to the SMP, and it is hardly helpful for transferring the learned knowledge to another SMP — because of the inevitably different characteristics or structure of the SMP. Another reason is to effectively experiment with the classification only on the textual content created by individuals without relying on any external behavioral trace. If it is possible to distinguish depressed individuals from control based only on their produced textual content, then we expect that any added behavioral footprint will improve the performance scores.

More specifically, Ricard et al. (2018) [Ricard et al., 2018] was the only study to compare user-generated content with community-generated content. Their experiments show evidence that community-generated content does indeed help improve performance scores as opposed to only using linguistic features.

6.3.3.2. Classification

The extracted features are used to feed classification models. The classification method will automatically classify different entities based on how features are computed: a user, a post, or even a community. All related works deal with supervised learning approaches and label their dataset according to Section 6.3.2.

It is important to note that if they labeled the data using a psychometric test, the obtained score naturally results in a user-level labeling process [Tsugawa et al., 2015; Cheng et al., 2017; Ricard et al., 2018; Mukta et al., 2022]. As the score obtained is usually an integer number, the related works often consider a threshold to split users into two classes: depressed and non-depressed — or other mental health disorders. For example, the BDI splits the score into four categories of intensity of depressive symptoms: 0–13 for minimal intensity, 14–19 for mild intensity, 20–28 for moderate intensity, and 29–63 for severe intensity. Furthermore, in the psychiatric literature, moderate and severe categories are related to a depressed individual [Gorenstein et al., 2011]. It is similar to other psychometric tests — such as CES-D —, where they establish optimal cutoffs to distinguish depressed from non-depressed individuals.

On the other hand, several studies relied on individuals self-reporting depression (or other mental health conditions) to separate into two groups — one with the mental health disorder and the control group [Shen et al., 2018; Orabi et al., 2018a; Trotzek et al., 2018a; Aragón et al., 2019; dos Santos et al., 2020; Bucur et al., 2023]. When considering self-reporting the mental health condition, researchers use the premise that

the individual is suffering from that condition and label them as of the positive group. For the negative class, researchers often scrape data from random users across the SMP or collect data from forums that discuss general topics, such as subreddits about movies, food, or news [Orabi et al., 2018a; Trotzek et al., 2018a; Aragón et al., 2019].

One exciting approach to gathering control group data is to collect data from individuals who talk about the mental health disorder but do not suffer from any mental health disorder [dos Santos et al., 2020]. Using posts from the "depression" subreddit as the positive class and posts from subreddits such as news, movies, or food as the negative class might induce the model to learn simple correlations that result in high metric scores. As such, the model could learn to correlate the occurrence of words such as "depressed" and "unhappy" with depression simply because those words are more likely to occur in the depression subreddit. Collecting data from individuals who shall use the same — prominent — words that depressed individuals use but are not suffering from depression makes the task more challenging for classification models — and more akin to the real world.

Classifying the user directly (user-level) can attribute the mental disorder directly to the user. However, one issue is relying on aggregating tricks to obtain the user-level feature vector. When doing feature engineering, it is straightforward to aggregate data; for example, when using n-grams or BoW, it is enough to sum the frequencies across all posts of a single user to obtain the frequency vector for a single user. Additionally, when using LIWC, the researcher only needs to aggregate the frequency for each category for each post, resulting in a user-level feature vector. Aggregating manually engineered features across time might result in a loss of information for classification. Several related works used aggregated engineered features data to train user-level classifiers [Tsugawa et al., 2015; Shen et al., 2018; Cheng et al., 2017; Ricard et al., 2018; Trotzek et al., 2018a; Aragón et al., 2019; dos Santos et al., 2020].

However, aggregating features is less effective for low-dimensional neural distributed representations, such as word2vec or BERT. Previous works have presented a theoretical and practical framework for constructing a Multiple Instance Learning (MIL) methodology [Mann et al., 2022]. One study implements a user-level classification that leverages the MIL paradigm [Bucur et al., 2023] — i.e., without aggregation tricks. We note, however, that none of the mentioned works gives formal MIL specifications for their tasks.

In the context of post-level classification, it might seem too granular to classify a single post. However, there are a few applications wherein such granular classification is desirable, such as identifying a post written by a person at risk of suicide [Coppersmith et al., 2018]. In such circumstances, using the latest post written by the individual to measure suicide risk is necessary. However, depression is a condition whose diagnosis criteria require two weeks to observe symptoms. Consequently, a post-level classification will inevitably lose the temporal component when analyzing mental health conditions.

When examining user-level classification, there remain valid concerns regarding the model's training process over the user-level feature vector. Specifically, the model learns to establish a linear or otherwise correlation with the class based on the aggregated feature-engineered vectors (user-level feature vector). However, it is worth noting

that this feature vector remains static — it is not learned during training, just statically computed before it. As a result, it does not consider local temporal variations, such as the change in usage of emotional words from one week to another. Even if we let the model adjust the user-level feature representation during training, it will not compute the user-level representation based on the post-level representations. In contrast, employing the Multiple Instance Learning (MIL) methodology [Mann et al., 2022] allows the model to ascertain the best user-level representation during the training phase, provided that the feature extractor can be fine-tuned. The learning procedure dynamically determines the most effective user-level feature vector based on post representations.

Regarding supervised machine learning algorithms, the most frequently used classical machine learning techniques are the Support Vector Machines (SVM) [Tsugawa et al., 2015; Cheng et al., 2017; Wongkoblap et al., 2018; Aragón et al., 2019; Liu et al., 2019], Logistic Regression [Bagroy et al., 2017; Ricard et al., 2018; Trotzek et al., 2018a; Wongkoblap et al., 2018; Liu et al., 2019; dos Santos et al., 2020], Random Forests (RF) [Reece and Danforth, 2017; Liu et al., 2019; Mukta et al., 2022], Decision Trees (DT) [Wongkoblap et al., 2018; Liu et al., 2019], Naïve Bayes (NB) [Shen et al., 2018; Wongkoblap et al., 2018], or tree-based ensemble algorithms, such as AdaBoost and LightGBM [Mukta et al., 2022]. These methods have the advantage of being low resource intensive and often provide some degree of interpretability — such as using the coefficients of a linear SVM.

However, these methods often reach a performance saturation point, showing minor improvements even with bigger datasets. As such, deep learning algorithms are often used to overcome this limitation, such as Long Short-Term Memory (LSTM) [Orabi et al., 2018a; Coppersmith et al., 2018; Cha et al., 2022; Bucur et al., 2023; Wu et al., 2023], CNN [Orabi et al., 2018a; Trotzek et al., 2018a; Cha et al., 2022], and Transformers [Bucur et al., 2023]. Nevertheless, a few works do not use either LSTM or CNN for user-level classification, but as a sequence (of words) classification [Coppersmith et al., 2018; Trotzek et al., 2018a; Cha et al., 2022; Wu et al., 2023]. Only two of the selected related works use either LSTM or Transformer to classify individuals based on the sequence of posts [Orabi et al., 2018a; Bucur et al., 2023].

6.4. Enacting Change: Principles and Directives for Socio-Ethical Machine Learning Models for Screening Mental Disorders

To screen depression automatically, researchers often investigate Machine Learning (ML) methods that rely on social media publications to learn patterns associated with depression. Those models offer an alternative way to large-scale screening depression that could guide mental health administrators to create better policies. Although one could argue that such models have been successful to a degree, they are still experimental or impose a significant risk to use in the real world. As these models are inherently embedded in social systems, it is imperative to discuss the general role of technology and the existing issues around it that impact — and are impacted by — social aspects. Consequently, social and technological components are not isolated in society. We must understand the main social catalysts that lead to the innovation of technological solutions. This intertwined relationship between the said “social” and “technology” often results in sociotechnical

systems [Selbst et al., 2019]. We could perceive society as a continuously evolving sociotechnical system because of our increasing overreliance on technology.

Technology has imposed its presence in arguably all aspects of life: from leisure to work, gaming to video streaming, and augmented reality to shopping. Its obtrusive characteristics, nonetheless, challenge society in many ways, such as data privacy and insidious changes in socioeconomic structures. As these tools are present in many aspects of life, they act in elusive ways. For instance, it can change people's votes without their awareness [Epstein and Robertson, 2015]. The subtlety nature of some technologies acts purposedly to change our *psyche* in some direction. In this sense, they can be used as a mechanism to control opinion according to power relations, showing as political tools with moral implications.

AI, and more specifically, Machine Learning (ML), casts even more challenges with its high ability to address complex problems. With the recent advances in ML and subsequent implementations of those new technologies in production, we, as a society, are falling behind in understanding the real consequences these new tools bring, especially when they are taken for granted as purely beneficial and neutral. However, as researchers, we must understand the impacts of the technologies we create. Furthermore, we should understand the social and political consequences of the wide adoption of these technologies imbued with *psyche-changing* capacities.

The inquiry about how a technological advancement could potentially change inherent human behavior was first documented in Plato's *Phaedrus*. Plato's discourse investigated the influence of a now-incorporated and widely used technology in society: writing[4]. Despite its practical benefits, Plato asked himself whether writing would weaken an individual's memory capacity. Indeed, once, the ancient Greeks could recite a significant part — or entirely — of the Iliad by memory [Jaeger, 2001; Foley, 2007; Parry, 1933].

Individuals in our modern society rely more and more on new technological tools. The immediate consequence, as theorized by Plato, is the effect on the ability it will directly replace or partially replace — such as memory. For example, the very nature of preprocessing input data to standardized formats dehumanizes and deskills humans, removing any contingency and making everything predictable; the consequence: it promotes hegemonic behavior that might homogenize creativity [Burrell and Fourcade, 2021]. This is crucial because it does not happen individually, as technology and AI is widely used. This will inevitably impact society at all levels, not only the specialized individuals directly using the technology, which ultimately will disturb the socioeconomic framework. Are the creators of such tools, governments, or even the public sphere aware that those changes will profoundly impact society?

Still in Plato's *Phaedrus*, the old god Thoth, the creator of many arts, argues that his creation — the writing art — benefits individuals because it will improve their memory and wisdom. However, the other god discussing with Thoth argues that the "inventor of an art is not always the best judge of the utility or inutility of his inventions to the users of them" [Plato et al., 1952]. This inquiry is more modern than ever: Society as a whole — not only specialists but laypeople too — need to take part in the public discourse

[4]Analogy borrowed from [Ballesteros, 2020].

and take action about impactful technologies in their lives. Even though Plato's inquiry proved correct, the undeniable benefits of writing vastly outweigh its problems, as he left us many texts. However, we need to pose the same question that Plato once investigated, with a few more: Will this new technology change our behavior? If so, what are the consequences? Do the benefits far outweighs the new problems created by using such technology? In what context should the proposed technology be used to avoid any form of moral violation? If technology can influence users' psyches, how do we ensure those in power do not misuse it for their benefit? Are these technologies fair and just? Are they addictive?

Previous works have discussed the ethical impacts surrounding those questions for predicting mental health state. However, we stress that previous works discussed more on the perspective of conducting ethical and moral research guidelines [Conway, 2014; Benton et al., 2017; Chancellor et al., 2019], or understanding the general population's ethical opinion about using social media data for research [Mikal et al., 2016; Fiesler and Proferes, 2018]. Inspired by Selbst et al. (2019) [Selbst et al., 2019] ripple effect trap[5], we take another route: we are particularly interested in the humanistic, social, political, and ethical concerns that impact, and are impacted by, the creation and usability of models to screen depressed individuals deployed on social contexts.

