Moral Framing of Mental Health Discourse and Its Relationship to Stigma: A Comparison of Social Media and News

Shravika Mittal
smittal87@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Munmun De Choudhury
munmund@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

ABSTRACT

Mental health discussions on public forums influence the perceptions of people. Negative consequences may result from hostile and "othering" portrayals of people with mental disorders. Adopting the lens of Moral Foundation Theory (MFT), we study framings of mental health discourse on Twitter and News, and how moral underpinnings abate or exacerbate stigma. We adopted a large language model based representation framework to score 13,277,115 public tweets and 21,167 news articles against MFT's five foundations. We found discussions on Twitter to demonstrate compassion, justice and equity-centered moral values for those suffering from mental illness, in contrast to those on News. That said, stigmatized discussions appeared on both Twitter and News, with news articles being more stigmatizing than tweets. We discuss implications for public health authorities to refine measures for safe reporting of mental health, and for social media platforms to design affordances that enable empathetic discourse.

CCS CONCEPTS

• **Human-centered computing → Collaborative and social computing**.

KEYWORDS

moral foundation theory, mental health discourse, twitter, news media, BERT, stigma

ACM Reference Format:
Shravika Mittal and Munmun De Choudhury. 2023. Moral Framing of Mental Health Discourse and Its Relationship to Stigma: A Comparison of Social Media and News. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3544548.3580834

1 INTRODUCTION

Mass media impacts the thinking, behavior, and emotions of the general population: Bandura [6] noted media to "serve as socializing agents that aid in construction and perpetuation of perceptions and learned behaviors." When it comes to mental health, in the absence of actual experience with people with mental illness, individuals have been known to often rely on print media (newspapers, journals, or magazines), television shows, and films as important sources of information about mental health [91]. In recent years, with the proliferation of social media use, people have been using social media to self-disclose, seek support, raise awareness about mental health, and combat stigma [25, 102].

The widespread prevalence of mental health discussions on social media platforms and news media implies that they are not only reflecting public attitudes and values, but also increasingly shaping societal perceptions in the public sphere. For instance, prior research has found that 'frames' of mental illness not only inform the public *what* to think about, but *how* to think about it [57] – a concept derived from the sociological notion of 'Framing' [33]. Consequently, negative framings of mental health may create misperceptions, myths, and hostile attitudes toward those with mental illness and their caregivers. In turn, this can have negative consequences for people with mental illness, a group that already experiences widespread human rights violations, social disadvantages, and systemic inequities. Consequently, facing potential or real discrimination in employment, education, and healthcare, coupled with the fear of being labeled 'mentally-ill', individuals with mental health challenges may avoid self-disclosure or seeking help and treatment [144]. In contrast, positive framings may influence the development and cultivation of benevolent views, reduce stigma, and contribute to a change in public attitudes. This is especially important given growing support for policies that seek to shift the mental healthcare model from institutionalization of sufferers to a community care approach [147].

Given the significance of understanding media framing of mental health, researchers from many fields, such as communication, journalism, and public health have studied the topic over the years. This research has discovered that, unfortunately, in news media, journalists often adopt a stigmatizing frame in reference to mental illness by associating it with violent criminal behavior [23, 143]; other times they adopt an overly sensational framing to increase readership [131]. It is also less likely for news stories to present information on medical breakthroughs or personal victory accounts on mental illnesses [23]. Although nascent, some social computing research has observed similar problematic (e.g., flippant or mocking) framings on social media as well [71, 115]. These research have revealed overly insensitive or brutish allusions of mental health in news and social media, however, to our knowledge, the underlying moral values in these descriptions have not been explored.

Morality has shaped how we extend mental healthcare for several decades [98]. Hence, understanding framings of mental health in public discourse through the lens of underlying moral values can shine a light on the roots of specific framings, explain which moral arguments may carry weight with certain consumers, and

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3580834

how specific moral positions may have differential impact. Additionally, Kleinman and Hall-Clifford [75] emphasized a need to understand how the moral standing of individuals and groups impacts the transmission of stigma associated with a phenomenon. Corrigan and Penn [22] said that "in terms of mental illness, stigmas represent invalidating and poorly justified knowledge structures that lead to discrimination." There is a need to recognize how social and cultural experiences, portrayed via moral values, create stigma, since stigma is a social, interpretive, or cultural process [75]. They further suggested that by "focusing on how local values enacted in people's lives affect stigma, we will be able to create more effective and measureable anti-stigma interventions." Thus looking at stigma via the lens of moral experience, or what is most at stake for actors in a local social world, Yang et al. [153] said that it could "[provide] a new interpretive lens by which to understand the behaviors of both the stigmatized and stigmatizers, for it allows an examination of both as living with regard to what really matters and what is threatened." In essence, with knowledge of the moral framings of mental health, it may be possible to craft measures that reduce the negative portrayal of mental disorders, allowing individuals to disclose and seek help without fear and shame.

In this paper, we explore the framings of mental health discourse in social media and news via the lens of the Moral Foundation Theory [60]. The theory was proposed by a group of social and cultural psychologists to study how notions of morality vary across individuals, relationships, institutions, or cultures. The theory consists of five foundations represented via virtue-vice facets: Care/Harm, Fairness/Cheating, Loyalty/Betrayal, Authority/Subversion, and Sanctity/Degradation, that extend the 'three ethics' (autonomy, community, divinity) used to describe morality [124]. Additionally, following [75], we investigate the relationship between internal attitudes (or moral frames) and negative perceptions (or stigma) associated with mental health discourse, i.e., which virtue or vice facets of the five foundations relate to mental health stigma. Focusing on two forms of public mental health discourse – Twitter and News – we address three research questions:

RQ1: What are the moral foundation framings of mental health discussions on Twitter and News?
RQ2: How stigmatized are these discussions?
RQ3: How do moral foundations and stigma relate to each other in these two types of mental health discourse?

To answer these research questions, we collected data from Twitter and various news sources using 30 pre-defined, expert-curated mental health related keywords. In total we worked with 13,277,115 public tweets and 21,167 public news articles on mental health, all of which were shared/published in concurrent timeframes (2020-21). We created vector representations or embeddings – contextual representations of language – using a BERT-based framework [28] for the five moral foundations and a WordNet [92] based new lexical resource to capture language around approval or stigma of mental illness. These representations were then scored against the sentence level textual embeddings for tweets and news articles. Interestingly, our analysis for RQ1 reveals that mental health discussions initiated by the general public on Twitter align more with the positive dimensions of all five moral foundations than those initiated by journalists on News. That is, Twitter's mental health discourse is inclusive, kind, and justice-focused. Despite this, both tweets and news articles used highly stigmatized framings (RQ2), and stigma was more prevalent in those tweets and articles that adopted vicious (negative) moral framings (RQ3).

Through our approach and findings, this paper thus offers the following contributions. This paper makes the first attempt to understand human moral values in the context of mental health discourse, and we do so by adopting a theoretically-grounded approach to understanding morality, i.e., Moral Foundation Theory. Furthermore, this work is uniquely positioned as it looks at cross-media analysis of mental health framings. Our findings bear implications for safe-reporting guidelines that can shape public discourse of mental health, for journalists as well as for the general public. We also highlight technical artifacts that could be incorporated in social media platforms to facilitate mental health discussions that are more inclusive, kind, and equitable. This paper additionally provides a first of its kind human- and empirically-validated dictionary for stigma, available for use by the broader community.

Ethics Statement. In this paper, we utilize public Twitter posts and news articles. As an observational study of retrospectively gathered data and without any interaction with the authors of these content, our research did not qualify as "human subjects research," per our Institutional Review Board guidelines. Nevertheless, we followed best practices in our analysis [19], such as working with deidentified data, and refraining from sharing raw or personally identifiable data in any form. All quotes in this paper are paraphrased to reduce traceability and potential harm to those who authored the analyzed data. *The paper contains descriptions of mental illness and suicide, which may be triggering to some readers, thus we suggest caution and use of self-care resources in reading this work.*

2 BACKGROUND AND RELATED WORK

2.1 Morality, Stigma, and Mental Health

Moral issues around mental health and mental illness have been of interest to scholars since several decades. A large part of this conversation has centered around the moral responsibilities to society from those with mental illness, as well as what would count as an adequate and sensitive societal response to misdoings by such individuals [16, 99]. As societal treatment of those with mental illness has evolved over time, so have attitudes toward the moral underpinnings of mental health. 19th and early 20th century conceptions often framed mental illness as a "character flaw," sometimes as a "deadly sin;" [38, 106, 146] in many parts of the world, acts of suicide were considered criminal activities punishable in a court of law [93]. For a long time, suicide thus continued to represent in the eyes of doctors and psychiatrists an act that stands in opposition to family, work, religious and other social values. In fact, to psychiatrists practicing mental health help during this period, suicide meant "an abandonment of one's duties to society, to the state and to the sanctity of life, values that were integrated into their etiological studies of this act" [150, pp. 1]. Hence the moral stance of the day said legal and religious sanctions against a person who attempted suicide were warranted and justified [82].

Fortunately, more progressive views have emerged in the present century [98]. Moral guidance today discourages and sometimes disapproves of framings for sufferers of substance use disorders as "addicts" or those with schizophrenia as "schizophrenics." Journalistic guidelines in recent years, to fight stigma, suggest reporting of loss of life due to a suicide as an individual "dying by suicide", rather than the previously commonly used criminalized framing of "committing a suicide" [7]. Still, scholars today argue that the erstwhile moral arguments against mental health concerns, particularly suicide, which existed for centuries, shaped early psychiatric theories and discourse on suicide, and continue to thrive in present day suicidology [79]. MacDonald [86] argued that these moral arguments are at the crux of the medicalized conceptualization of mental health, as it is through these moral values that medicine was able to "appropriate" these acts. It is therefore unsurprising that some experts consider these problematic moral positions responsible for our limited understanding of the risk factors driving adverse outcomes like suicide even today – a persistent issue that continues to hamper help-seeking and prevention efforts [41]. Specifically, Yampolsky and Kushner [150] posited that it is imperative to lay bare the moral values on which the medicalized assumptions about causes of mental illness have been based, such as sexuality, religious practice, criminality and excessive alcohol use; according to Esquirol [34], these used to be called *causes morales* in 19th century France and oftentimes, a causal chain was drawn between social, moral and psychological causes of mental illness: "The more civilization is developed, the more the brain is excited, the more one's susceptibility is active, the more one's needs increase, the more one's desires are imperious, the more there are causes for sorrow, the more mental alienation is frequent, the more suicides there must be." In short, moral values are inherently intertwined in the way we conceptualize and understand mental health in the society.

The need to adopt a morality lens to understand general perceptions of mental illness is further underscored by the prevailing social stigma around mental illness [23]. Those suffering from mental disorders constitute among the most stigmatized, disenfranchised, impoverished, and vulnerable members of society [49]. Stigma influences the cultural and systemic expectations of how to respond to people in distress; it also shapes beliefs about personal accountability and agency, and in turn, help-seeking behaviors [70]. For instance, some existing narratives may see those with mental illness as morally weak "benefits scroungers," forcing them to choose between "taking accountability" and "control" of their actions and emotions, or accepting a more passive, ill and "defective" role in order to get support [137]. Consequently, scholars have emphasized the recognition of the moral narratives that underpin both mental health care and processes of reform towards sufferers [152].

Taken together, a study that explores moral underpinnings of popular mental health discourse is warranted, as such discourse shapes public views on mental health and mental illness. This paper seeks to close this gap through the study of two prominent channels of mental health discourse today – social and news media.

2.2 Media Framings of Mental Health Discourse

Given the prevalence of mental health discussions on several media platforms, there has been prior work that investigates how the general public (in social media platforms) or journalists (in newspapers) refer to mental health concerns. This branch of research stems from the sociological foundations of "Framing" [33] – frames "select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described." Prior works have adopted framing theory to study hate groups on social media [109], re-design political crowdfunding campaigns [29], and identify bias in news media [100]. In our context, framing suggests that the representation of mental health on social media or news media impacts the cognitive thinking of their consumers; in journalistic inquiry, it is known that inappropriate framing of suicide events may result in copycat suicides, formally called the Werther Effect [43]. In general, stigmatizing framings in mass media can exacerbate feelings of oppression among those with mental illness, as stigma can interfere with their social integration, violate their civic rights, self-image, and family life, and could result in employment and housing discrimination [23]. We provide an overview of prior research on mental health framings in news media and Twitter.

Wahl [144] looked at newspaper framings to observe that "dangerousness" was most commonly associated with stories on mental illness. A comprehensive study that reviewed mass media's role in shaping mental health stigma revealed a similar finding [76]. Interestingly, Gwarjanski and Parrott [57] observed that online news framings on schizophrenia impacted online social behavior of their readers. News articles that associated schizophrenia with violent and criminal behavior received stigmatizing reader comments. Conversely, news articles with a stigma-challenging representation of schizophrenia received stigma-challenging reader comments. It was also observed that the readers were more likely to self-disclose their personal experiences with mental health in the presence of stigma-challenging news frames. There have also been attempts to conduct cross-culture studies to see how news coverage for depression differs between the English and the Spanish languages [145]. The works discussed here adopt a qualitative annotation-based framework, limiting their analysis to a few news articles. In contrast to these, our study utilizes a well-grounded BERT-based framework that can process large-scale data and capture complex semantic context in news. Additionally, though these works highlight the social disadvantages associated with stories on mental illness such as forms of aggression, incompetency, and reactions to adversity, they do not uncover the journalists' moral underpinnings.

In the light of the presence of problematic framings in mass media, prior work has highlighted training programs that can better inform the journalists on how to report mental health discussions to reduce negative implications. Corrigan et al. [23] investigated the impact of reading a positive, neutral, or negative journalism article on mental illness. The authors observed that positive articles prompted the readers to portray affirming attitudes ("recovery, empowerment, and self-determination") toward individuals with mental illness. Unsurprisingly, Australia's National Mental Health Strategy emphasises on appropriate media portrayal of mental health. In this context, several resources have been curated to facilitate responsible reporting. Using one such resource, Francis et al. [42] evaluated mental health discussion on non-fiction media across nine dimensions (e.g., "does the item emphasize the illness rather than the person?"). They found that most media items were of good quality though only a few included relevant help services. This finding highlights the importance of training programs or relevant resources that the journalists can refer to better frame mental health on mass media. A similar finding was noted by Sumner et al. [131], who examined how news shared on Facebook adhered to suicide-reporting guidelines. They found such news articles to describe more harmful, rather than protective elements, although articles with safe-reporting guidelines tended to be shared more.

Research has also been conducted to explore how the general public talks about mental health on social media platforms. While studying posts on two hashtags: #depression and #schizophrenia, Reavley and Pilkington [112] observed that though a majority of tweets were supportive, a proportion of them consisted of stigmatized framings. Another Twitter study on #schizophrenia observed that the use of the adjective "schizophrenic" was often negative, sarcastic, or medically inappropriate [71]. Language that mocks or trivializes mental health conditions has also been found to be prevalent on Twitter [115]. Another study explored negative and positive framings around bipolar disorder and other mental illness on Twitter [15]. Stigmatizing tweets received fewer retweets compared to those about personal experiences. However, these social media studies analyze the framing of mental health discourse using a discrete scale of measurement, i.e., they perform a qualitative analysis to rate a small sample of tweets as either supportive or stigmatizing. They do not explore the internal attitudes employed by the tweet authors towards mental health. Recent works on AI-generated responses to online mental health discussions highlight the importance of using empathic frameworks that go beyond the binary positive-negative framings and consider attitudes of compassion, warmth, and internal understanding [97, 120]. Our paper addresses this concern by providing a framework that quantifies moral frames of compassion, justice, and equity-centered values in mental health discourse, via Moral Foundation Theory.

Closest to this paper is the work of Pavlova et al., who sought to identify "mental health frames" beyond the largely studied binary stigma or stigma-challenging framings [105]. Seven mental health frames were observed, which were discovered via Latent Dirichlet Allocation: 'Awareness', 'Feelings and Problematization', 'Classification', 'Accessibility and Funding', 'Stigma', 'Service', and 'Youth'. Each of these frames was further explored to identify general sentiment using the Linguistic Inquiry and Word Count (LIWC) program [107]. This investigation revealed that mental health discourse is often used to problematize social issues and focuses on common mental illnesses like depression or anxiety.

Our work builds on this body of work. We contrast prior approaches to understand mental health framing in media by taking a theoretically-grounded approach – we harness the Moral Foundation Theory [60], as it provides a systematic way to explain variation in human moral reasoning. Notably, we provide a first approach to objectively quantify stigma in language, and by assessing its manifestation in mass media, our work discovers new connections between stigmatization of mental illness and the morality that underlies media discourse. Furthermore, this is the first work that performs a cross-platform analysis, between social media and news media, to analyze moral and stigmatized framings of mental health.

2.3 Social and News Media Studies Using the Moral Foundations

The Moral Foundation Theory (MFT) [59] has been a core framework to study the functioning of individuals, relationships, or institutions. MFT proposes that several innate and universally available psychological systems are the foundations of "intuitive ethics" [60]. Consequently, a lot of these studies have looked at moral differences between conservatives and liberals, which highlight that though liberals limit their usage to two foundations (care/harm and fairness/cheating) conservatives endorse all the five foundations [52].

Of relevance here, MFT has been adopted in studies of online media: Dehghani et al. [27] explored linguistic differences between weblogs of conservatives and liberals in the context of a controversial political issue in the United States, "Ground Zero Mosque", using the Moral Foundation dictionary [44]. Other studies employed the MFT to understand stances towards U.S. politicians [116], differences in U.S. immigration policy debates [55], political behavior and framing [69, 113], attitudes during the George Floyd protests [110], sentiment towards Asians during the COVID-19 pandemic [74], as well as underlying tonality of social media conversations [73]. MFT has additionally been applied to study human values based on their social media interaction. Notably, Kalimeri and colleagues [72] used MFT to study the moral values of individuals that "liked" pro or anti-vaccination Facebook pages. Authors observed that those resilient to vaccination portrayed anti-authoritarian values, and those in support of vaccination portrayed values of family and care. Prior research has also employed MFT to study moral framings in news media. A significant amount of work applied MFT to understand partisan differences across news outlets on issues of public interest, such as climate change, police violence, and vaccination [46, 94, 119]. Carvalho et al. [18] designed a fake news classification tool by assessing the difference in moral foundations to contrast between reliable and low-reputation sources. MFT has additionally been applied to study political discourse in news media on Mosque construction controversies [13], environmental attitudes [35], and Oklahoma Sharia amendment campaign [14]. We advocate that exploring moral values is essential for understanding internal attitudes humans portray while framing content on social and news media platforms, an aspect unexplored in the context of mental health discourse, which forms a basis for our current study.

3 DATA

3.1 Twitter

Among the different social media platforms, Twitter has been established to demonstrate many attributes of a news media [81]; at the same time, researchers have noted that Twitter is often employed as a broadcasting mechanism to raise awareness around mental health and fight stigma [25]. For our RQs, we deemed Twitter and News to therefore be comparable and adequate platforms in terms of understanding moral frames around mental health discourse.

3.1.1 Data Gathering Approach. We used the *focalevents*[1] Application Programming Interface (API) to collect public tweets on mental health. Our search query filtered the tweets using 30 pre-defined mental health related keywords (e.g. anxiety, depression, bulimia,

[1] https://github.com/ryanjgallagher/focalevents

Table 1: Number of tweets (#Tweets) and news articles (#Articles) matching each mental health keyword in our collected Twitter (13,277,115 tweets) and News (21,167 news articles) datasets. A tweet or news article may contain multiple keywords, contributing to the count for each of them. To validate the datasets we report the precision (PR), proportion of relevant tweets/news articles, on a random sample of 50 tweets and 20 news articles for each keyword.

Keyword	Twitter		News	
	#Tweets	PR (50)	#Articles	PR (20)
anxiety	1997522	0.94	3218	0.95
depression	1289793	0.92	6691	1.00
mental health	2337163	0.96	1860	0.95
mental illness	365259	0.96	330	0.95
mental disorder	54649	0.96	54	0.95
bipolar	85121	1.00	95	0.90
bpd	53945	1.00	1287	1.00
ptsd	87941	1.00	21	0.90
paranoia	86204	0.84	242	0.95
schizophrenia	27148	0.96	71	1.00
schizophrenic	18386	1.00	24	0.90
schizo	52240	1.00	92	1.00
panic attack	182691	0.84	113	0.90
panic	1044765	0.80	1175	0.85
anxiety attack	76945	0.90	11	0.82
social anxiety	61317	0.92	36	0.90
self harm	74775	0.94	11	1.00
self-harm	25596	0.92	103	0.95
eating disorder	70951	0.96	139	0.90
binge eating disorder	830	0.94	3	1.00
anorexia	24715	0.90	20	0.95
anorexic	15330	0.92	2	1.00
bulimia	4898	0.88	13	0.92
bulimic	1944	0.88	5	1.00
unwanted	219620	0.82	520	0.80
stress	205572	0.80	585	0.85
depressed	2327089	0.94	8170	0.95
depressing	423255	0.94	331	0.95
suicidal	965692	0.96	216	1.00
suicide	1685428	0.90	1409	0.90

Table 2: General statistics of the Twitter and News datasets.