As such, we rely on social sciences, philosophy, economy, and politics studies to develop an analytical framework discussing three main challenges of deploying models to screen depressed individuals based on social media data. Therefore, we contribute with the following an analytical framework: (a) First, we explore the challenges of the social system where the ML model is embedded, understanding how the model impacts the sociotechnical system and vice versa (Section 6.4.1); (b) Second, we investigate how the ML models impact different stakeholders to varying degrees (Section 6.4.2); (c) Finally, we investigate how data inequity and misrepresentation create inherently biased models that are ultimately dangerous for screening depressed individuals (Section 6.4.3). Finally, by observing the existing tensions in the aforementioned analytical framework, we propose strategies to mitigate the presented issues, such as mental health, data, and digital literacy. Given the gap in this interdisciplinary research, we contribute to approximate computer scientists to social sciences, giving more emphasis on the social aspect.

6.4.1. Those who rule and who are ruled

In this Section, we focus on the social, political, economic, ethical, and philosophical aspects that underpin the application of ML models to screen for depressed individuals using social media data. By doing this, we expect computer scientists to have a broader view of the fundamental social aspects underlying the deployment of models in the wild. We do not expect this Section to be an exhaustive enumeration of all social issues but rather to shed light on existing issues that computer scientists often overlook. Hence, we start by discussing that commercial interests are aligned with collecting behavioral data, which is essential to virtually any technology today. Next, we discuss how data creates a power relation deepening surveillance and control. Finally, we discuss the relationship of

[5]Defined by Selbst et al. (2019) as "Failure to understand how the insertion of technology into an existing social system changes the behaviors and embedded values of the pre-existing system". Here, however, we argue that this change is bidirectional.

commercial interests and data to the task of screening for depressed individuals on social media platforms.

The technology industry had a turning point just after the dot-com bubble. After firms failed to deliver profitable business, they found what would be called the "new oil": data [Burrell and Fourcade, 2021]. They found that the digital trace, or behavioral data, that users left behind using a system could generate profit. Companies started to use this data to push directed advertisements for their users while reaping a share of advertisers' payment. Furthermore, some firms also started to sell the data they had to other firms[6].

The new proposed way to monetize over digital trace data proved financially successful. Firms deliberately eased access to their systems to facilitate unimpeded entry, ultimately attracting a larger user base and generating more digital traces, which led to more profit. This powerful mechanism rapidly boosted firms' growth, leading to higher profits and stabilizing their market share and dominance. Companies that arrived earlier in technology gained so much power that they are now accused of antitrust violations [Kolhatkar, 2021]. Nevertheless, the monopoly is not only financial: they hold information and computing power, equating to political power in the era of *infocracy*. The consolidation of the called "Big Techs" resulted in conglomerates of firms in which no other firm can compete, or if that is the case, they buy the competitors or force them to give up [Noble, 2018].

Consolidated companies have collected users' digital traces for more than two decades. Previous AI methods have not effectively handled big data sets, be it for lack of hardware power or because the method's performance did not scale well with data. After the rebirth of neural networks under the garment of deep learning [Aggarwal et al., 2018], the Big Techs could now put their vast data sets under deep learning to improve their predictive capabilities and explore new avenues they could not access. For example, the studies mentioned in Section 6.3 collect user-generated content through social media, be it with explicit consent or not. However, individuals frequently do not know their data could be used in such a way, although they agree with the social media platform terms, often without paying proper attention to the small letters and large texts. As a result, users' data generate value for third-party entities without the *explicit consent* of the data creator — the data is being used to train models for screening depressed individuals. The implicit consent through accepting those platforms' agreement terms is too vague and frequently ambiguous. Users only know that their data could be shared or used to train models without explicitly knowing the model type and task it solves. Many individuals are unaware that their data can be used for such endings or think deleting previous posts will solve the problem [Mikal et al., 2016]. Furthermore, ML models are known to be vulnerable to attacks that target extracting information from the training set [Salem et al., 2019; Hu et al., 2022; Huang et al., 2022], which imposes even more risks for users whose data are used for training such sensitive models.

The data available for those companies are mainly personal or behavioral. They use this data to create what Byung-Chul Han calls *psychometrics*: a method to gener-

[6]Interestingly, the practice of surveying and selling data to other interested firms is much older than expected. For example, during his first job in 1907 in Portugal, Fernando Pessoa was tasked to collect information from firms to sell to other firms around the world [Zenith, 2022].

ate a personality profile [Han, 2022]. The vast data feeds deep learning models trained to predict the users' behavior or predilections. Based on the profiles, the company can better discriminate users' interests. For instance, it is possible to determine a group of individuals with depressed-related behavior through clustering techniques or topic modeling [Resnik et al., 2015; Dipnall et al., 2017]. When implying that individuals' behavior is similar to those of a group, we incur the risk of determining the individual identity as being equivalent to those of the group.

Consequently, we negate two fundamental principles of self-determination for identity: justification and control [Engelmann et al., 2022]. When firms group individuals by their psychometrics, they often do not disclose or justify that automatic decision. Thus, users do not know what firms "think" about them and have no access or chance to modify that "opinion". This problem is aggravated by the fact that firms arbitrarily choose the amount of data to determine whether an individual pertains to a group of individuals with depressed-related behavior. The amount of information that determines whether an individual is depressed or not is hardly written in stone, and it will be different for each specific case. By limiting an entire life that encloses unconscious and conscious actions through deterministic mathematical formulations, we crystallize human behavior to a set of mechanical rules by ignoring the contingency of life. Therefore, those systems must let individuals self-determine their identity by justifying and controlling how they are portrayed, especially for sensitive "classifications" such as screening for depressed individuals.

While firms are driven by commercial interests and market dominance, the "side-effects" of applying those technologies on a large scale should not be negligible. Even if the firm does not misuse their predictions, they could sell the psychometrics data to other firms — a standard practice in the market [Noble, 2018]. Furthermore, as users surf the web, their data could be retroactively associated with digital traces from other sites, yielding more power to those institutions that can collect and use this data to train more effective models. The ethical implication is, "Do they own the right to sell our information, which other digital systems could use to create enhanced psychometrics"?

In essence, the individual is no longer entitled to privacy, and firms use data they can collect for their commercial interests. Although digital systems often disclose how they use data in their agreement terms, technological tools transform individuals into hostages. Even if the individual wants to use something other than the system, they are left behind with only a few choices, which are frequently deficient [Noble, 2018; Burrell and Fourcade, 2021]. Thus, the most widespread tool is consistently improved because it is widely used: they generate more psychometrics data to train better models, generating better-directed advertisements, increasing profits and market dominance in a perpetual cycle. This is especially the case for Google products, where they transform its users in the *cybertariat*: "individuals that perform a continuum of unpaid, micro paid and poorly paid human tasks" [Burrell and Fourcade, 2021] on tasks such as verifying if there is a crosswalk in the given images.

Although companies' commercial interests sound distant from screening for depressed individuals, we contend that these elements form the core political, economic, and social fabric that underpins predicting depression status, among many other tasks.

Because companies and applications optimize to favor profits, they often rely on techniques that touch human cognition, emotion, and behavior — such as making addictive applications to keep users engaged. Bombarding users with microtargeted advertisements; exposing individuals to an ever-increasing competitive life; killing the alterity by creating Filter Bubbles[7]; excess of positivism — notably in social media platforms; societal norms and expectations such as hyper-productivity and self-optimization; selling behavioral data to interested third parties. These forces make individuals more exposed to stress and depression [Han, 2015]. While ML models for screening depressed individuals pose a solution to this increasing mental health issue, it is also inherently embedded into the commercial interests of the same society that generated this problem in the first place. As a result, we must first ask ourselves what is the best long-term strategy: create ML models to large-scale screen depressed individuals or fundamentally change societal structure to avoid stress, burnout, and depression?

Arguably, one common approach to both problems is education. From one side, society must be educated to understand insidious and predatory commercial practices to protect themselves. For example, in a focus group, despite individuals knowing that Twitter data is publicly available, they think that deleting posting history will protect them because they lack the knowledge that the data could be collected and saved in other personal databases [Mikal et al., 2016], or collected through paid API's and reselling services. In terms of the general population, individuals need to understand the principles and mechanisms that govern technology because they lack fundamental knowledge about data permanence. As such, individuals are surprised when they are told that their social media posts could be used to find the prevalence of depression [Mikal et al., 2016]. For the personal aspect, we will further discuss the education aspect in Section 6.4.2.

In our application of interest, screening for depressed individuals is not always a stand-alone model. Typically, it could be embedded into a larger system as its primary objective or a collateral effect of trace data collected through system usage. The embedded model is explicitly trained to screen possibly depressed individuals, often using standard psychological tests or self-report data as labels, as explored in Section 6.3. The objective and scope of such models are naturally restrained, and thus, it is easier to understand the ethical implications such models might pose. However, identifying possibly depressed individuals based on the digital trace they left behind poses a significant risk to privacy and the capacity of individuals to self-determine.

However, different environments create different opportunities and goals. Social or humanitarian goals are often absent in a firm where data is used for profit. On the other hand, in places that should be more welcoming, such as schools and universities, the goal of screening for depression is to benefit the students. Since the sample of undergraduate and graduate students is often three to six times more exposed to depression than the general population [Ibrahim et al., 2013; Evans et al., 2018], it is vital to find depressed students to conduct an effective intervention. In a welcoming, affectionate setting, where the university's goal is not to profit above everything else, there are legitimate social or

[7]Eli Pariser's concept about recommender systems that select ideas and news that individuals might like, or agree, based on their digital behavior patterns. As a result, individuals are predominantly exposed to information that aligns with their existing beliefs and relies less and less upon alternative viewpoints to shape opinion — alterity.

humanitarian objectives to help students instead of exploring their data for profit — although only sometimes valid, since there are private universities as well. Furthermore, in a controlled setting with clear goals to help students, we expect data collection and processing to follow strict ethical guidelines, such as the approval of the institutional review boards (IRBs). With that, volunteers to participate in the study (or deployed system) explicitly concede permission to use their data and that by providing the data, they explicitly benefited from the resulting system.

Unlike the university environment, firms are optimized for best matching the advertisement with their users. To do that, they need to discriminate well among various profiles, one of them possibly being the "depressed" profile. Depending on the technique they use, the profile is latent, which means that they do not know *a priori* if he or she is interested in "drama", "technology", "sports", or "depressed" content. However, they know well enough to predict predilections — a proxy for identity. If they take the time to evaluate groups of individuals with similar interests or personalities, they might find a group tied to depressive content; in the end, they could further explore and benefit from this particularity. More specifically, a social media platform company aims to improve engagement by neglecting the consequences it brings. For example, social media platforms increase engagement based on provocative and divisive publications, which opens the door to several conspiracy theories with manipulative content and false information [Rauchfleisch and Kaiser, 2020; Fisher, 2022]. When creating a space that amplifies divisiveness, addiction, and envy, social media platform companies lay the fundamental triggers for causing and sustaining depressive symptoms. Hence, when screening depressed individuals on social media, we may want to solve a problem inherently amplified by how commercial interests are intertwined with social media platforms.

Furthermore, with the psychometrics information, firms can disclose their findings to other subsidiary companies — improving their services —, but can also sell the same information to interested third parties. Even when users know that the platform sells their information, they often need to learn that their digital trace is being processed to find much richer information, such as psychometrics. In this scenario, users must judge whether they want their personality screened by predatory practices, but before that, they should be educated on how this technology works. While the general population is not educated, firms mine information arguably freely[8].

From another standpoint, a system incurs the risk of broadcasting inaccurate information. For example, one individual might be wrongly screened as depressed by the system. Next, driven by financial interests, the firm might commercialize identity information to other firms or redistribute it to its subsidiaries. It is clear that selling inaccurate information is not only prejudicial to the buyers — or the seller —, but it may cause harm to the individual whose misjudgment of the model directly affects. Furthermore, the user frequently has no mechanism to change how its identity is rendered nor control over its commercialization: they lack the justification and control capacities for self-determination [Engelmann et al., 2022][9]. On the other hand, if the firm and its sub-

[8]Although we have General Data Protection Regulation (GDPR) and Lei Geral de Proteção de Dados Pessoais (LGPD) to protect individuals' privacy and information, it is hard to keep a vast and unregulated space like the internet under control.

[9]The reader could argue that users can stop using the system. However, some systems are pervasive

sidiaries keep the false prediction, it might create an echo chamber of disinformation in all systems owned by the firm that initially committed the mistake.

This situation is undesirable by itself and aggravated when the general population widely uses a system — notably when the system is seen as a public resource[10]. The harm is not only on the individual; however, it can shift the perception of other users of the tool — which will inevitably promote hegemonic representations about an individual or a group of individuals. In that sense, Byung-Chul Han also argues that algorithmic operations have the potential to totalitarianism [Han, 2022]: as the classic totalitarianism thoroughly explains the past, present, and future through a straightforward truth to the detriment of multiple possibilities, ML models poses the same issues through black-box predictions. The model prediction is a new truth. By promoting a unitary truth — or worldview —, widely believed to be neutral and unbiased, we reach a state of totalitarianism by inadvertently burying other opinions and reinforcing the dominant standpoint, similar to what happens with recommendation systems that create Filter Bubbles. The dominant culture's views and opinions, including race hierarchy and power relations, often surmount the opinions of the marginalized groups. Even worse, the dominant's opinions of the marginalized groups often prevail over the opinions of the marginalized groups have about themselves[11]. Thus, depressed individuals suffer from social stigma because the dominant view of the mental disorder is not portrayed nor publicly disseminated the way individuals who have this mental disorder expect. The constant disinformation about the mental disorder in conformity with the rise of political ideologies of meritocracy and neoliberalism[12] results in a dominant view that does not comply with how depressed individuals see themselves. The representation of information is rendered as a function of the dominant view. Popularity is vital because AI and generalized technology are heavily based on statistical processes. However, popularity does not equate to truth.

If groups of individuals are being misrepresented in technological systems, how should we, as a society, approach this problem? First, we should educate society about the underlying functions of such digital systems so they know what is happening with the digital traces they left behind. Next, accountability is of utmost importance for harm caused to marginalized groups. However, the harms such as "black girls" portrayed as pornographic and sexualized girls in Google Search are often said to be simply a "glitch" that they can "fix" [Noble, 2018]. However, the "glitch" culminates in predatory practices that misrepresent identity and the self. Despite the huge impacts, companies such as Google bypass by "fixing" the "glitches" and explain that they are not guilty because this is an "anomaly" in the system [Noble, 2018]. Even worse, a study has demonstrated how to manipulate people's votes by changing the ranking in search results [Epstein and Robertson, 2015]. However, if the developers and owners of those systems are not accountable, since there is no policy to intervene, and public opinion perceives such systems as neu-

enough to cause social exclusion when users do not use them. Both social media and online search tools are examples.

[10]Google Search is an example of a widely used tool seen as a public resource.

[11]See [Noble, 2018] for an example of how a search with the keywords "black girls" used to portray black girls as the stereotype of sexualized individuals in 2011.

[12]This is notably true if we consider the dominant ideologies of software engineers from Silicon Valley, which are responsible for the vast majority of systems used by individuals across the globe. See [Noble, 2018] and [Burrell and Fourcade, 2021] for more information.

tral [Burrell and Fourcade, 2021; Noble, 2018], who will be accountable for such actions? Depression is already stigmatizing and isolating [Chancellor et al., 2019]; thus, how do we deal with technologies that statistically infer the presence of depression that is not detrimental to social media users? Although the Samaritan's Radar app was created to prevent suicide, bad actors were bullying and stalking vulnerable individuals [Lee, 2014]. It is crucial to emphasize the social media platforms' responsibility to their users' mental health. Moreover, the suggestion that "glitches" are beyond human control equates to denying the problems originated from multiple sources, such as dataset bias, algorithmic malfunction, and unethical corporation attitudes.

While firms are optimized for commercial interests and market dominance, can they capitalize on public resources widely used by society? How do we make the creators of digital systems, such as software engineers, CEOs, and programmers, accountable for the impact of the systems they create? While we, as a society, do not create a better public policy to intervene in such questions, we are left behind with manipulative software that amplifies the hegemonic worldviews that users seldom have other options to replace. Furthermore, firms are structuring tacit overarching narratives through technology. Although implicit, they develop pernicious unexpected outcomes that are not socially or politically discussed.

Restricting the scope and application of the model that works on screening for depressed individuals is one step towards more fairness in such a sensitive problem. On the one hand, institutions that control the data and the resulting model use them to optimize profit. On the other hand, we have institutions whose primary goal is to create models from provided data (explicitly) and whose primary benefits are returned to the individuals who provided the data themselves. We advocate that the second case is crucial for benefiting those who help create the predictive system for screening mental health social media users. In contrast, the first case might create severe consequences for the individuals who provide the data: predatory neoliberal practices, such as pushing improved marketing to users or selling services to users who contributed with unpaid work to create those said services.

6.4.2. Technologically Mediated Behavior

This Section discusses the impacts of potentially deployed ML models to screen depressed individuals on stakeholders. The stakeholders are mental health administrators, public policymakers, individuals who are the target of ML models, social media platform administrators, psychologists, or psychiatrists. Thus, we first discuss the fundamental challenges of technology that change daily behavior and routine to discuss more deeply the role of ML models as mechanisms that change perception and behavior for the case of screening depression using social media data.

Since elementary school, we have received feedback based on our assignments, exams, and quizzes. The feedback is essential to improving students' performance, but it also influences how students perceive themselves in the world. For example, if a student frequently receives low-grade scores, it can significantly impact their opinion, ability, and perception of themselves in society [Festinger, 1954; Burrell and Fourcade, 2021].

Nonetheless, the feedback loop is not restricted to education. The presence of

feedback loops in day-to-day work and relationships has been a common element. However, with technology being more pervasive in people's lives, the feedback takes the shape of constant visualizations, assessments, scores, and recommendations [Burrell and Fourcade, 2021]. In other words, we can monitor how many steps we walk, how many hours we sleep, our body temperature, and our heart rate. While those metrics are often helpful, they promote a continuous and intermittent feedback loop that users are sometimes unprepared for.

Metrified systems also create another effect: users might rely on them to have constant feedback. However, as Byung-Chul Han argues, algorithms exclude "the possibility of the experience of contingency" [Han, 2022], which have always been an essential tool for many scientific discoveries in history. By relying on contingency and through trial and error, humans constantly arrive at surprising conclusions which leveraged unpredicted situations. On the other hand, metrified feedback encourages comparison to other individuals, often called "social comparison" [Festinger, 1954]. This phenomenon lowers the barrier to comparing oneself to friends or acquaintances, which once was subjective, but now takes place as — supposedly — objective and metrified comparisons. The *apparently* objective comparisons can severely impact individuals' self-perceptions, possibly aggravating self-esteem and subjective well-being, which might reverberate into depression and anxiety symptoms [Verduyn et al., 2017; Hwnag, 2019; Sharma et al., 2022].

Consider, for example, the use of Grammarly[13]. It often notifies users to change phrases to a more positive tone. Although it may provide some practical value for both experienced and inexperienced users, it has the potential to *nudge* our writing capacity in a particular direction: writing positively. It also has another component: its *opinion* is implicitly saying that the person is writing *negatively*. At the same time that the machine's *opinion* seems subjective (the text *seems* negative, write it more positively), it wears the mantle of objectivity because the suggestions are mathematically driven and thus seem objective. While our opinion of our ability sometimes depends on other opinions — human opinions —, unambiguous criteria, such as mathematical formulae and metrics, provide a clear path to comparison: "I am better because my score is 10 and yours is 5". By suggesting to write positively, what are the impacts it could create? Will individuals see themselves as negative writers? Will people start writing more positively as a consequence? If society widely uses such a system, what are the consequences of suggesting everyone write positively? Another example, for instance, is when users are asked to make moral decisions. In ethical-sensitive domains, it is common to elicit the ethical values of stakeholders since any ML model will affect the stakeholders differently in varying degrees. Often, users guide their ethical values based on past and present information. However, when faced with predictive information, i.e., the opinion resulted from the prediction of an ML model or human expert, individuals' ethical preferences are directly impacted [Narayanan et al., 2022]. Moreover, they found evidence that humans prefer to rely more on the predictions of the ML model than a human expert [Narayanan et al., 2022]. In another experiment, even in light of the evidence of privacy and security issues, popular social media users did not delete nor change the privacy settings, which shows more evidence that individuals think they are immune to manipulation [Hinds et al.,

[13]A technological tool that assists users in detecting English errors while suggesting rewriting phrases for different tones.

2020]. Those experiments support the idea that ML copilots — although helpful for making decisions — have the potential to *nudge* behavior, i.e., manipulate opinion, possibly even about oneself. The worrying part is that this opinion is hegemonic and insidious. Given the same input, the model will take the same opinion (prediction).

Hannah Arendt argues that the sense of reality is mediated through shared knowledge among the community [Ballesteros, 2020]. She says that although we perceive the world through our point of view, understanding is reaffirmed because we have others perceiving the world the way we do. In that way, suppose the machine is the "other" — anthropomorphizing the machine —, by yielding constant feedback to users, the users might change the sense of reality to be coopted by the machine. General technology and AI are not only changing our perception of the world and sense of reality, but they also change how we are perceived — the said technologies of reputation, such as social media [Ballesteros, 2020]. As a result, users think that the technology they are using is innocuous, while the system acts on a pre-reflexive level [Han, 2022] to *nudge* behavior.

More formally, B. F. Skinner elaborated on the operant conditioning mechanism to explain how new behavior is learned in non-human animals. Based on positive and negative rewards, animals can keep or extinguish behavior accordingly. Animals learn new behavior through reinforcers (rewards, often food) and a reinforcement schedule, which delivers the reinforcers based on pre-defined rules. Skinner and his colleagues found that variable-time reinforcement schedules are the most effective for maintaining and strengthening behavior because the unpredictability of reward keeps the subject constantly engaged, anticipating the subsequent reinforcer [Staddon and Cerutti, 2003]. Although most experiments validate the theory for non-human animals, they also have been shown to have similar results for humans [Staddon and Cerutti, 2003].