	Mean	Median	Std dev
Tweets/user	2.51	1.00	7.29
#Words: tweets	22.51	22.00	11.64
#Words: news	992.45	744.00	1062.77

etc.), refer Table 1, derived from prior research on social media and mental health [20, 67]; in these works, the keyword set was rigorously curated and validated through consensus generation among public health experts. Particularly, Choi et al. [20] carefully explored five online data sources – Google, YouTube, Twitter, Reddit, and Tumblr – to identify the keywords listed in Table 1. We acknowledge that this keyword set may not represent the entire space of conventional mental health discourse, but they do capture the most common or frequent textual cues used by people in mental health discussions on online platforms [39, 54, 89]. This resulted in the collection of 22,149,834 tweets. The tweets were collected for 20 months from March 2020 to October 2021. Since collection of all possible tweets matching the said keywords was not practical (some of the keywords were present in potentially millions of tweets because of their rather general nature), we extracted one week worth of tweets for each month. These extraction weeks in any given month were chosen randomly to avoid temporal bias.

3.1.2 Filtering and Data Cleaning. The focus of this study was to examine the underlying moral and stigmatized framings in the broader context, i.e., to explore how people express general perceptions of mental health and not how they self-disclose or frame their own mental health issues. As a result, following collection, we filtered and fine-tuned this corpus by removing those tweets that had personal self-disclosures of a mental health condition. Additionally, this step was performed to make a more fair comparison between Twitter and News – since news articles are unlikely to discuss one's own mental health challenges, equivalently, we focused on non-personal presentations of mental health topics in tweets. Specifically, we filtered tweets containing self-disclosures like "I was diagnosed with (anxiety | depression | schizophrenia)", "I used to self-harm", "I had a (panic attack | anxiety attack)" etc. Refer to Table A2 for a complete list of self-disclosure Twitter search queries used in this study. Such queries have been used in multiple prior works to identify self-reported postings about diagnosis or experience of a mental illness or its symptoms [12, 21, 56]. After removing such tweets we were left with 17,442,077 tweets. We further removed tweets that originated from organizational accounts using *Humanizr* [90]. With the removal of organizational tweets we were left with 13,277,115 tweets posted by 7,169,239 unique users, with an average of 2.51 tweets per user. Table 1 provides a distribution of tweets corresponding to each mental health keyword. Table 2 provides an overview of number of tweets per user and length of tweets. Here are some examples of paraphrased tweets in our dataset: "There are many treatment options for depression. The most important step is to reach out for help if you need it.", "People with schizophrenia are more likely to be infected.", and "We should stop associating people's volatile behavior with bipolar disorder."

3.2 News

Next, we started with six publicly available News datasets to compile news articles on mental health. These included, the BBC News dataset [53], the AG News classification dataset [154], the MIND dataset [149], the News Aggregator dataset [31], the All the news 2.0 dataset [138], and the Harvard Dataverse dataset [78]. From these, we only considered the last two datasets that contained articles published within the timeframe of our Twitter dataset i.e., from March 2020 to October 2021. We temporally aligned the timeline of our Twitter and News datasets to make a valid comparison of moral framing between them, avoiding the confounding impact of time on mental health attitudes or writing styles. Table 3 provides an overview of the two news article datasets we used for this study. These either contain the actual news articles or their headlines, and have been used in prior research, offering face validity of use:

(1) *All the news 2.0 dataset*: This dataset spans over 27 news publishers. Prior research has used it for identifying patterns of political polarization in media [58], perceived differences

Table 3: An overview of the two publicly available news datasets adopted to compile news articles on mental health.

Dataset	#Articles	Type of text	Top 4 Source(s)/Publisher(s)
All the news dataset 2.0	2688878	Headline, Article	Breitbart, New York Post, CNN, Washington Post, and 23 others
Harvard Dataverse	1244184	Headline	Australian Broadcasting Corporation

between human and machine generated news articles [135], and news recommendation [126].

(2) *Harvard Dataverse*: This is a publicly available dataset on Kaggle, a renowned online data science community. It has been used for fraudulent news headline detection [85].

The news articles in both the datasets were filtered using the same set of 30 mental health related keywords that were used to filter the tweets, as mentioned earlier. Table 1 provides a distribution of news articles corresponding to each mental health keyword. In total, we were able to collect 21,167 news articles on mental health. Our dataset contains news articles that are textually represented using (a) only headline and (b) space separated headline and article body. This choice was made in reference to prior works that use headline-only news datasets, in addition to those with article excerpts, for language understanding tasks like text classification [2], event detection [37], and news recommendation [111, 148, 149]. Table 2 provides an overview of the length of news articles collected.

3.3 Dataset Quality Check

To validate our Twitter and News datasets, we randomly sampled 50 tweets and 20 news articles corresponding to each of the 30 mental health keywords. The first author hand annotated them as relevant or irrelevant, drawing upon her expertise and familiarity with social media content. The annotations were then discussed with the second author for agreement and consensus. It should be noted that for keywords with under 20 news articles the entire set was annotated for dataset validation. Table 1 provides precision scores indicating the proportion of relevant tweets and news articles in the random sample drawn for each keyword.

4 METHODS

4.1 Representation of Tweets and News Articles

The short length of tweets and the news headlines for some of the news articles may not allow sufficient linguistic context that could capture nuances in the moral and stigma framing of mental health discourse [8, 9]. Therefore, we first extracted the sentence level BERT embeddings to get vector representations for tweets and news articles. To capture the framing style and context of the textual content, we represented it using sentence level embeddings instead of averaged word embeddings, since the latter does not consider the relationships between words.

4.2 Representation of Moral Foundations

To understand moral framing of mental health on Twitter and News corresponding to our RQ1, we utilized the widely used and validated Moral Foundation dictionary [44]. It consists of words or phrases for each of the five moral foundations [59]: (1) Care/Harm, *underlying virtues of kindness, gentleness, and nurturance*; (2) Fairness/Cheating, *generating ideas of justice, rights, and autonomy*; (3) Loyalty/Betrayal, *encompassing virtues of patriotism and self-sacrifice for an identifying group*; (4) Authority/Subversion, *including virtues of leadership and followership, deference to legitimate authority and respect for traditions*; and (5) Sanctity/Degradation, *embodying religious notions of striving to live in an elevated, less carnal, more noble way*. There are two dictionaries for each moral foundation, representing its (1) virtue (e.g. Care) and (2) vice (e.g. Harm) facet. For instance, example words corresponding to the Care/Harm moral foundation include 'protect', 'compassion', 'consoled'; 'kill', 'threaten', 'destroy', while that for Sanctity/Degradation include 'sacred', 'purity', 'divine'; 'decay', 'sin', 'repulsive'.

We note that prior work has successfully utilized the above MFT dictionaries to understand a variety of human behaviors online, as described in Section 2.3. We used representational learning to augment purely dictionary based approaches, as these existing approaches often fail to capture linguistic context and nuanced writing style, both due to a reliance on exact matching of hand-curated corpus of fixed/limited words. Therefore, using the dictionaries mentioned above, we generated a representation for the positive and the negative facets of each moral foundation. In particular, we utilized a pre-trained large language model – BERT [28] – to generate word embedding representations for all the words present in a dictionary, along with their relevant synonyms extracted using WordNet [92], and then averaged them to create the final representation of each moral foundation dimension. For instance, through this computation we had two embedding vectors for the Care/Harm moral foundation, one for Care and the other for Harm.

Then, to obtain a single linguistic representation of a specific moral foundation, we followed the approach provided by Kwak et al. [80]. Kwak et al. [80]'s method is called FrameAxis, and it seeks to characterize language framing by introducing "microframes" that are essentially *semantic axes* [3, 127] or vector representations for two sets of antonymous words. The microframes are obtained by subtracting the embeddings of the two opposing poles. Existing work that explores moral framing of documents using MFT has adopted and validated the FrameAxis methodology to generate embedding representation for the five moral foundations [94, 113]. In other words, these works essentially subtracted the vectors for the virtue and vice facets of each moral foundation to obtain a single linguistic representation for the same. Drawing on these prior approaches, in our current study, the final representation for the Care/Harm moral foundation (and similarly others) was given by subtracting the negative dimension vector (using the vice dictionary) from the positive vector (using the virtue dictionary).

4.3 Approach to Compare Twitter and News

Next, to analyze the moral framing in tweets and news articles, we compared the vectors representing each tweet or news article against the five embedding vectors for the five moral foundations, obtained through the approach described in Section 4.2. This comparison was made using cosine similarity scoring, a standard practice in the use of large pre-trained embeddings [80, 94]. These scores range from -1 to 1, where a score closer to -1 represents that the tweet or news article aligns more with the negative dimension (vice) of the moral foundation and a score closer to 1 represents that the tweet or news article aligns more with the positive dimension (virtue). Then, we set a threshold of 0 cosine similarity (mid point of the [-1, 1] scale) to say that tweets or news articles with a score greater than 0 align with the positive (virtue) dimension of the foundation and those less than 0, with the negative (vice) dimension. Refer to Appendix A for details on validating our approach.

4.4 Operationalizing and Measuring Stigma

Recall that our second research question aims to assess the levels of stigma manifested in tweets and news articles, surrounding mental health discourse. Unlike the moral foundations, there are no available dictionaries or tools to operationalize and measure stigma. Consequently, here we describe an approach to do so.

Stigma is the prejudice and discrimination attached to a phenomena [22, 87]. Following the virtue-vice characterization of the five moral foundations in Section 4.2, we describe levels of stigma in language through an Approval/Stigma frame, "approval" being a commonly considered opposite of "stigma". Accordingly, we referred to existing literature on the conceptualization of stigma, along with stigmatized (stigma-challenging) framings in mass/social media to get seed keywords for stigma (approval) dimensions [50, 83, 105, 144]. These works have employed public health experts to perform an exploratory content analysis and identify stigmatizing/anti-stigmatizing words or phrases. We again used the WordNet tool to identify corresponding synonyms for the seed keywords, giving us one dictionary each for approval and for stigma. With these dictionaries, we computed a vector representation of the Approval/Stigma frame following the same procedure as used for the five moral foundations, discussed in Section 4.2. The final representation of the Approval/Stigma frame was obtained by subtracting the averaged negative dimension vector from the positive one. For the Approval/Stigma frame, refer to Appendix C for the keywords we curated in the respective Approval and Stigma dictionaries.

4.5 Human Evaluation

WordNet covers different contexts while extracting synonyms for a given input. For instance, the word 'kind' present in the dictionary for Care/Harm moral foundation added out of context synonyms such as 'sort', 'form', and 'variety'. To circumvent this issue, we manually inspected the expanded Moral Foundation and the new Approval/Stigma dictionaries to remove irrelevant words that got captured by WordNet. The Moral Foundation dictionaries were processed via qualitative evaluation by the paper's authors, referring the foundation definitions present in literature [51, 60] and until 100% consensus was reached on all words. For the evaluation of Approval/Stigma dictionaries, the authors referred to exemplars/definitions present in prior work on stigma [57, 105].