Moreover, operant conditioning can be found in online gaming, traditional gambling games, and slot machines [Deibert, 2019]. When applying the operant conditioning theory to social media platforms, users are conditioned to expect reinforcers, such as likes, comments, or private messages — social rewards — in a variable time interval. The social rewards are assumed to share neural mechanisms with non-social rewards [Lindström et al., 2021]. Consequently, users engage in a continual expectation of new notifications and interactions (reinforcers), which in turn consolidates new addictive behavior. The immediate nature of online communication exposes individuals to immediate gratification, further strengthening addictive behavior. Social media platform addiction has been demonstrated to share neurological mechanisms with substance abuse [Turel et al., 2014; Kupferberg et al., 2016], and individuals with social media addiction experience similar symptoms to substance-related addiction symptoms [Kupferberg et al., 2016]. In a North American national survey, 32% of people with a substance use disorder also have a Major Depressive Disorder (MDD) [Carey, 2019; Xu et al., 2020]. More broadly, we advocate that addictive behavior is not only restricted to social relation expectations — such as likes, comments, and followers —, since social media platforms offer various services and experiences. For example, Facebook users might be addicted to the platform because of its games or buying and selling goods in Facebook groups.

Firms, through technology, explore fundamental laws of humans' neurophysiology mechanisms by hacking our reward system. Based on that, firms create addictive

software designs to keep users engaged in their systems: with engagement, they generate more data to secure profits. The elaborated strategy acts upon the reward system's affective dimension because it is faster and more effective than a reasoned argument [Han, 2022]. Instigating excitement, emotions, and engagement through simple images (memes) is more straightforward than a 10-page essay.

While the system slowly directs the user behavior in a particular direction, users often believe these tools to be neutral and non-biased [Noble, 2018; Burrell and Fourcade, 2021]. This false perception creates an ominous problem: technology changes user behavior without the user's consent or awareness. Even more, it transforms the users into hostages because not using the digital systems equates to social isolation or unproductivity, which are deemed necessary in our fast and hyperconnected society. In the context of screening for depressed individuals, the same problems can happen. Individuals using any digital system today are prone to be tracked and submitted to psychometrics, as explained in Section 6.4.1. Therefore, in the context of commercial applications, firms have the potential to find latent variables connected to depressed individuals based on digital traces. When users engage with the system, they also receive feedback, such as a system that alerts them that their written text contains negative emotions or that they are frequently recommended to depressed-related content. The Filter Bubble is crucial to understanding how users are constantly exposed to no content other than the content recommended by the system. The reception of highly channeled content related to depression potentially reinforces the identity of the self as a depressed individual. Despite the efforts to leave the Filter Bubble, individuals will face other systems that also use their data to recommend the same content. This perpetual cycle provokes rumination, especially for negative thoughts, which have a significant role in worsening or maintaining depressive episodes [Cooney et al., 2010].

Moreover, even when individuals can see and control the predictions about their identity in social media systems, it could not be beneficial. Data transparency is ubiquitously considered a good feature; however, as we have been demonstrating through this study, the predictions of ML models have the power to shape opinion. Thus, there are two negative sides when explicitly showing predictive information to social media users: first, transparency paves the way for feedback loops that improve the firms' systems, which then return as paid services or enhanced marketing to users — resulting in unpaid work; second, data transparency could work against the capacity of individuals to self-determine once the predictive opinion is seen as objective and factual truth, thus overriding the individual's opinion about oneself [Engelmann et al., 2022]. In a focus group, one individual said "The fact that if it was an algorithm, and they were looking like, 'Hey, we think you're feeling low right now.' I feel like it might make me feel even more low" [Mikal et al., 2016]. From another angle, surveillance and data transparency are linked: "it is not people but information that is truly free" [Han, 2022]. The consequence is that people's data are truly transparent, while domination itself is not transparent, and neither the black-box nature of ML models is transparent [Han, 2022].

Differently, when screening for depression is not embedded into a larger system — such as a search tool or social media — the process will take a different principle. We argue that an ML model exclusively used to screen depressed users mediated through public entities and strict policies should produce a more secure outcome. The impact will

not be directly on the perception of the average user of digital systems but on who uses the model. The "user" (or stakeholder) will be general practitioners, psychologists, psychiatrists, healthcare administrators, and public policy makers. In that scenario, the problem associated with a lack of education in users is reduced to a few professionals whose education could include ethics in technology. Despite general practitioners being educated on the usage of ML models, they are still prone to changing their perception when using the model, which hints at the question: how will this model impact the perception of those professionals?

On the one hand, even if the average general practitioner is educated or trained to use the tool, they can blindly rely — often unintended — on the technology. In a hypothetical situation, the general practitioner did not expect the patient to be depressed, but the model has alerted about the possibility of depression. The general practitioner might be biased to agree with the model and treat the patient as depressed, although the model functions as a support mechanism and not as a decisive truth. Though the general practitioner gives the last word, the black-box nature of the model's decision cuts any aspect of communicative rationality [Han, 2022]. For example, peer discussion about a patient's diagnosis and treatment is essential to any medical residency. By relying on a decision without further explanation or dialogue, communicative rationality is eliminated from the process, which is so profoundly important to shape rationality and, thus, to reason about any object. Although the model might help raise awareness of the possibility of depression, it still can not offer any explanation or reason — at least not how they are conceivable today.

Another possible collateral effect is the *belief* in technology. With the generalized perception that any technology is improving in ways that challenge our rationality, technology can turn into artifacts that hold truthfulness. For example, a study demonstrates that American users often think search engines are an unbiased source of information and even believe that what is shown is true [Noble, 2018]. Once, those technologies were not perceived as trustworthy; however, as firms improve their services, perception slowly shifts to an ominous *belief* state. Currently, ML models to screen depressed individuals are limited by construction, both in terms of performance and in the input signals they leverage. However, better models will not only rely on textual or visual cues: they can also leverage the improvements in IoT, such as capturing data, but not limited to, for example, blood pressure, body temperature, voice tone, video, and audio. The future is uncertain, but as models improve performance scores, they might reach a state of *belief* where they are seen as inherently better than human capacity in the task. The risk is when general practitioners see their abilities overtaken by the model's predictive performance. What is the impact of ML models tasked to screen depressed individuals if they are deemed better than humans? Though this is not happening now, preparing for what might come is vital.

On the other hand, a different opinion (from the AI model) might bring attention to an otherwise forgotten or unpracticed ability. General practitioners might understand they lack the expertise to diagnose depression while the tool helps them — because it might raise the possibility of depression while they are not seeing it. It might promote the inverse of what was discussed earlier: general practitioners will prepare themselves to better diagnose depression in light of evidence that their ability is deficient. The tool here is merely a support mechanism to help the general practitioner. The perception shift here

is beneficial to the user's ability.

Improving general practitioners' ability in primary care benefits society as a whole. In primary care, patients constantly complain about many symptoms, such as lack of concentration, sleep disturbances, forgetfulness, back pain, or headaches. These physical symptoms presented to general practitioners are often somatization processes associated with depression. Frequently, patients attribute emotional distress through physical manifestations of pain rather than psychological processes [Lipowski, 1990], which are called somatizations. Thus, allied to the fact that general practitioners often misdiagnose depression — only 10%–60% of cases are correctly diagnosed [Löwe et al., 2008] —, they prescribe unnecessary exams and medications based on these physical manifestations rather than the primary cause — depression. Therefore, inadequate treatment will bring risks to the patient's health without any benefit. At the same time, it will incur a waste of public resources, burdening the public health system. By relying on ML models to help identify the possibility of depression, the general practitioner will depend less on invasive diagnostic tests and treat the disorder accordingly without wasting public resources. Similar to what happened with Plato's writing inquiry, the benefits might vastly outweigh the issues.

In conclusion, we should worry about the entities who create the technology, but we also need to understand the impacts on users. One technology might cause significant harm to democracy while providing benefits to users, yet the users might be blind to such harm. Users should be educated, but more is needed to engage private corporations to include ethical processes in their pipeline.

6.4.3. Data, Inequity, and Misrepresentation

"Democracy is degenerating into infocracy" [Han, 2022]. Information — or data — has been commodified; consequently, all internet users have also been commodified since they are data producers. Although the internet is simple to join — you only must have a device and an internet connection —, it is far from a just and fair place. The internet is open, but often, those who use it and produce data are from specific hegemonic social hierarchies propagated through the digital medium. The evident problem of misrepresentation in traditional media is carried and amplified to the world of the internet [Noble, 2018]. The hegemonic representation (or opinion) will prevail over the representation of minorities' opinions due to firms' commercial interests and statistical processes. For example, the — English — data used to train GPT-2 and GPT-3 is heavily based on Reddit content, in which 67% of users are men between 18 and 29 years old [Bender et al., 2021]. Moreover, the content used to train GPT-3, for example, is highly filtered to remove unintelligible data or undesirable content, which might contain, among others, content created by marginalized communities that should not be filtered [Bender et al., 2021].

Such heavily skewed data is not only a matter of statistics but should open our eyes to the hegemonic data that originates hegemonic representations in technological systems. While general-purpose systems and ML models rely on statistics, in which case the most prevalent data will be emergent in any such system, the issue of misrepresentation will persist. Those said "neutral systems" are not neutral [Noble, 2018; Burrell and Fourcade, 2021]. Though exact and deterministic, the mathematics behind the systems is situated in a complex social body. Their mathematical processes are inherently political because

they are capable of changing behavior, and therefore they are also moral and ethical. This fight for representation hits directly into marginalized communities.

The excluded — or underrepresented — data is not only statistically insignificant to emerge but also opens another avenue to marginalization. The predominance of content related to marginalized communities on the internet is often not how marginalized communities self-perceive themselves. To illustrate this situation, a Google search with the keywords "black girls" in 2011 yielded several pages with porn content [Noble, 2018]. From hundreds of thousands of pages that could be returned to "black girls", the Google algorithm decided to return porn content; this is not unintentional. This could be a direct consequence of commercial interests drawn from the broad appeal of the hegemonic opinions circulated online. In other words, Google search users intend to see this type of content by searching "black girls". However, it could be a combination of both. Either way, this is not how black girls see themselves. As a consequence, historical issues of enslavement and portraying black women as sexual objects are still perpetuated through digital systems by hegemonic representations [Noble, 2018].

Users of social media platforms also need to be educated on how their data are being factually used. They lack knowledge about data permanence, thinking that deleting their posting history will render them immune to tracking, post-processing, or aggregate analysis [Mikal et al., 2016]. If users were to be educated on how their data are being used in post-processing mechanisms and how much profit is generated from their content production, they could safeguard themselves in a way that could significantly change how companies handle data.

Another issue related to social media data is the bias related to content production. As aforementioned, social media users tend to be skewed toward young adults, although the number of elderly individuals has been increasing [Center, 2021]. As such, models to screen depressed individuals trained with as much data are prone to perform better for young adults simply because they contain the most active individuals in online communication. In contrast, a sample of elderly individuals might not enjoy the same performance benefits from the model as young adults because they are not as present online. Another issue might be related to digital literacy, as elderly individuals still need to be educated on digital communication. As code (or programs) sets the range of usability, it enables and disables individuals and groups [Youmans and York, 2012]. Although social media platforms are open and accessible, underrepresented groups might not feel welcome in such places for various reasons, such as moderation practices or cyberbullying [Bender et al., 2021]. Hence, ML models relying on datasets where structural — coded — gatekeeping mechanisms impair a wide array of individuals to participate are inevitably prone to be biased. Therefore, using various data sources, looking particularly for sources where underrepresented groups are more prone to use, is one way of helping bridge the performance gap between different samples. Mainly, understanding where the sample of depressed individuals is posting or feel most comfortable posting is critical to accessing abundant sources of data to create better and robust ML models for screening depressed individuals.

While firms are optimized for commercial interests, they seldom reserve time or interest to understand their systems' "bugs" or "glitches". As data is the fundamental

commodity for those firms, they optimize their systems to keep users engaged and productive; in other words, they generate more data. For instance, racist, divisive, or conspiracy theory content promotes increased engagement, securing higher profits. Moreover, even when software engineers actively moderate such content, commercial interests inevitably stand in the way. It is easier and more profitable to keep users engaging freely — with possibly divisive content — than restricting them, resulting in less content [Roberts, 2016]. From interviews with commercial content moderators, Sarah T. Roberts found that despite the employees arguing that specific content — blackface, for instance — is undesired, the company decided to keep it [Roberts, 2016]. The threshold for keeping or removing content is a careful balance between profit and the company's public image. Moreover, even entire teams on machine-learning ethics have been dismantled, as happened with Twitter in 2022 [Chayka, 2022].

As the prominent data will be created through addictive practices and well-engineered algorithms to keep users engaged, the online data will inevitably be biased. If researchers do not carefully evaluate the data they are feeding to ML models, it will carry the same bias and misrepresentations in the data. Even worse, the model could amplify the misrepresentation and cause unpredicted harm, mainly directed at marginalized communities.

For the task of screening depressed individuals, there are key elements that distinguish who will benefit more or less from such technology. For example, as data online is prominently in English, native English speakers will have the advantage of having models trained on more data than native Brazilian speakers. Not only quantity determines model performance, but also quality. Native speakers of not-so-famous languages are left with fewer and low-quality data samples, resulting in less desirable performant ML models. Even if we consider transferring the knowledge from one ML model trained on a more extensive dataset to start training another ML model to work with a small dataset, we suffer from the portability trap [Selbst et al., 2019]. It means the social context in which the ML model is trained, such as the data and its type (which media), where it was produced (which social media), and from which sample (e.g., median age, nationality) might mislead or harm when used to train models to be embedded in other social contexts. Thus, it is imperative to correctly model the social context and possible ethical implications when transferring knowledge from one model to another.

Furthermore, using pre-trained models — trained on large datasets of generalized text or images — to improve the performance of the downstream task might pose other risks. First, as we have argued, those said "generalized" texts or images often suffer from bias and misrepresentation issues. Second, the said "pre-trained" models, also named "foundation models" [Bommasani et al., 2021], are a standard block used in many classification systems. As foundation models are trained on a massive corpus of data through deep learning algorithms, they will not only eclipse less frequent content but can also amplify the hegemonic content. As the current ML models for screening depressed individuals heavily rely on foundation models to extract good textual or visual representations for classification [Trotzek et al., 2018b; Orabi et al., 2018b; Aragón et al., 2019; Mann et al., 2020, 2022; Bucur et al., 2023], they also inherit the bias and misrepresentations included in the foundation model. On top of that, fine-tuning the foundation model to specialize it to the domain data, a standard practice in many automatic systems that screen individuals' data for depression, does not necessarily eliminate these issues — on the

contrary, it could perpetuate and amplify them.

As we approach a state of singularity in Computer Vision and Natural Language Processing (NLP) by using the same foundation models for various tasks, we risk perpetuating and spreading bias and misrepresentations to many aspects of the field. Furthermore, though unexpected emergent properties that arise in models are sometimes beneficial, there are times they may create undesirable collateral effects. For example, the emergent property of in-context learning that emerged from GPT-3 training, which is the capacity to adapt the language model to a downstream task based on a given context prompt as input, is attractive because it helps solve many tasks with higher accuracy and adaptability. However, it also might be the cause of hallucinations. As emergent properties are primarily unpredictable and need extensive empirical validation, it is too unstable to be reliable for screening depressed individuals using social media data.

Furthermore, GPT-3 and GPT-4 lays the groundwork for the vast improvement of Natural Language Generation (NLG). Differently from what is argued by [Bommasani et al., 2021], all data is not only created by people, nor only by people for other people. GPT-3 or GPT-4, a Large Language Model (LLM), can automatically generate data for diverse tasks with high capability. The internet, full of racist, homophobic, and divisive content, now faces another challenge: the content created by advanced NLG models, such as GPT-3 and GPT-4. Those models not only perpetuate the misrepresentations of the internet but also accentuate them by automatically creating content quicker and more effectively — though not always correct — than any human being. The risk of such stochastic parrots [Bender et al., 2021] for screening depressed individuals is that they could automatically generate "depressed-related" content. The boundaries for human-made content are blurry since we now have LLMs capable of creating human-like textual and visual content. Consequently, we advocate practitioners not using an LLM to "augment" datasets for screening depressed individuals, at least as conceivable today, because of the many inherent risks associated with the content they produce.

6.4.4. Enacting Change

From previous sections, we can observe the persistent tension between opposing interests or motivations that can explain the main challenges in screening MDD using social media data. In this Section, we explore two key areas to discuss further the above tensions for screening depressed individuals in social media: education and data protection laws. Finally, we conclude this Section by stressing a few suggestions highlighted from the discussed tensions and the proposed critical areas of discussion.

6.4.4.1. Education

There are two main ways of conceiving AI and education today: We can use AI to help educate individuals or educate individuals to prepare for the age of AI — here, we will focus on the latter. The same could be said for mental health disorders, where individuals still need to learn or improve their knowledge about them. For 193 member states in the United Nations, Quality Education (SGD-4) and Good Health and Well-Being (SGD-3) are among the 17 shared Sustainable Development Goals (SDG).

Education will be highly impacted by AI, be it by newly incorporated ways to learn and educate or by new ways of interacting and using technology that inevitably change the desired abilities when companies hire personnel. Nonetheless, the current United Nations SDG report states how the COVID-19 pandemic highly impacted education and advocates for higher investments in education and embracing technology to improve education [of Economic and Affairs, 2023]. While severe primary educational concerns remain, digital literacy, fundamental for benefiting and avoiding AI's dangers, is still largely lacking [of Economic and Affairs, 2023]. Digital literacy can be defined as "a set of skills required by 21st Century individuals to use digital tools to support the achievement of goals in their life situations" [Reddy et al., 2020].

We add to the above definition the necessity to understand how Information Technology (IT) companies are commercially motivated and how the data produced by individuals can be used for several purposes, as discussed in Section 6.4.1. Because companies explore ways to maximize profit at the expense of social and humanitarian goals, citizens must inform themselves that companies are consciously creating ways to explore human psychological vulnerabilities to expand engagement. Therefore, behavior is altered to promote engagement and addiction, which possibly culminates in anxiety and depression in expectation of social rewards, as discussed in Section 6.4.2.

Mental health disorders are still largely stigmatized and isolating [Chancellor et al., 2019]. As such, pushing the education boundaries to help citizens understand and recognize mental health disorders is crucial. To that, Mental Health Literacy (MHL) is essential to help promote recognizing, managing, and preventing mental health disorders [Jorm et al., 1997]. It can be defined as "understanding how to obtain and maintain positive mental health; understanding mental disorders and their treatments; decreasing stigma related to mental disorders; and enhancing help-seeking efficacy (knowing when and where to seek help and developing competencies designed to improve one's mental health care and self-management capabilities)" [Kutcher et al., 2016]. MHL shapes how lay individuals understand mental health disorders and seek appropriate help — otherwise, individuals might not even know they have depression.

For the tensions we have discussed in this text, we focus on how to educate students — or society, more broadly — to prepare for the age of AI, notably through the lens of applications to screen depressed individuals as one of the fundamental ways to enact change. Hence, digital and mental health literacies are essential for individuals to preserve autonomy and privacy and empower individuals in political contexts. We will explore two notable applications to understand better the potential of MHL and digital literacy.

For example, the Samaritans, a philanthropic organization that provides help to individuals with emotional distress and suicide ideation, created an app in 2014. This app, called "Radar", was aimed at Twitter users to detect signs of suicide or depression [Hsin et al., 2016]. However, the app was designed to alert the user when one of their friends exhibits the said signs of suicide or depression, even if their friends did not opt-in to participate. The signs of depression were determined by text matching with pre-determined phrases such as "I am sad" and "I want to kill myself", i.e., based on a keyword list. Such a simplistic approach to detection is easily prone to errors and misclassifications. This

raises serious concerns regarding privacy and accuracy because users are being analyzed without explicit consent, and the accuracy is questionable. Even worse, bad actors downloaded the app to encourage targeted individuals by the app — possibly suicidal — to take their lives.

Similarly, Facebook also offers the service of suicide prevention[14]. Facebook implemented two main strategies for suicide prevention, identifying potential individuals by (1) reactive reporting and (2) proactive reporting [Gomes de Andrade et al., 2018]. Reactive reporting is based on the collaboration of Facebook users that can "flag" whether a post demonstrates suicidal ideation, similar to reporting a post for violating community rules. The user who reported the post is then prompted to offer help directly or to delegate the responsibility to Facebook human evaluators. However, the Facebook staff noted that several posts containing suicidal ideation were not being reported. As such, they developed proactive reporting based on ML models. The model can identify posts from users at-risk, which, based on a threshold, a report is sent to the Community Operations to review the post and send resources, if applicable [Gomes de Andrade et al., 2018].

The primary issue learned from the Samaritans' Radar app fiasco is that privacy is vital: Who can access the information about screening depressed individuals on social media? We argue that access to this information needs to be very carefully controlled by mental health professionals. The user did not explicitly opt-in (agree to give explicit consent) to participate and might not even know they have been screened for depression and suicide risk. Equally important, the Facebook staff can infer suicidal ideation based on human and automatic evaluations. Are Facebook users aware that their friends can report their posts for containing the risk of suicide? Moreover, as a consequence, Facebook can learn about users' mental states by relying on the evaluation of Facebook users and the Community Operations personnel. While the Facebook service for suicide prevention is not as open to bad actors as happened to the Samaritans app, the service is embedded in a social media platform with commercial interests.

As discussed in Section 6.4.1, social media platforms, — and other mega-technological corporations — allied to advertisement strategies to influence individuals to purchase products may rely on vast psychometrics data. Based on this data, they could act like bad actors like the Samaritans' Radar app. Not only that, they have the potential to trigger and induce depression, stress, and anxiety, as we discussed in Section 6.4.2. The question is, do they have the right to induce such information based on self-generated data by their users?