To further validate our Approval/Stigma dictionaries the first author manually annotated a random sample of 100 tweets and 50 news articles in our dataset, labeling their alignment with approval or stigma. The annotations were discussed for agreement and consensus with the second author. We then compared the approval/stigma labels assigned by our BERT-based framework, utilizing the curated Approval/Stigma dictionaries, against the hand-annotated ground truth labels. Our framework was able to achieve high precision (Twitter: 0.89; News: 0.92) and recall (Twitter: 0.92; News: 0.94) on the annotated random samples.

5 RESULTS

5.1 Differences in Moral Frames on Twitter and News

As mentioned in Section 4.3, to analyze the moral framing of mental health in tweets and news articles (RQ1), we used cosine similarity scoring to compare the vectors representing each tweet or news article against the five embedding vectors for the five foundations.

Table 4 provides raw distribution statistics of cosine similarity scores for both Twitter and News. Essentially, we compare how tweets and news articles align with the five moral foundations, i.e., whether they lie more towards the virtue side of a foundation or towards the vice side of a foundation. It can be observed that Twitter's mental health framings, across all the five moral foundations align more with the virtue facet, in terms of having a larger proportion of tweets that align with the positive dimension (virtue), having a cosine similarity score greater than 0. For instance, more than three-quarters of the tweets use Fairness related framings of mental health (75.73% tweets), while only a quarter use otherwise (24.27%). In other words, the ratio of the number of tweets falling towards the positive dimension, to the number of tweets falling towards the negative dimension is greater than 1 for all the five moral foundations; it is as high as 11.67 for the Care/Harm moral foundation, and at a minimum almost twice (1.94) for the Authority/Subversion dimension. On the other hand, the news articles skew more towards the negative or the vice dimension. For instance, for the Loyalty/Betrayal moral foundation, the ratio is only 0.14, indicating that far more news articles (87.95%) demonstrate the vice facet of this foundation in their writing, rather than the virtuous one (12.05% articles only). Further, as indicated in Table 4, Mann-Whitney U-tests on the cosine score distributions for Twitter and News show significant differences across all the foundations.

A summary of the above patterns is captured in the frequency distribution plots shown in Fig. 1. The histogram for Twitter is shifted towards the right of that for News, indicating that Twitter uses more virtue oriented moral framings, compared to News.

Consider the following exemplars that received a highly positive/negative score for three of the five moral foundations. The tweet below, which aligns with the virtue facet of Care/Harm foundation, tries to nurture a positive environment by rejecting dismissive characteristics associated with depression:

> "Depression has nothing to do with being weak or lazy." (Care: Paraphrased tweet)

On the other hand, the following news article excerpt that skewed towards the vice facet of Care/Harm foundation generalizes individuals with a mental disorder to get penalized for harmful behavior:

> "Today, mentally ill [...] are likely to be arrested, incarcerated, suffer solitary confinement or rape in prison and commit another crime once released. [...] people

Table 4: Distribution statistics of tweets and news articles based on setting a threshold on 0 cosine similarity score. +ve (-ve) represents the tweets or news articles that have a cosine similarity score greater (lesser) than 0 (the selected threshold value), indicating positive/virtue and negative/vice alignment to the corresponding moral foundation. Ratio indicates the ratio between the number of tweets/news articles that align with the virtue facet of a moral foundation (e.g., Care) to those that align with the vice facet of the same moral foundation (e.g., Harm). Mann-Whitney U-tests were performed to compare the cosine similarity score distributions for Twitter and News across the five moral foundations ($p < 0.1$: '*', $p < 0.05$: '', $p < 0.01$: '***').**

Moral Foundation	**Twitter**			**News**		
	Count (+ve; -ve)	% (+ve; -ve)	*Ratio* (+ve/-ve)	*Count* (+ve; -ve)	% (+ve; -ve)	*Ratio* (+ve/-ve)
Care/Harm**	12229522; 1047593	92.11; 7.89	11.674	7383; 13784	34.88; 65.12	0.536
Fairness/Cheating**	10055374; 3221741	75.73; 24.27	3.121	4740; 16427	22.39; 77.61	0.289
Loyalty/Betrayal***	10537975; 2739140	79.37; 20.63	3.847	2551; 18616	12.05; 87.95	0.137
Authority/Subversion***	8752996; 4524119	65.93; 34.07	1.935	1295; 19872	6.12; 93.88	0.065
Sanctity/Degradation**	9484552; 3792563	71.44; 28.56	2.501	3766; 17401	17.79; 82.21	0.216

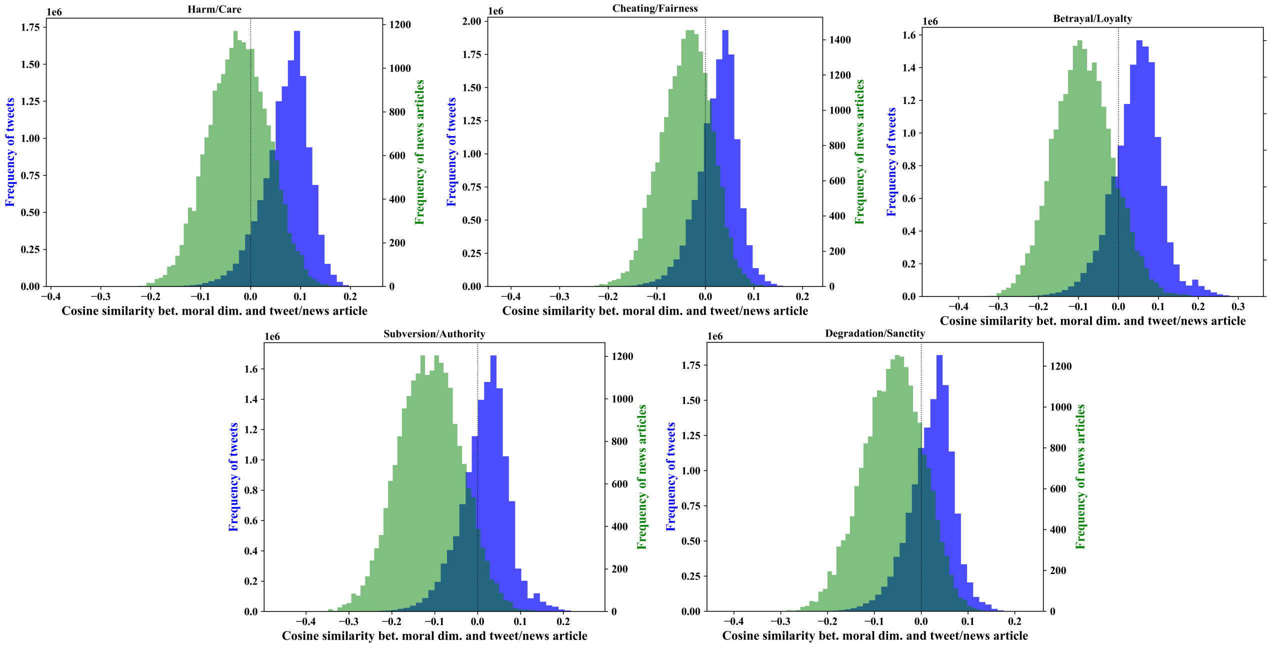


Figure 1: Distribution of cosine similarity scores for tweets (left vertical or Y-axis) and news articles (right vertical or Y-axis) across the five moral foundations. The dotted line at 0 cosine similarity represents the threshold used for defining the alignment of textual content towards the positive (virtue) or negative (vice) dimension of a foundation.

> in federal prison have a history of mental disorder." (Harm: Paraphrased news article excerpt)

As mentioned earlier, this observation is in line with previous research on newspaper framings that found the usage of "dangerousness" as an attribute in stories on mental illness [144].

For the Fairness/Cheating moral foundation, the following news headline aligns with the virtue facet and displays values of justice and equality while discussing mental health.

> "True 'parity' in mental health requires change in attitude." (Fairness: Paraphrased news headline)

The following tweet, associated with Cheating dimension, describes those with mental disorder to receive unwarranted benefits.

> "People with a mental disorder have an unfair advantage. They can use it as an excuse to get out of things." (Cheating: Paraphrased tweet)

Lastly, the following tweet that aligns with the Loyalty dimension tends to highlight in-group framings and encourages people to foster an inclusive community for suicidal individuals.

> "Please let us not dismiss people who are suicidal. Let us be there for them, talk to them, check in on them." (Loyalty: Paraphrased tweet)

Conversely, tweet aligning with betrayal "others" people with schizophrenia, differentiating them from "sane" individuals.

> "[City Name] is crawling with mentally ill people who are schizophrenics, instead of sending them to mental

> health hospitals. All attacks are done by someone who is clearly not a sane person, be it drugs or schizophrenia." (Betrayal: Paraphrased tweet)

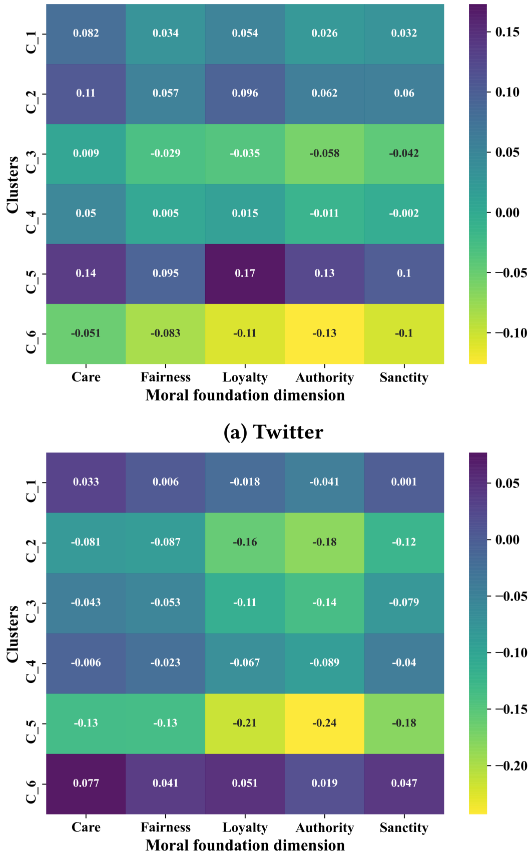
Table 5: Interaction between different moral foundations after performing clustering on the Twitter and News datasets. The moral foundation interactions (low, balanced, or high) are defined by dividing the color axis shown in Fig. 2 into score ranges. For instance, for Twitter we have < -0.05: low, $(-0.05, 0)$: moderately low, $(0, 0.05)$: balanced, $(0.05, 0.10)$: moderately high, and > 0.10: high. Statistical analysis using Kruskal-Wallis H-tests reveal significant differences across the score distributions for the five moral foundations in each cluster ($p < 0.05$: '*', $p < 0.01$: '', $p < 0.001$: '***').**