With rapid AI development, institutions have yet to create better public policies. However, one of the best ways to avoid these pernicious practices is to increase society's digital and mental health literacy through education strategies. In a similar vein, the World Health Organisation has underscored the fact that health literacy[15] is a better predictor of health than many other factors, such as income and employment status [Furnham and

[14]Although it is not explicitly designed to screen depressed users, we note that the comorbidity of suicide ideation and depression is reported to be the highest among several mental disorders [Henriksson et al., 1993]

[15]MHL was inspired by health literacy, which could broadly be defined as "the ability to gain access to, understand, and use information in ways which promote and maintain good health" [Jorm et al., 1997].

Swami, 2018]. Although there is still a need for a body of evidence to state the same thing for mental health literacy, it is clear that there is an urgency to simultaneously increase the digital and mental health literacies to improve over the many discussed challenges. Therefore, by understanding how commercial interests are aligned with practices to addict and promote engagement, citizens need to protect their interests in favor of their (mental) health.

6.4.4.2. Data and Legislation

Data protection laws came to help in a longstanding issue regarding IT companies exploiting personal data without transparency on how they were using it or by obtaining explicit consent from individuals. Two notable protection laws are GDPR for the European Union (EU) and the Brazilian LGPD. Now, companies must comply with their countries' data protection regulations and the data protection laws of the countries where their companies are partners. Although both regulations improve towards protecting social and human rights in the era of AI, their implementations happened very recently — GDPR in 2016, and LGPD in 2020 —, and the effect of these laws are in its infancy.

Both data protection laws share the concept of "controller" and "data subject" ("*titular*" in LGPD). The controller is "any legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of processing personal data" [Parliament and of the European Union, 2016]. In LGPD, the data subject is defined as "a natural person to whom the personal data that are the object of processing refer to". Furthermore, personal data can be classified as sensitive, which refers to data concerning health, ethnic origin, political opinion, and genetic or biometric data when related to a natural person, among others.

Both LGPD and GDPR assume that any processing of personal data must be directly communicated and only conducted over explicit approval from the data subject. At the same time, processing personal sensitive data in GDPR and LGPD are forbidden except when the data subject provided explicit consent; however, there is a legal basis for processing sensitive data even without explicit consent from the data subject. Notably, among the seven exceptions in LGPD, we stress two of them most related to mental health disorders: (1) "to protect the health, exclusively, in a procedure carried out by health professionals, health services or sanitary authorities"; (2) protecting life or physical safety of the data subject or a third party. Processing personal sensitive data for health interests is also supported by GDPR but only under the responsibility of a professional obliged to professional secrecy [Parliament and of the European Union, 2016].

The data related to the mental sphere includes the possibility and risks of identifying mental disorders and the identity of data subjects. As such, when the controller can obtain psychometrics data to single out individuals, it violates privacy under the GDPR and LGPD laws. However, as Ienca et al. [Ienca and Malgieri, 2022] pointed out related to GDPR legal basis, "the data revealing thoughts or memories are not automatically sensitive data just because they refer to the 'mental sphere"'. When the data is not classified as sensitive, the possibilities to process and commercialize the data are less restricted than sensitive data. Mainly, non-sensitive data in LGPD can be processed for new purposes

— such as financial gains —, while subsequent processing for sensitive health data is explicitly prohibited for commercial gains.

Although GDPR explicitly provides a legal basis for mental ill-health issues, based on the "data concerning health" that explicitly includes mental health, there is no particular mention of mental health in LGPD, only to the broader "health" term. Moreover, while both laws provide specific guidelines to guarantee the data subject's privacy and self-determination, there is still a gap, especially for LGPD and mental health disorders, that legal entities can explore on a legal basis, even if such actions may not be ethically or morally sound. As suggested by other authors, this gap could be explained by the lack of more specific categories in processing sensitive data, such as including "emotions", "desires", "thoughts" (textually) as data that could be used to identify a natural person uniquely or infer several other personal dimensions [Ienca and Malgieri, 2022].

Therefore, as discussed in Sections 6.4.1 and 6.4.2, legal entities are in a position of obtaining, processing, and inferring depression and identity lawfully. Although other obligations apply, such as asking for consent, there is a clear gap in the legislation that allows treating such data as non-sensitive. Even when psychometrics data might be considered sensitive, both GDPR and LGPD allow processing them under the responsibility of a professional subject (in the case of health-related services) or by acquiring explicit consent from the data subject as defined in Article 9(2)(a) and 9(2)(h) in GDPR.

For the case of social media platforms, while legal entities can capture psychometrics data to infer other attributes lawfully, such as mental health states, users agree to Terms of Services based on what is called "weak consent regimes" [Ienca and Malgieri, 2022]. Social media platform users consent without reading the Terms of Services, or even when they read, it often contains complex language and specialized terminology that individuals do not understand [Obar and Oeldorf-Hirsch, 2020]. Consequently, obtaining consent from social media platform users under LGPD's Article 7 (and 11 for personal sensitive data) and GDPR's Article 6 (and 9 for personal sensitive data) is straightforward.

Although there might be legitimate interest from data subjects to self-track depression or suicide ideation, we note that the gap in treating textual thoughts and emotions allows legal entities to subsequently process the information for commercial interests, such as using psychometrics for advertising. Legal entities can proceed under this path lawfully, especially under LGPD, which is a legitimate interest only to the legal entities [Ienca and Malgieri, 2022]. However, we note several risks associated with collecting and inferring depression in an individual, as discussed in Sections 6.4.1 and 6.4.2. Besides these risks, Chancellor et al. (2019) [Chancellor et al., 2019] mentions the risks of advertisement for prescription drugs, credit score based on mental health state, and health insurance raising premiums because of existing mental disorders. Are there enough reasons for legal entities to lawfully share and process self-generated data, especially under the risks of manipulating behavior and opinion and amplifying existing prejudice?

While data protection laws are still improving, especially in light of the rapid AI development, individuals also need to understand how their data could be used without the scope of their legitimate interests. Data protection laws enforce privacy, but we argue that it is still insufficient to protect mental data, as it could be used to infer mental health disorders or other mental states [Ienca and Malgieri, 2022]. Remarkably, the two

mentioned laws have different approaches to mental health disorders, which is another relevant point of discussion — while GDPR is more robust to mental health data, LGPD is more flexible.

Thus, similarly to digital and mental health literacies, there is still an urgency to educate individuals to understand how legal entities are commercially motivated to use their data. Therefore, data literacy is another important aspect of today's digital world, where general-purpose technology, specifically social media, is present in virtually all aspects of life. Individuals must understand and be able to protect their self-determination and legitimate interests and reliably reject abusive data collection practices. Another way to help individuals in this direction is to promote legislation and Terms of Services in a simplified (plain language) way so that they can further comprehend their privacy rights and how legal entities are using their data.

6.4.4.3. Discussion

Based on the three main areas of tension we explored in Sections 6.4.1, 6.4.2, and 6.4.3, and the two main approaches to mitigate the tensions in Sections 6.4.4.1 and 6.4.4.2, here we propose practical suggestions to mitigate the political, social and ethical challenges of screening depressed individuals in social media.

As we have discussed, data protection laws need to be more comprehensive to protect individuals in terms of inferring mental states from raw visual and textual self-generated content. Consequently, policymakers, public administrators, and legislators should consider including mental data as personal sensitive data. More importantly, we advocate explicitly including mental health and mental data into LGPD to prevent misuse and promote the privacy and safety of individuals. In this way, the mental data used to predict MDD, for example, would be strictly used either (1) under the explicit consent of the individual as a tool for self-assessment and with the legitimate use only for that purpose; or (2) to use for purposes of research or medical diagnosis, the provision of health care or treatment conducted by a professional obliged to a professional oath.

Although current data protection laws offer some security and privacy to personal sensitive data, using the data for only the two scenarios is very restrictive. As such, we call for a public debate to discuss privacy versus the greater social good. On the one hand, less privacy would improve the performance of ML models since they rely on more data. On the other hand, more privacy restricts the performance — and generalizability — of ML models or concentrates the power to a few legal entities with power, as currently happens with the Big Techs. Furthermore, there is the public versus private debate: Should the government oversight and regulate more or less? Many restrictions and oversight would result in an Orwellian monitoring [Mikal et al., 2016], but too few restrictions could lead to predatory and unethical practices.

One way to enforce restrictions without creating an Orwellian society is to change the dynamics underpinning predatory practices that alter behavior and, consequently, opinion and thoughts. As such, Governments could focus on the main strategies legal entities are using that impose several risks to the mental health of individuals. This suggestion relies on the fact that depression and other mental health disorders in the era of

technology might correlate with technological addiction [Han, 2015; Lin et al., 2016; Primack et al., 2017]. Consequently, dedicating efforts to changing the dynamics that motivate corporations might improve collective mental health more extensively than dedicating resources to screen depressed individuals using ML models. We call for public opinion and more research towards this open debate.

Another suggestion is that Governments need to ensure the right actors are held responsible for their actions. Accountability is essential to ensure that legal entities comply with the legislation and data protection laws. When Big Tech giants say it is "a bug" or "a glitch" in the system, it is not. It is a critical problem in the system that relies on the cacophony of prejudice or errors in human-generated data — and now, the LLM data.

There is also the concern of using personal data to infer depression. Users of social media platforms might agree that using individual data to infer mental disorders, such as depression, is undesirable. However, if the platform could create a mechanism to create population-based algorithms instead of user-based algorithms, we could advance research by understanding the behavior of individuals suffering from depression, for example. Thus, such findings can help public policymakers create better policies and strategies to improve services and social welfare. When using a population-based method and ensuring that no individual can be singled out, we incur fewer risks of violating privacy. We advocate creating models that can work on aggregate-level data without access to individual-level data in ways that would be impossible to identify individuals.

When implementing the ML model, the debate of intervention versus observation must also be considered. For research, intervention is complex and limited, given the unstable nature of the developed models and limited resources. For commercial or governmental interests, deploying models on real-life situations might help if aligned with intervention strategies. Consequently, training the relevant personnel on the inherent risks, benefits, and optimal utilization of observation tools for effective intervention is crucial. The goal of this entire process should be focused on the appropriate intervention strategies that will ultimately benefit society.

Education is fundamental to the professionals involved in the observation and intervention aspects of screening depressed individuals on social media. However, society must also be educated. On the one hand, individuals lack fundamental knowledge about data literacy; on the other hand, they also lack appropriate knowledge about mental health literacy. While technology consistently relies on data to create improved services, microtargeted advertisements, and addictive interfaces and mechanisms, individuals often lack the knowledge of minimum safeguards. As such, they use predatory services at the expense of their mental health. We advocate for improving mental health and data literacy.

6.5. Final Considerations and Future Perspectives

The use of artificial intelligence (AI) in supporting the diagnosis, treatment, and prognosis of mental disorders, especially depressive and anxiety disorders, presents significant potential and challenges. It is crucial to reflect on the current achievements and future directions in this interdisciplinary field.

AI's capacity to analyze large amounts of data from social media and other digital

sources offers a unique opportunity for early detection and intervention of mental health issues. Integrating machine learning models with psychological and psychiatric practices can improve the accuracy and efficiency of screening mental disorders, leading to more personalized and timely treatment plans. Additionally, AI-driven tools can provide continuous monitoring and support, aiding individuals in managing their conditions more effectively.

However, several challenges remain. The ethical implications of using AI in mental health care are vital. Privacy, data security, and informed consent issues are especially crucial when dealing with sensitive personal information. Researchers and practitioners must navigate these concerns carefully to ensure that the deployment of AI tools respects individuals' rights and complies with regulatory standards such as the General Data Protection Regulation (GDPR) and the *Lei Geral de Proteção de Dados* (LGPD) in Brazil.