Cluster	Score range	Interaction between moral foundations
Twitter		
C_1*	$[0.03, 0.08]$	moderately high care, rest balanced
C_2*	$[0.06, 0.11]$	moderately high for all foundations
C_3**	$[-0.06, 0.01]$	low authority, balanced care
C_4*	$[-0.01, 0.05]$	balanced for all foundations
C_5**	$[0.10, 0.17]$	high for care and loyalty
C_6**	$[-0.13, -0.05]$	low for all foundations
News		
C_1*	$[-0.04, 0.03]$	high care, rest moderately high
C_2**	$[-0.18, -0.08]$	low loyalty and authority
C_3***	$[-0.14, -0.04]$	varying across foundations
C_4**	$[-0.09, -0.01]$	balanced loyalty and authority
C_5**	$[-0.24, -0.13]$	low for all foundations
C_6*	$[0.02, 0.08]$	high for all foundations

5.2 Interpreting the Moral Frames

To further explore the varied forms of moral framing of mental health discourse on Twitter and News, we performed a clustering analysis. In order to cluster the tweets and the news articles we represented an individual data point (tweet or news article) using a five element feature vector, having the cosine similarity score of a tweet or news article against each of the five moral foundations. Once we had these feature vectors, we used the k-means clustering algorithm to cluster the data points in our Twitter and News datasets. To get the optimal number of clusters that we should use to set the value of k in k-means clustering we studied the Elbow curve (using distortion score for different values of k or number of clusters) and the Silhouette curve (using silhouette score for different values of k). In our data, the optimal number of clusters was found to be 6 for k-means clustering, for both Twitter and News.

After obtaining the clusters, we averaged the cosine similarity scores for tweets or news articles belonging to one particular cluster for each of the five moral foundations. Using this, we obtained the heatmap distribution for Twitter and News, shown in Fig. 2. We also performed a statistical analysis using the Kruskal-Wallis H-tests [77] to explore differences in cosine similarity score distributions across the five moral foundations, for each of the six clusters. This analysis revealed significant differences, as summarized in Table 5.



(a) Twitter

(b) News

Figure 2: Heatmap distribution of the level of virtue/vice corresponding to each of the five moral foundations. Distribution shown over the clusters of (a) tweets and (b) news articles obtained via the k-means algorithm.

Through these heatmaps we can explore interactions between different moral foundations, as tabulated in Table 5 for Twitter and News. Using these, we can see how different moral foundations appear when we frame tweets or news articles on mental health. For instance, we see that for both Twitter and News, clusters exist for which tweets and news articles score high for the Care/Harm and Loyalty/Betrayal foundations (C_5 for Twitter and C_6 for News), that is, in these clusters, the tweets and articles tend to frame mental health via the virtuous facet of Care/Harm as well as that of Loyalty/Betrayal. Take the following exemplars within these clusters. Kindness and gentleness are a hallmark of the virtue facet of Care/Harm, and we note that the tweet below speaks in inclusive language, expressing solidarity and support for those suffering from depression. Similarly, the news article headline and excerpt below speaks of veteran suicide in an empathetic light and recognizes a particular veteran who died by suicide for his contributions towards protecting his country – a framing considered virtuous within the Loyalty/Betrayal moral foundation.

> "As someone with a history of depression, I hope you're all well. Please don't forget to take time for yourself. Productivity is great, but so is treating your body and mind right. Just a reminder to take care of your mental health always." (Paraphrased tweet)

> "Veteran commits suicide in parking lot of [City Name]: Veteran [...] killed himself [...] after reportedly being turned away for emergency care. [...] His work for his fellow veterans might not be finished yet. "Your death is not in vain. Through your tragedy, may the bureaucrats change policies to help others that were in your situation," read one message left on the funeral home's website, quoted by the [City Name] press." (Paraphrased news headline/article excerpt)

Next, we observe that there are clusters that have elements (tweets or news articles) scored low across all the five moral foundations (C_6 for Twitter and C_5 for News), meaning they frame mental health using the vice facet of all the moral foundations. Exemplars are given below. In the following tweet, individuals with mental illness are accused to have a high propensity to commit criminal activities and the author implicitly assumes that such individuals are violent and aggressive, and therefore should be feared or avoided. Such a framing espouses Sanctity/Degradation, whose vice facet implies that "the body is a temple which can be desecrated by immoral activities and contaminants" [59] – in this case, due to an individual's underlying mental illness. The framing also maps with the vice facet of Care/Harm, as individuals with mental illnesses are "othered" for their presumed dangerous behaviors. Similar observations can be made for the news excerpt below, which ascribes unpredictable and antisocial attributes to depressed individuals. This framing shows disregard and apathy towards the structural factors that may trigger a suicide attempt, thus aligning with the vice facet of Care/Harm. Further, an inability to treat suicide attempters with altruism, compassion, or with a restorative justice approach seems to be in alignment with the vice facet of the Fairness/Cheating moral foundation.

> "People with schizophrenia commit rape, are mentally weak and disrespect the law." (Paraphrased tweet)

> "Strange seasonality of violence: Why April is 'the beginning of the killing season' [...] owing to those who are depressed, suicidal [...] increase in sunlight improves mood and energy just enough for suicidal people to make plans and follow through. [. . .] their rage may build as they see people out having fun together in groups. [...] That highlights discrepancies between those who are socially healthy and those who aren't." (Paraphrased news headline/article excerpt)

These similarities across Twitter and News reflect on how an individual may approach writing textual content. It is likely that a tweet or news article aligning more with the vice facet of one foundation may also negatively relate to the others, and vice versa.

5.3 Comparing Stigma in Twitter and News

Next, corresponding to RQ2, to analyze the moral framing of tweets and news articles against the Approval/Stigma dimension, we computed the cosine similarity scores, similar to the method adopted for the moral foundations. We still used 0 cosine similarity score as a threshold to determine whether a tweet or news article aligns more with the positive (> 0) or the negative facet of Approval/Stigma.

We observed that 5,528,120 (7,748,995) tweets aligned with the positive (negative) dimension of the proposed Approval/Stigma frame. This shows that although Twitter discourse on mental health was largely around virtuous framing across all the five moral foundations, it was not the case for the Approval/Stigma dimension. A larger proportion of tweets aligned more with the negative Stigma facet than the positive Approval facet, indicating the discourse to lean more around stigmatizing frames rather than ones that were more inclusive or accepting of mental illness. A similar trend can also be observed for the news articles where 3,011 (18,156) news articles aligned with the positive (negative) facet of the new Approval/Stigma moral foundation. Still, upon comparing mental health discourse on Twitter and News, we find that the ratio of tweets aligning with Approval to tweets aligning with Stigma (0.713) is higher than the ratio of news articles aligning with Approval to those aligning with Stigma (0.166), suggesting that Twitter, like the above five moral foundations, uses more positive framings supporting acceptance of mental health challenges, rather denouncing, shaming, or viewing mental health experiences with dishonor.

5.4 Establishing Associations between Moral Foundations and Stigma

In response to RQ3, to examine how the five moral foundations relate to the proposed Approval/Stigma dimension in framing of mental health discourse, as shown in Fig. 3, we plot the cosine similarity scores for the Approval/Stigma foundation against all the other five moral foundations. In the figure, each blue dot located at a (x, y) coordinate corresponds to a tweet or a news article such that, x (y) represents the cosine similarity score for the tweet/news article with a moral foundation (Approval/Stigma dimension). Looking at the plots it can be clearly said that there is a positive correlation between the scores for the Approval/Stigma dimension and each of the five moral foundations. A linear model fit to each of the scatter plots also indicates that the relationship between the Approval/Stigma frame and the other five foundations is very similar across Twitter and News (captured through the slope of the regression line). In simple terms, it implies that when a tweet or a news article discusses mental health, if the underlying moral foundations are framed more virtuously, then the tone is also indicative of a more approving framing, rather than a stigmatizing one.

Slopes of the linear regression fits, represented by the coefficient of x in Fig. 3, provide an indication of the relationship strength between the Approval/Stigma frame and the five moral foundations. For Twitter, Care/Harm (0.65) and Fairness/Cheating (0.66) share the strongest positive association with Approval/Stigma frame. In contrast, Loyalty/Betrayal (0.38) foundation has the least positive association. Similarly, for News, Fairness/Cheating (0.76) has the highest and Loyalty/Betrayal (0.38) has the least positive relationship with Approval/Stigma frame. This implies that, for both Twitter

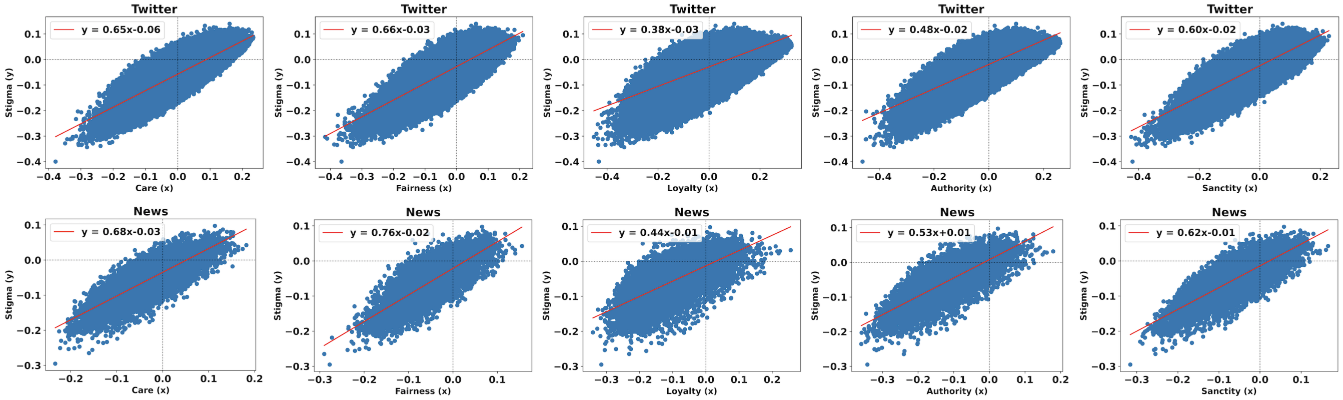


Figure 3: Scatter plots to capture the relationship between the cosine similarity scores of the Approval/Stigma dimension and the five moral foundations. The red lines indicate the regression line or the best linear fit for the distributions.

and News, a virtuous Fairness/Cheating (Loyalty/Betrayal) framing is the most (least) indicative of an approving framing.

Below we provide examples that associate virtuous facets of Care/Harm and Fairness/Cheating with approving framings. In the following tweet, the author disapproves of those who are dismissive of or deride mental health challenges faced by others, as it can be a barrier to treatment and help seeking. The author also advocates candid discussions about the issue, as a way of showing support, solidarity, and justice. This framing not only uses virtuous Care/Harm and Fairness/Cheating framings, but its suggestions are well situated in the stigma literature as well [23]. Complementarily, the news excerpt below attempts to normalize teen mental health challenges and calls for an inclusive, ecological approach involving others as a way to tackle the challenge. This excerpt thus not only uses a positive framing using Care/Harm and Fairness/Cheating, but also champions educating and making parents and clinicians more aware as a means of destigmatizing teen mental illness. In both the examples there are strong indications of the approval (or the anti-stigma) frame such as, 'exacerbate the stigma' and 'not associate depression with adolescence'.

> "When you mock those who talk openly about their mental health struggles, you exacerbate the stigma surrounding it for others. This stigma discourages people from asking for the help and support they need. Mental health isn't a joke. No one should have to suffer in silence." (Paraphrased tweet)

> "Is a Teen Depressed, or Just Moody?: [...] "When it comes to your child, statistics don't matter, what matters is your particular child," he said. "Pay attention to worry signs." [...] Electronic media usage is not a cause of depression [...] that's how they connect to their peer group, that's how they get their support [...], the message to parents and pediatricians is that we should not associate depression with adolescence or substance use and make them feel equal." (Paraphrased news headline/article excerpt)

Next, we include examples that associate virtuous facets of Loyalty/Betrayal foundation with approving framings. Both the tweet and news excerpt use language indicative of the virtuous Loyalty/Betrayal facet. They make use of solidarity- and empathy-oriented phrases such as "make them feel a part of the community", "be more empathetic and inclusive", and "more alike than different". However, unlike the examples presented above, they do not explicitly question mental health stigma. The approval framings are subtle, advocating support ("be there for them") and acknowledging hardships and life struggles, often attributed to the experience of mental illness ("gone through a lot").

> "We should not force a depressed person into saying they are not depressed. That's easy. Instead, we should be there for them and make them feel a part of the community." (Paraphrased tweet)

> "[...] came back with PTSD. They have gone through a lot, seen death one after the other. [...] A lot more can be done. We can be more empathetic and inclusive when we realise how much more alike we are than different. (Paraphrased news headline/article excerpt)

In contrast, vice oriented framings of mental health on Twitter and News also use a stigmatizing tone, as given in the pair of examples below. Both the tweet and the news excerpt use the vice framing of Sanctity/Degradation by attributing moral contamination as causes of mental illness and suicide. The stigma literature notes such framing to map to prejudiced and hostile attitudes towards the sufferers of mental illness, which in turn can impose unfair burdens on a group who are already at social disadvantage [23].

> "Such a chaotic generation. Standing on no morals while they scroll TikTok all day wondering why they can't find happiness. They're depressed because they are inherently weak and impure." (Paraphrased tweet)

> "suicidal thoughts are 'a big sign of mental weakness' [...] admission of suicidal ideation [...] is a sign of 'mental weakness' [...] no one should live life in fear and impose that fear on other people." (Paraphrased news headline/article excerpt)

Lastly, we found tweets with virtuous but stigmatized framings of mental health (essentially blue dots in the bottom right quadrants of Fig. 3). In the following tweet, though the author uses positive framings of the Care/Harm foundation while introducing the subject, they use a stigmatized attitude while describing the subject's mental illness such as "threatened" and "snap" [112].

> "Despite having BPD [...] was funny, productive, kind, and caring mom. Her mental illness threatened it all. [...] catastrophize and snap. (Paraphrased tweet)"

6 DISCUSSION

This research was able to reveal significant differences between Twitter and News in terms of framings of mental health discourse, analyzed quantitatively using a framework inspired by the Moral Foundation Theory. We found that tweets had a greater tendency to align with the positive (virtue) facet of all the five moral foundations compared to news articles. And then, somewhat alarmingly, when tweets and news articles used negative moral framings of mental health topics, they also tended to use stigmatizing language. Presence of stigmatized language framings in our work forms a parallel with existing research on news media and Twitter in the context of mental disorders, as noted earlier in the works of Joseph et al. [71], and Gwarjanski and Parrott [57]. This indicates that, while negative moral framings on Twitter might be on the minority, they are still present and engender other problematic framings.

Overall, this work innovated by providing a first study that uncovered the moral underpinnings and their relationship to stigma within conventional and social media discourse around mental health. As also noted in Section 2.3, although MFT has been appropriated to study a wide variety of sociopolitical phenomena online [27, 55, 69, 116], to our knowledge, this theory had not been utilized in the context of mental health as yet, a gap that is important to fill, because of how media narratives shape formal and informal caring (or the lack thereof), for those with mental health struggles [150]. Further, this theory has also not been harnessed to characterize the nature of stigma; again a gap, when filled, could lead to improved media guidelines espousing the values of acceptance and inclusion of people with mental illness in the broader society. Finally, our computational approach as well as the lexical resource we developed on stigma, together could be adopted in other health research to understand degrees and nature of marginalization of particular medically disenfranchised groups, or even to design interventions to mitigate stigma in popular discourse.

This Discussion section, accordingly, seeks to unpack what might contribute to the specific moral or immoral framings of mental health we observed in the two forms of media, the implications of these observations, and then some potential directions to mitigate them through technology-mediated means.

6.1 What Might Explain the Specific Mental Health Framings on Twitter and News?

In this subsection, we discuss what factors might be driving the platform-specific characteristics of mental health framings we observed on social media (driven by general public) and news (driven by professional journalists). We anticipate that Twitter's platform affordances, in particular its broadcasting nature, could be responsible for users to employ virtuous internal attitudes while framing mental health discourse. On the other hand, news media sensationalism [95] and political leanings may influence journalistic moral underpinnings, resulting in vice-oriented mental health framings.

6.1.1 Twitter: Technical Affordances and Shifting Norms. Twitter is a micro-blogging social media platform where individuals tend to broadcast their thoughts and opinions publicly, and therefore, their posts often tend to have a wide reach among diverse and even "imagined" audiences [84]. Scholars like Marwick and Boyd [88] have used imagined audiences as a concept to understand social media users' decision-making behind what is appropriate and relevant to share when they recognize that it is impossible to know who the actual audience is. As noted by scholars [141], this affordance of the platform encourages individuals to be more cautious or aware of their engagement. Context collapse on a public platform [88] is likely to espouse additional caution on the part of social media authors. This is because users may find it challenging to write in a way that is able to attend to the varied views and stances of so many different people in their social network [32]. Together, these conscious choices may be contributing to the patterns of framings we see on Twitter, i.e., having a high proportion of tweets aligning with the virtue facets of the moral foundations.

In addition, social media users often adopt a performative "lowest common denominator" [65], due to the "spiral of silence," [61] as well as the emergent "cancel culture" [104]. Mental health being a sensitive topic, can trigger contentious views and the topic is often embroiled in politically thorny issues in the U.S., such as gun control [125]. After all, Benning [10] noted how renowned psychiatrist and iconoclast critic, Thomas Szasz, wrote that only physical illnesses are real and that mental diseases are "counterfeit and metaphorical illnesses" [132]. As such, discussions of mental health issues on social media may portray a virtuous morality, as observed in this study, as a way to avoid unpopular opinions, potential confrontation, or unpleasant social exchanges.

Various underlying platform norms of Twitter may further motivate users to share their perspectives on a sensitive issue like mental health with inclusive, kind, and equity-driven moral values, as our findings revealed. Stupinski [130] already noted that the term "mental health" had risen in its frequency of use between 2012 and 2018. These may indicate shifting norms and increasing acceptability in people's attitudes toward mental health on Twitter. Moreover, multiple research papers have reported a normalization of mental health discourse on Twitter since the inception of the COVID-19 pandemic [68]. Social media platforms emerged as prominent platforms for these discussions, as physical distancing and lockdown policies made many individuals hunker down in their homes, taking away opportunities for in-person interactions [117]. The time period of our analysis overlaps with the first two years of the pandemic. Hence, this study reflects moral framings on the Twitter platform at a time when attitudes towards mental health are already evolving to have compassion and inclusion.

6.1.2 News: Clickbait Journalism and Political Bias. Since the past several decades, journalistic training has encompassed adequate accommodations for inclusive and non-stigmatizing mental health reporting. In particular, journalists are usually guided by a toolbox consisting of a set of guidelines that they should follow while reporting on mental health conditions [63, 136]. For instance, 'Achieving the balance' [63] is a resource kit that is utilized by Australian professionals in suicide and mental health reporting. Similarly, the Carter Center has a journalism resource guide on behavioral health [136]. Still, our findings show that news articles align more with negative moral framings of mental illness, contributing complementary evidence of harmful framings as observed by other researchers [131].

There might be a variety of reasons behind the findings pertaining to framing of mental health on news media. In recent years, traditional news media has struggled to stay relevant especially among younger, urban populations, facing stiff competition from social media [123]. Many news media have been obliged to embrace web platforms, using online advertising as the primary revenue model. Within such a model, in a competition for more eyeballs, many journalists and news agencies are relying on "clickbait" tactics [96]. In our study, clickbait attempts to fit within the limited attention spans of many readers may indicate insensitive framings of mental health. After all, Armstrong et al. [5], in interviews with media professionals reporting on suicide found "It's a battle for eyeballs and suicide is clickbait." What is alarming is, as our results in Section 5 suggest, such tactics may result in demonstrating moral foundations that perpetuate othering and stigma as well, such as that people with schizophrenia are "dangerous" and "violent", or that people with depression are just "not trying hard enough."

A yet another reason behind the negative moral framing of mental health on News could stem from particular media house's underlying political ideology. News platforms, from those covering legitimate news to those using viral and yellow journalism are trying their best to cover narratives that their (polarized) audience expects to see [140]. But can this political bias explain the negative moral framings of mental health in news media? Recently, Munsch et al. [101] showed that liberalism and conservatism are associated with qualitatively different psychological concerns, notably those linked to morality. Strupp-Levitsky et al. [128] complementarily found that political conservatives align more with the "binding" moral foundations (in-group loyalty, respect for authority, and purity) which, in turn, are associated with epistemic and existential needs to reduce uncertainty, threat, and system justification tendencies. In contrast, liberals demonstrate the "individualizing" foundations (fairness and avoidance of harm), which are more associated with empathic motivation. Our findings show that the binding foundations fare more prominently in news articles, bringing about a less compassionate framing of mental health. This could reflect the presence of more conservative-leaning news articles in our corpus, an aspect that may be explored more thoroughly in future research.