Another critical area for future research is the development of more robust and inclusive AI models. Current models often face limitations due to biased training data, which can result in inaccurate predictions and worsen health disparities. Efforts should be directed towards creating diverse and representative datasets and implementing fairness-aware algorithms that mitigate biases and promote equitable outcomes across different demographic groups.

Interdisciplinary collaboration is essential to advance the field of AI in mental health care. Partnerships between computer scientists, mental health professionals, ethicists, and policymakers can promote the development of innovative solutions that are both technically sound and ethically responsible. Furthermore, involving patients and the broader public in the design and implementation of AI tools can enhance their acceptance and effectiveness.

Looking ahead, integrating AI with other emerging technologies such as wearable devices, virtual reality, and telehealth platforms holds great promise. These technologies can complement AI applications by providing real-time data, immersive therapeutic experiences, and remote care options, making mental health services more accessible and responsive to individual needs.

In conclusion, the application of AI in mental health care represents a transformative approach to addressing the growing burden of mental disorders. By continuing to explore new methodologies, address ethical and technical challenges, and foster interdisciplinary collaboration, we can harness the full potential of AI to improve mental health outcomes and enhance the well-being of individuals and communities worldwide.

References

Charu C Aggarwal et al. Neural networks and deep learning. *Springer*, 10(978):3, 2018.

Yemi Aina and Jeffrey L Susman. Understanding comorbidity with depression and anxiety disorders. *The Journal of the American Osteopathic Association*, 106(5 Suppl 2):S9–14, 2006.

Mario Ezra Aragón, Adrian Pastor López-Monroy, Luis Carlos González-Gurrola, and Manuel Montes. Detecting depression in social media using fine-grained emotions. In *Proc. of the 2019 NAACL-HLT, Volume 1 (Long and Short Papers)*, pages 1481–1486, 2019.

American Psychiatric Association et al. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013.

Shrey Bagroy, Ponnurangam Kumaraguru, and Munmun De Choudhury. A social media based index of mental well-being in college campuses. pages 1634–1646. Association for Computing Machinery, 2017. ISBN 9781450346559. doi: 10.1145/3025453.3025909. URL https://doi.org/10.1145/3025453.3025909.

Alfonso Ballesteros. Digitocracy: Ruling and being ruled. *Philosophies*, 5(2):9, 2020.

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

Aaron T Beck, Robert A Steer, and Gregory Brown. Beck depression inventory–ii. *Psychological Assessment*, 1996.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proc. of the 2021 ACM FAccT*, pages 610–623, 2021.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Adrian Benton, Glen Coppersmith, and Mark Dredze. Ethical research protocols for social media health research. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pages 94–102, 2017.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Ana-Maria Bucur, Adrian Cosma, Paolo Rosso, and Liviu P. Dinu. It's just a matter of time: Detecting depression with time-enriched multimodal transformers. In Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo, editors, *Advances in Information Retrieval*, pages 200–215, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-28244-7.

Jenna Burrell and Marion Fourcade. The society of algorithms. *Annual Review of Sociology*, 47: 213–237, 2021.

Theadia L Carey. Use of antidepressants in patients with co-occurring depression and substance use disorders. *Antidepressants: From Biogenic Amines to New Mechanisms of Action*, pages 359–370, 2019.

Pew Research Center. Social media use in 2021, 2021. URL https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/. Accessed: 2023-08-14.

Junyeop Cha, Seoyun Kim, and Eunil Park. A lexicon-based approach to examine depression detection in social media: the case of twitter and university community. *HUMANITIES & SOCIAL SCIENCES COMMUNICATIONS*, 9, 6 2022. doi: 10.1057/s41599-022-01313-2.

Stevie Chancellor and Munmun De Choudhury. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*, 3(1):1–11, 2020.

Stevie Chancellor, Michael L Birnbaum, Eric D Caine, Vincent MB Silenzio, and Munmun De Choudhury. A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 79–88, 2019.

Kyle Chayka. A twitter employee's account of surviving layoff day. https://www.newyorker.com/culture/infinite-scroll/a-twitter-employees-account-of-surviving-layoff-day, 2022. Accessed: 2023-01-25.

Qijin Cheng, Tim M H Li, Chi-Leung Kwok, Tingshao Zhu, and Paul S F Yip. Assessing suicide risk and emotional distress in chinese social media: A text mining and machine learning study. *JOURNAL OF MEDICAL INTERNET RESEARCH*, 19, 6 2017. ISSN 1438-8871. doi: 10.2196/jmir.7276.

David A Clark. Cognitive restructuring. *The Wiley handbook of cognitive behavioral therapy*, pages 1–22, 2013.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1485–1497, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://aclanthology.org/C18-1126.

Mike Conway. Ethical issues in using twitter for public health surveillance and research: developing a taxonomy of ethical concepts from the research literature. *Journal of medical Internet research*, 16(12):e290, 2014.

Rebecca E Cooney, Jutta Joormann, Fanny Eugène, Emily L Dennis, and Ian H Gotlib. Neural correlates of rumination in depression. *Cognitive, Affective, & Behavioral Neuroscience*, 10(4): 470–478, 2010.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 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*, pages 31–39, 2015.

Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. Natural language processing of social media as screening for suicide risk. *Biomedical informatics insights*, 10: 1178222618792860, 2018.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *Proc. of the ICWSM*, volume 7, 2013.

Vanessa Borba de Souza, Jéferson Campos Nobre, and Karin Becker. DAC stacking: A deep learning ensemble to classify anxiety, depression, and their comorbidity from reddit texts. *IEEE J. Biomed. Health Informatics*, 26(7):3303–3311, 2022. doi: 10.1109/JBHI.2022.3151589. URL https://doi.org/10.1109/JBHI.2022.3151589.

Ronald J Deibert. Three painful truths about social media. *J. Democracy*, 30:25, 2019.

Sahraoui Dhelim, Liming Chen, Sajal K Das, Huansheng Ning, Chris Nugent, Gerard Leavey, Dirk Pesch, Eleanor Bantry-White, and Devin Burns. Detecting mental distresses using social behavior analysis in the context of covid-19: A survey. *ACM Comput. Surv.*, 6 2023. ISSN 0360-0300. doi: 10.1145/3589784. URL https://doi.org/10.1145/3589784.

Joanna Frith Dipnall, JA Pasco, Michael Berk, LJ Williams, Seetal Dodd, FN Jacka, and D Meyer. Why so glumm? detecting depression clusters through graphing lifestyle-environs using machine-learning methods (glumm). *European Psychiatry*, 39:40–50, 2017.

Wesley Ramos dos Santos, Amanda M. M. Funabashi, and Ivandré Paraboni. Searching brazilian twitter for signs of mental health issues. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 6111–6117. European Language Resources Association, 2020. URL https://aclanthology.org/2020.lrec-1.750/.

Severin Engelmann, Valentin Scheibe, Fiorella Battaglia, and Jens Grossklags. Social media profiling continues to partake in the development of formalistic self-concepts. social media users think so, too. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 238–252, 2022.

Robert Epstein and Ronald E Robertson. The search engine manipulation effect (seme) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences*, 112(33):E4512–E4521, 2015.

Teresa M Evans, Lindsay Bira, Jazmin Beltran Gastelum, L Todd Weiss, and Nathan L Vanderford. Evidence for a mental health crisis in graduate education. *Nature biotechnology*, 36(3):283, 2018.

Ethan Fast, Binbin Chen, and Michael S Bernstein. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 4647–4657, 2016.

Leon Festinger. A theory of social comparison processes. *Human relations*, 7(2):117–140, 1954.

Casey Fiesler and Nicholas Proferes. "participant" perceptions of twitter research ethics. *Social Media+ Society*, 4(1):2056305118763366, 2018.

Max Fisher. *The Chaos machine: the inside story of how social media rewired our minds and our world*. Hachette UK, 2022.

John Miles Foley. "reading" homer through oral tradition. *College Literature*, pages 1–28, 2007.

Sigmund Freud. Notes upon a case of obsessional neurosis. *Standard edition*, 10, 1909.

Guanghui Fu, Changwei Song, Jianqiang Li, Yue Ma, Pan Chen, Ruiqian Wang, Bing Xiang Yang, and Zhisheng Huang. Distant supervision for mental health management in social media: Suicide risk classification system development study. *JOURNAL OF MEDICAL INTERNET RESEARCH*, 23, 6 2021. ISSN 1438-8871. doi: 10.2196/26119.

Adrian Furnham and Viren Swami. Mental health literacy: A review of what it is and why it matters. *International Perspectives in Psychology*, 7(4):240–257, 2018.

Norberto Nuno Gomes de Andrade, Dave Pawson, Dan Muriello, Lizzy Donahue, and Jennifer Guadagno. Ethics and artificial intelligence: suicide prevention on facebook. *Philosophy & Technology*, 31:669–684, 2018.

C Gorenstein, WY Pang, IL Argimon, and BSG Werlang. Manual do inventário de depressão de beck–bdi-ii. *São Paulo: Editora Casa do Psicólogo*, 2011.

Paul E Greenberg, Andree-Anne Fournier, Tammy Sisitsky, Mark Simes, Richard Berman, Sarah H Koenigsberg, and Ronald C Kessler. The economic burden of adults with major depressive disorder in the united states (2010 and 2018). *Pharmacoeconomics*, 39(6):653–665, 2021.

Byung-Chul Han. *The burnout society*. Stanford University Press, 2015.

Byung-Chul Han. *Infocracy: Digitization and the crisis of democracy*. John Wiley & Sons, 2022.

Markus M Henriksson, Hillevi M Aro, Mauri J Marttunen, Martti E Heikkinen, ET Isometsa, Kimmo I Kuoppasalmi, JK Lonnqvist, et al. Mental disorders and comorbidity in suicide. *American journal of psychiatry*, 150:935–935, 1993.

Joanne Hinds, Emma J Williams, and Adam N Joinson. "it wouldn't happen to me": Privacy concerns and perspectives following the cambridge analytica scandal. *International Journal of Human-Computer Studies*, 143:102498, 2020.

Honor Hsin, John Torous, and Laura Roberts. An adjuvant role for mobile health in psychiatry. *JAMA psychiatry*, 73(2):103–104, 2016.

Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54 (11s):1–37, 2022.

Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. Are large pre-trained language models leaking your personal information? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2038–2047, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL `https://aclanthology.org/2022.findings-emnlp.148`.

Ha Sung Hwnag. Why social comparison on instagram matters: Its impact on depression. *KSII Transactions on Internet and Information Systems (TIIS)*, 13(3):1626–1638, 2019.

Ahmed K Ibrahim, Shona J Kelly, Clive E Adams, and Cris Glazebrook. A systematic review of studies of depression prevalence in university students. *Journal of psychiatric research*, 47(3): 391–400, 2013.

Marcello Ienca and Gianclaudio Malgieri. Mental data protection and the gdpr. *Journal of Law and the Biosciences*, 9(1), 2022.

Werner Jaeger. A formação do homem grego. *São Paulo: Fontes*, 2001.

Samireh Jalali and Claes Wohlin. Systematic literature studies: database searches vs. backward snowballing. In *Proceedings of the ACM-IEEE international symposium on Empirical software engineering and measurement*, pages 29–38, 2012.

Anthony F Jorm, Ailsa E Korten, Patricia A Jacomb, Helen Christensen, Bryan Rodgers, and Penelope Pollitt. "mental health literacy": a survey of the public's ability to recognise mental disorders and their beliefs about the effectiveness of treatment. *Medical journal of Australia*, 166(4):182–186, 1997.

Sean W Kelley and Claire M Gillan. Using language in social media posts to study the network dynamics of depression longitudinally. *Nature communications*, 13(1):870, 2022.

Sheelah Kolhatkar. Lina khan's battle to rein in big tech. https://www.newyorker.com/magazine/2021/12/06/lina-khans-battle-to-rein-in-big-tech, 2021. Accessed: 2023-09-06.

Aleksandra Kupferberg, Lucy Bicks, and Gregor Hasler. Social functioning in major depressive disorder. *Neuroscience & Biobehavioral Reviews*, 69:313–332, 2016.

Stan Kutcher, Yifeng Wei, and Connie Coniglio. Mental health literacy: Past, present, and future. *The Canadian Journal of Psychiatry*, 61(3):154–158, 2016.

Dave Lee. Samaritans pulls 'suicide watch' radar app. *BBC News*, 11 2014. URL `https://www.bbc.com/news/technology-29962199`. Accessed: 2023-08-12.

Liu Yi Lin, Jaime E Sidani, Ariel Shensa, Ana Radovic, Elizabeth Miller, Jason B Colditz, Beth L Hoffman, Leila M Giles, and Brian A Primack. Association between social media use and depression among us young adults. *Depression and anxiety*, 33(4):323–331, 2016.

Björn Lindström, Martin Bellander, David T Schultner, Allen Chang, Philippe N Tobler, and David M Amodio. A computational reward learning account of social media engagement. *Nature communications*, 12(1):1311, 2021.

Zbigniew J Lipowski. Somatization and depression. *Psychosomatics*, 31(1):13–21, 1990.

Xingyun Liu, Xiaoqian Liu, Jiumo Sun, Nancy Xiaonan Yu, Bingli Sun, Qing Li, and Tingshao Zhu. Proactive suicide prevention online (pspo): Machine identification and crisis management for chinese social media users with suicidal thoughts and behaviors. *JOURNAL OF MEDICAL INTERNET RESEARCH*, 21, 6 2019. ISSN 1438-8871. doi: 10.2196/11705.

Bernd Löwe, Robert L Spitzer, Janet BW Williams, Monika Mussell, Dieter Schellberg, and Kurt Kroenke. Depression, anxiety and somatization in primary care: syndrome overlap and functional impairment. *General hospital psychiatry*, 30(3):191–199, 2008.

Paulo Mann, Aline Paes, and Elton H Matsushima. See and read: detecting depression symptoms in higher education students using multimodal social media data. In *Proc. of the ICWSM*, volume 14, pages 440–451, 2020.

Paulo Mann, Elton H Matsushima, and Aline Paes. Detecting depression from social media data as a multiple-instance learning task. In *Proc. of the ACII*, pages 1–8. IEEE, 2022.

Priya Mathur, Amit Kumar Gupta, and Abhishek Dadhich. Mental health classification on social-media: Systematic review. Association for Computing Machinery, 2023. ISBN 9781450399937. doi: 10.1145/3590837.3590946. URL `https://doi.org/10.1145/3590837.3590946`.

Jude Mikal, Samantha Hurst, and Mike Conway. Ethical issues in using twitter for population-level depression monitoring: A qualitative study. *BMC Medical Ethics*, 17, 4 2016. ISSN 14726939. doi: 10.1186/s12910-016-0105-5.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Md. Saddam Hossain Mukta, Salekul Islam, Swakkhar Shatabda, Mohammed Eunus Ali, and Akib Zaman. Predicting academic performance: Analysis of students' mental health condition from social media interactions. *BEHAVIORAL SCIENCES*, 12, 6 2022. doi: 10.3390/bs12040087.

Saumik Narayanan, Guanghui Yu, Wei Tang, Chien-Ju Ho, and Ming Yin. How does predictive information affect human ethical preferences? In *ACM Conference on AI, Ethics, and Society*, 2022.

Thin Nguyen, Dinh Phung, Bo Dao, Svetha Venkatesh, and Michael Berk. Affective and content analysis of online depression communities. *IEEE Transactions on Affective Computing*, 5(3): 217–226, 2014.

Safiya Umoja Noble. Algorithms of oppression. In *Algorithms of Oppression*. New York University Press, 2018.

Jonathan A Obar and Anne Oeldorf-Hirsch. The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society*, 23(1):128–147, 2020.

United Nations. Department of Economic and Social Affairs. *The Sustainable Development Goals: Report 2023*. UN, 2023.

Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. Deep learning for depression detection of twitter users. pages 88–97, 2018a.

Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. Deep learning for depression detection of twitter users. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 88–97, 2018b.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

Wenjing Pan, Bo Feng, and Cuihua Shen. Examining social capital, social support, and language use in an online depression forum: social network and content analysis. *Journal of Medical Internet Research*, 22(6):e17365, 2020.

Javier Parapar, Patricia Martín-Rodilla, David E. Losada, and Fabio Crestani. Overview of erisk 2021: Early risk prediction on the internet. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12880 LNCS:324 – 344, 2021. doi: 10.1007/978-3-030-85251-1_22. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85115867066&doi=10.1007\%2f978-3-030-85251-1_22&partnerID=40&md5=a5a71fb7eea7c4abea8e8208da6f1649. Cited by: 19.

European Parliament and Council of the European Union. General data protection regulation (gdpr), 2016. URL https://gdpr-info.eu/. Accessed: 2023-08-17.

Milman Parry. Whole formulaic verses in greek and southslavic heroic song. In *Transactions and Proceedings of the American Philological Association*, pages 179–197. JSTOR, 1933.

James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577, 2003.

Plato et al. *Phaedrus*, volume 275. Bobbs-Merrill Indianapolis, 1952.

Brian A Primack, Ariel Shensa, César G Escobar-Viera, Erica L Barrett, Jaime E Sidani, Jason B Colditz, and A Everette James. Use of multiple social media platforms and symptoms of depression and anxiety: A nationally-representative study among us young adults. *Computers in human behavior*, 69:1–9, 2017.

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

Adrian Rauchfleisch and Jonas Kaiser. The german far-right on youtube: An analysis of user overlap and user comments. *Journal of Broadcasting & Electronic Media*, 64(3):373–396, 2020.

Pritika Reddy, Bibhya Sharma, and Kaylash Chaudhary. Digital literacy: A review of literature. *International Journal of Technoethics (IJT)*, 11(2):65–94, 2020.

Andrew G Reece and Christopher M Danforth. Instagram photos reveal predictive markers of depression. *EPJ Data Science*, 6, 2017. doi: 10.1140/epjds/s13688-017-0110-z. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85027011738&doi=10.1140%2fepjds%2fs13688-017-0110-z&partnerID=40&md5=12591ab9e3de234208ab344a688f4cec.

Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. Beyond lda: exploring supervised topic modeling for depression-related language in twitter. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 99–107, 2015.

Benjamin J Ricard, Lisa A Marsch, Benjamin Crosier, and Saeed Hassanpour. Exploring the utility of community-generated social media content for detecting depression: An analytical study on instagram. *JOURNAL OF MEDICAL INTERNET RESEARCH*, 20, 6 2018. ISSN 1438-8871. doi: 10.2196/11817.

Robert H Rice. Cognitive-behavioral therapy. *The Sage encyclopedia of theory in counseling and psychotherapy*, 1:194, 2015.

Esteban A. Ríssola, David E. Losada, and Fabio Crestani. A survey of computational methods for online mental state assessment on social media. *ACM Trans. Comput. Heal.*, 2(2):17:1–17:31, 2021. doi: 10.1145/3437259. URL https://doi.org/10.1145/3437259.

Sarah T Roberts. Commercial content moderation: Digital laborers' dirty work. 2016.

Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. In *26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24-27, 2019*. The Internet Society, 2019.

Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 59–68, 2019.

Aparna Sharma, Kavish Sanghvi, and Prathamesh Churi. The impact of instagram on young adult's social comparison, colourism and mental health: Indian perspective. *IJIM Data Insights*, 2(1):100057, 2022.

Tiancheng Shen, Jia Jia, Guangyao Shen, Fuli Feng, Xiangnan He, Huanbo Luan, Jie Tang, Thanassis Tiropanis, Tat-Seng Chua, and Wendy Hall. Cross-domain depression detection via harvesting social media. In *Proc. of the 27th Int. Joint Conf. on Artificial Intelligence (IJCAI-2018)*, pages 1611–1617, 2018.

Ruba Skaik and Diana Inkpen. Using social media for mental health surveillance: A review. *ACM Comput. Surv.*, 53, 6 2020. ISSN 0360-0300. doi: 10.1145/3422824. URL `https://doi.org/10.1145/3422824`.

John ER Staddon and Daniel T Cerutti. Operant conditioning. *Annual review of psychology*, 54 (1):115–144, 2003.

Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.

Marcel Trotzek, Sven Koitka, and Christoph M Friedrich. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Trans. on Knowledge and Data Engineering*, 2018a.

Marcel Trotzek, Sven Koitka, and Christoph M Friedrich. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Trans. on Knowledge and Data Engineering*, 2018b.

Sho Tsugawa, Yusuke Kikuchi, Fumio Kishino, et al. Recognizing depression from twitter activity. In *Proc. of the ACM CHI*, pages 3187–3196, 2015.

Ofir Turel, Qinghua He, Gui Xue, Lin Xiao, and Antoine Bechara. Examination of neural systems sub-serving facebook "addiction". *Psychological reports*, 115(3):675–695, 2014.

Philippe Verduyn, Oscar Ybarra, Maxime Résibois, John Jonides, and Ethan Kross. Do social network sites enhance or undermine subjective well-being? a critical review. *Social Issues and Policy Review*, 11(1):274–302, 2017.

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

Akkapon Wongkoblap, Miguel A Vadillo, and Vasa Curcin. A multilevel predictive model for detecting social network users with depression. pages 130–135, 2018. ISBN 978-1-5386-5377-7. doi: 10.1109/ICHI.2018.00022.

En-Liang Wu, Chia-Yi Wu, Ming-Been Lee, Kuo-Chung Chu, and Ming-Shih Huang. Development of internet suicide message identification and the monitoring-tracking-rescuing model in taiwan. *JOURNAL OF AFFECTIVE DISORDERS*, 320:37–41, 6 2023. ISSN 0165-0327. doi: 10.1016/j.jad.2022.09.090.

Le Xu, Jun Nan, and Yan Lan. The nucleus accumbens: A common target in the comorbidity of depression and addiction. *Frontiers in neural circuits*, 14:37, 2020.

William Lafi Youmans and Jillian C York. Social media and the activist toolkit: User agreements, corporate interests, and the information infrastructure of modern social movements. *Journal of Communication*, 62(2):315–329, 2012.

Richard Zenith. *Pessoa: uma biografia*. Schwarcz S.A, 2022.