6.2 Implications of Viceful and Stigmatizing Framings on Marginalization

Our findings offer important implications that are relevant to understand media trust, or the lack thereof, among marginalized communities. Scholars have defined marginalization as a "lack of integration and the status as an 'outsider'; with respect to dominant cultures" [11]. Marginalized communities, including those with mental illness, are confronted with issues resulting from their social identity, such as exclusion, invisibility, misrepresentation, and hate speech, not just in offline contexts, but increasingly so, online [26, 103, 144]. This is because while topics representing their voices and needs may be present in their social bubbles, the broader media narratives are perceived to fuel discrimination, and other systemic barriers and biases [36]. When mental health issues, on news or in social media, are framed with viceful moral foundations and exhibit stigmatizing stances, as our results showed, such discourse may further erode media trust. Such skepticism has been connected to medical disenfranchisement, compounding access disparities for people with mental health struggles [64]. Gibney [48] noted a "crisis of trust" that emerged particularly during the COVID-19 pandemic, deepening beliefs that those marginalized in mental health care are not "going to be treated fairly and equally in the healthcare system, and that [...] the system is not only flawed but actively out to get them." Consequently, the negative stereotypes and unfounded fears based on misperceptions included in the paraphrased exemplars of tweets and news article excerpts provided in Section 5, risk leading to resistance to community based treatment programs, as well as underfunding of mental health research and facilities.

Related to the issue of media skepticism is the issue of media framings shaping consumers' mental models of news sources. We have witnessed a remarkable increase in distrust of public institutions [133]. Legacy media is often perceived to be part of the overall social and political elite formation [118], adopting an approach that abandons the less advantaged and promotes the interests of the elite, in turn, introducing a new element to traditional political populism, both from right-wing and left-radical factions [40]. Negative moral framings on media, as observed in our work, are likely to call a media outlet's reliability and impartiality into question, aligning with mental models that deepen cynicism about social and political institutions, including mental health ones on psychiatric treatment and institutionalized care. In the following subsection, we suggest possible approaches to address these implications.

6.3 Recommendations to Support Improved Mental Health Framing

In the past few decades, there have been remarkable attempts to raise mental health awareness and abate stigma through a variety of public health campaigns, such as through advocacy work, giving face to the struggle of mental illness, celebrity disclosures, or by tapping into public psyche via social influencers [129]. And still, despite these efforts, our findings reveal a mismatch between how people think about mental health – as indicated in tweets and news articles – and what public health messaging on mental health has been striving for. We argue that these findings call for more work to be done to change the current framings of mental health in the public consciousness.

To prevent morally unsound and stigmatized reporting from harming communities marginalized by mental illness, media professionals should uphold their ethical duty and defend the rights of these populations, as well as speak with truth and compassion. Specifically, our findings may be utilized to underscore that journalistic resource guides are further strengthened, taking inspiration from the social psychology literature, like the MFT, to promote improved strategies for reporting mental health conditions. There is evidence that modification of reporting on suicidal behavior is feasible and can be effective [62]. Several studies have measured the style of media reporting about suicide before and after recommendations for media were launched [45] or before and after trainings for editors and journalists were given [134]. This indicates that adapting recommendations for improved moral framing of mental health on news media can have practical positive outcomes on the ground. Journalists are already suggested to prevent the use of derogatory language in writing as it contributes to the negative attitudes about mental illness that keep people from seeking treatment [131]. However, these guidelines do not foreground the need to consider the writer's underlying moral stance. Since our findings show that a lack of awareness or reflection of morality can aggrandize stigma, we suggest journalists to use a Care, Fairness, or Sanctity based moral frame that prevents narratives linking mental illness with violence or portraying people with mental health problems as dangerous, criminal, evil, or disabled and unable to live responsible, fulfilling lives. Choosing to center news framing on these moral foundations will also ensure that value-neutral terms are used and over-simplification of circumstances is prevented, such as preferring to say "she had struggles managing work and school, ever since she was diagnosed with an anxiety disorder," instead of "she has missed work and school because she suffers from anxiety." Our findings also call for new constructive partnerships between media professionals and mental health experts – a direction that echoes views of other scholars [5]. Additionally, with the emergence of *computational* or *automated* journalism, there is an increased emphasis on adopting practices to ensure accurate, impartial, and transparent reporting [1]. Some suggested ways are to publicise curation mechanisms [30] and explore algorithmic accountability [4]. These strategies are shown to evoke normative arguments and enhance emotional engagement [30], aspects that fit well with exacerbating stigma in news media mental health reporting.

Moreover, the pervasive presence of stigma calls for relevant resource guides that could be useful to both the general public (on social media) and the professional journalists (on news media) and can help to reduce negative stereotypical framings associated with mental disorders and be more aware of the social consequences of specific framings. AI-mediated communication in healthcare is already supporting a variety of tasks that allow individuals to seek and receive help, help medical professionals connect with patients, and provide therapeutic outcomes for treating mental illness [17, 47, 121]. Framing guidelines for mental health will need to permeate the design of these tools, so that the end users can receive appropriate algorithmic recommendations or behavioral nudges on moral framing, while expressing their views or reporting information on mental health. Nudging is a well-researched and well-practiced approach in behavior change [142], and in recent social computing research, algorithmic nudging has been suggested as a viable way to build more functional online support communities, such as by providing suggestions for adequate linguistic accommodation to a community's norms [122] or recommending writing styles conducive to self-disclosure [151]. Adopting similar techniques, in our work, algorithmic nudging may imply providing (near real-time) suggestions for virtuous moral foundation based alterations to text that includes a particular negative moral framing, thus also potentially facilitating the subconscious learning of writing in a destigmatizing manner. Specifically, such alterations could follow recent works on empathic re-writing [121] – transforming low empathy content to higher empathy – to support better online mental health conversations. However, such algorithmic adjustments warrant careful formulation, for instance by referring to theoretical models in psychotherapy research [24, 139], to avoid introduction of further negative attitudes and information distrust. Finally, with emerging research around conversational AI to support technology-mediated interactions, such guidelines will need to be considered even more carefully in the design of chatbots, as they are increasingly advocated for mental health [97].

Next, the strong association between negative moral framings and stigma in both Twitter and News not only bears implications for better reporting guidelines as elaborated above, but also opens up opportunities for platforms to empower and educate users to use more morally uplifting, destigmatizing language. They can surface pointers to educational resources and information campaigns, such as from the National Alliance on Mental Illness [114], to make users more aware and thoughtful in their mental health discourse. Social media platforms may even algorithmically promote content about campaigns of advocacy groups that seek to change the way people think and act about mental health challenges. Such efforts will need to be harmonious with recent efforts by platforms like Twitter to support "healthy conversations"[2]. For instance, without silence speech altogether, stigmatizing frames may be algorithmically de-emphasized, while inclusive frames may be prioritized in users' timelines. This way, mental health disclosers would feel that there is no shame in talking about their situation, and that social media platforms are indeed the "safe spaces" [155] they aspire to be.

6.4 Limitations and Future Work

We note some limitations in this research. First, although we look at various widely used publicly available news datasets, there is an absence of relevant datasets or news article repositories specific to mental health. Future research could further expand our dataset with more relevant news articles. Second, our current work only looks at the moral foundation based framing of the original tweets and news articles. It would be interesting to see how the framing of original posts influences the audience by studying their reactions shared on the respective platforms, or if and how specific moral or (de)-stigmatizing framings spread in the social network. This extension could further strengthen the motivation of our research by seeing whether or not certain language framings have an impact on consumers' perception. Finally, this work only looks at one social media platform, Twitter. In future, it would be insightful to see whether or not the current findings hold when this is extended to other social media platforms like Facebook or Snapchat that are close-knit, perhaps motivating people to not be too cautious of the content they post due to reaching a more homogeneous audience.

7 CONCLUSION

Problematic mental health framing in mass media has widespread negative impacts – it can exacerbate discrimination and prejudice that can make it difficult for individuals to admit to taking and benefiting from treatment. Through this work, we sought to understand the moral foundations of the general public and journalists surrounding mental health discourse on social and news media.

[2] https://about.twitter.com/en/our-priorities/healthy-conversations

Driven by a BERT-based embedding framework that was used to score Twitter posts and news articles against the five moral foundations within the Moral Foundations Theory, we found notable differences between the mental health moral framings on the two platforms. With a newly introduced language representation based lexical resource for Approval/Stigma, we further discovered that although tweets espoused more compassionate and justice-oriented moral values compared to news articles, stigmatizing framing was widely prevalent on both sources. In fact, when tweets or news articles wrote with more vicious morality, they amplified stigmatizing perspectives as well. Our research contributes to designing and augmenting safe reporting guidelines for mental health in mass media, and offers design opportunities to social media platforms to facilitate "healthy" and inclusive mental health conversations.

ACKNOWLEDGMENTS

Mittal and De Choudhury were partly supported through NIH grant R01MH117172 to De Choudhury. We thank the members of the Social Dynamics and Well-Being Lab at Georgia Institute of Technology for their guidance. We are grateful to the reviewers for their valuable feedback on this research.

REFERENCES

[1] Tanja Aitamurto, Mike Ananny, Chris W. Anderson, Larry Birnbaum, Nicholas Diakopoulos, Matilda Hanson, Jessica Hullman, and Nick Ritchie. 2019. HCI for Accurate, Impartial and Transparent Journalism: Challenges and Solutions. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3290607.3299007

[2] Leonidas Akritidis, Athanasios Fevgas, Panayiotis Bozanis, and Miltiadis Alamaniotis. 2019. A Self-Pruning Classification Model for News. In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*. 1–6. https://doi.org/10.1109/IISA.2019.8900751

[3] Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. 2018. SemAxis: A Lightweight Framework to Characterize Domain-Specific Word Semantics Beyond Sentiment. https://doi.org/10.48550/ARXIV.1806.05521

[4] Mike Ananny and Kate Crawford. 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society* 20, 3 (2018), 973–989. https://doi.org/10.1177/1461444816676645

[5] Gregory Armstrong, Lakshmi Vijayakumar, Anish V Cherian, and Kannan Krishnaswamy. 2020. "It's a battle for eyeballs and suicide is clickbait": The media experience of suicide reporting in India. *PloS one* 15, 9 (2020), e0239280.

[6] Albert Bandura. 2009. Social cognitive theory of mass communication. In *Media effects*. Routledge, 110–140.

[7] Susan Beaton, Peter Forster, and Myfanwy Maple. 2013. Suicide and language: Why we shouldn't use the 'C' word. *InPsych* (2013).

[8] Rami Belkaroui and Rim Faiz. 2015. Towards Events Tweet Contextualization Using Social Influence Model and Users Conversations. In *Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics* (Larnaca, Cyprus) *(WIMS '15)*. Association for Computing Machinery, New York, NY, USA, Article 3, 9 pages. https://doi.org/10.1145/2797115.2797134

[9] Rami Belkaroui and Rim Faiz. 2017. Conversational based method for tweet contextualization. *Vietnam Journal of Computer Science* 4, 4 (01 Nov 2017), 223–232. https://doi.org/10.1007/s40595-016-0092-y

[10] Tony B Benning. 2016. No such thing as mental illness? Critical reflections on the major ideas and legacy of Thomas Szasz. *BJPsych bulletin* 40, 6 (2016), 292–295.

[11] Matthias Berndt and Laura Colini. 2013. Exclusion, marginalization and peripheralization. *Leibniz Institute for Regional Development and Structural Planning, Berlin (Working Paper* 49 (2013).

[12] Michael L Birnbaum, Sindhu Kiranmai Ernala, Asra F Rizvi, Munmun De Choudhury, and John M Kane. 2017. A Collaborative Approach to Identifying Social Media Markers of Schizophrenia by Employing Machine Learning and Clinical Appraisals. *J Med Internet Res* 19, 8 (14 Aug 2017), e289. https://doi.org/10.2196/jmir.7956

[13] Brian J. Bowe. 2018. Permitted to Build? Moral Foundations in Newspaper Framing of Mosque-Construction Controversies. *Journalism & Mass Communication Quarterly* 95, 3 (2018), 782–810. https://doi.org/10.1177/1077699017709253

[14] Brian J. Bowe and Jennifer Hoewe. 2016. Night and Day: An Illustration of Framing and Moral Foundations in the Oklahoma Shariah Amendment Campaign. *Journalism & Mass Communication Quarterly* 93, 4 (2016), 967–985. https://doi.org/10.1177/1077699016628806

[15] Alexandra Budenz, Ann Klassen, Jonathan Purtle, Elad Yom Tov, Michael Yudell, and Philip Massey. 2020. Mental illness and bipolar disorder on Twitter: implications for stigma and social support. *Journal of Mental Health* 29, 2 (March 2020), 191–199. https://doi.org/10.1080/09638237.2019.1677878

[16] Jessica Burns. 2013. A Restorative Justice Model for Mental Health Courts. *S. Cal. Rev. L. & Soc. Just.* 23 (2013), 427.

[17] Zoraida Callejas and David Griol. 2021. Conversational Agents for Mental Health and Wellbeing. In *Dialog Systems*. Vol. 22. 219–244.

[18] Flavio Carvalho, Helder Yukio Okuno, Lais Baroni, and Gustavo Guedes. 2020. A Brazilian Portuguese Moral Foundations Dictionary for Fake News classification. In *2020 39th International Conference of the Chilean Computer Science Society (SCCC)*. 1–5. https://doi.org/10.1109/SCCC51225.2020.9281258

[19] Stevie Chancellor, Eric PS Baumer, and Munmun De Choudhury. 2019. Who is the "human" in human-centered machine learning: The case of predicting mental health from social media. *Proc. CSCW* (2019).

[20] Daejin Choi, Steven A Sumner, Kristin M Holland, John Draper, Sean Murphy, Daniel A Bowen, Marissa Zwald, Jing Wang, Royal Law, Jordan Taylor, et al. 2020. Development of a machine learning model using multiple, heterogeneous data sources to estimate weekly US suicide fatalities. *JAMA network open* 3, 12 (2020), e2030932–e2030932.

[21] Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying Mental Health Signals in Twitter. In *Proc. CLPsych*. ACL, 51–60.

[22] P W Corrigan and D L Penn. 1999. Lessons from social psychology on discrediting psychiatric stigma. *Am Psychol* 54, 9 (Sept. 1999), 765–776.

[23] Patrick W. Corrigan, Karina J. Powell, and Patrick J. Michaels. 2013. The Effects of News Stories on the Stigma of Mental Illness. *JNMD* (March 2013).

[24] Mark H Davis et al. 1980. A multidimensional approach to individual differences in empathy. (1980).

[25] Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *ICWSM*.

[26] Munmun De Choudhury, Sanket S Sharma, Tomaz Logar, Wouter Eekhout, and René Clausen Nielsen. 2017. Gender and cross-cultural differences in social media disclosures of mental illness. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 353–369.

[27] Morteza Dehghani, Kenji Sagae, Sonya Sachdeva, and Jonathan Gratch. 2014. Analyzing Political Rhetoric in Conservative and Liberal Weblogs Related to the Construction of the "Ground Zero Mosque". *Journal of Information Technology & Politics* 11, 1 (Jan. 2014), 1–14.

[28] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. NAACL-HLT*. ACL, 4171–4186.

[29] Sanorita Dey, Brittany R.L. Duff, and Karrie Karahalios. 2022. Re-Imagining the Power of Priming and Framing Effects in the Context of Political Crowdfunding Campaigns. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 127, 22 pages. https://doi.org/10.1145/3491102.3502084

[30] Nicholas Diakopoulos and Michael Koliska. 2017. Algorithmic Transparency in the News Media. *Digital Journalism* 5, 7 (2017), 809–828. https://doi.org/10.1080/21670811.2016.1208053

[31] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[32] Robin IM Dunbar. 1992. Neocortex size as a constraint on group size in primates. *Journal of human evolution* 22, 6 (1992), 469–493.

[33] Robert M Entman. 1993. Framing: Towards Clarification of Fractured Paradigm (Journal of Communication). *PP/S* (1993).

[34] Etienne Esquirol. 1838. *Des maladies mentales considérées sous les rapports médical, hygiénique et médico-légal*. Vol. 1. chez JB Baillière.

[35] Matthew Feinberg and Robb Willer. 2013. The Moral Roots of Environmental Attitudes. *Psychological Science* 24, 1 (2013), 56–62. https://doi.org/10.1177/0956797612449177 PMID: 23228937.

[36] Christopher J Ferguson. 2021. Does the Internet Make the World Worse? Depression, Aggression and Polarization in the Social Media Age. *Bulletin of Science, Technology & Society* 41, 4 (2021), 116–135.

[37] Dèlia Fernàndez-Cañellas, Joan Espadaler, David Rodriguez, Blai Garolera, Gemma Canet, Aleix Colom, Joan Marco Rimmek, Xavier Giro-i Nieto, Elisenda Bou, and Juan Carlos Riveiro. 2019. VLX-stories: Building an online event knowledge base with emerging entity detection. In *ISWC*.

[38] Suman Fernando. 2014. *Mental health worldwide: Culture, globalization and development*. Springer.

[39] Jessica L. Feuston and Anne Marie Piper. 2018. Beyond the Coded Gaze: Analyzing Expression of Mental Health and Illness on Instagram. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 51 (nov 2018), 21 pages. https://doi.org/10.1145/3274320

[40] Catherine Fieschi and Paul Heywood. 2004. Trust, cynicism and populist anti-politics. *Journal of political ideologies* 9, 3 (2004), 289–309.

[41] Scott J Fitzpatrick. 2014. Re-moralizing the suicide debate. *Journal of bioethical inquiry* 11, 2 (2014), 223–232.

[42] Catherine Francis, Jane Pirkis, Catherine Francis, Jane Pirkis, R. Warwick Blood, David Dunt, Philip Burgess, Belinda Morley, Andrew Stewart, and Peter Putnis. 2004. The Portrayal of Mental Health and Illness in Australian Non-Fiction Media. *Australian & New Zealand Journal of Psychiatry* 38, 7 (July 2004).

[43] Andreas Frei, Tanja Schenker, Asmus Finzen, Volker Dittmann, Kurt Kraeuchi, and Ulrike Hoffmann-Richter. 2003. The Werther effect and assisted suicide. *Suicide and Life-Threatening Behavior* 33, 2 (2003), 192–200.

[44] Jeremy A Frimer. 2019. Moral Foundations Dictionary 2.0. https://doi.org/10.17605/OSF.IO/EZN37

[45] King-wa Fu and Paul Siu Fai Yip. 2008. Changes in reporting of suicide news after the promotion of the WHO media recommendations. *Suicide and Life-Threatening Behavior* 38, 5 (2008), 631–636.

[46] Dean Fulgoni, Jordan Carpenter, Lyle Ungar, and Daniel Preoţiuc-Pietro. 2016. An Empirical Exploration of Moral Foundations Theory in Partisan News Sources. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association (ELRA), Portorož, Slovenia, 3730–3736. https://aclanthology.org/L16-1591

[47] Hannah Gaffney, Warren Mansell, and Sara Tai. 2019. Conversational Agents in the Treatment of Mental Health Problems: Mixed-Method Systematic Review. *JMIR Mental Health* 6, 10 (Oct. 2019). https://doi.org/10.2196/14166

[48] Michael Gibney. 2020. Healthcare inequity creates 'crisis of trust' among US marginalized communities. https://www.spglobal.com/marketintelligence/en/news-insights/latest-news-headlines/healthcare-inequity-creates-crisis-of-trust-among-us-marginalized-communities-60807253

[49] Erving Goffman. 2009. *Stigma: Notes on the management of spoiled identity*. Simon and schuster.

[50] Erving Goffman. 2014. Stigma. In *Classic and Contemporary Readings in Sociology*. Routledge, 108–113.

[51] Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. Moral Foundations Theory. In *Advances in Experimental Social Psychology*. Vol. 47. Elsevier, 55–130.

[52] Jesse Graham, Jonathan Haidt, and Brian A. Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology* 96, 5 (May 2009), 1029–1046. https://doi.org/10.1037/a0015141

[53] Derek Greene and Pádraig Cunningham. 2006. Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering. In *Proc. ICML*. ACM Press, 377–384.

[54] Frances J. Griffith and Catherine H. Stein. 2021. Behind the Hashtag: Online Disclosure of Mental Illness and Community Response on Tumblr. *American Journal of Community Psychology* 67, 3-4 (2021), 419–432. https://doi.org/10.1002/ajcp.12483

[55] Ted Grover, Elvan Bayraktaroglu, Gloria Mark, and Eugenia Ha Rim Rho. 2019. Moral and Affective Differences in U.S. Immigration Policy Debate on Twitter. *Computer Supported Cooperative Work (CSCW)* 28, 3 (01 Jun 2019), 317–355. https://doi.org/10.1007/s10606-019-09357-w

[56] Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences* 18 (2017), 43–49. https://doi.org/10.1016/j.cobeha.2017.07.005 Big data in the behavioural sciences.

[57] Anna Rae Gwarjanski and Scott Parrott. 2018. Schizophrenia in the News: The Role of News Frames in Shaping Online Reader Dialogue about Mental Illness. *Health Communication* 33, 8 (Aug. 2018).

[58] Nick Hagar, Johannes Wachs, and Emőke Ágnes Horvát. 2021. Writer movements between news outlets reflect political polarization in media. *New Media & Society* (2021). https://doi.org/10.1177/14614448211027173 arXiv:https://doi.org/10.1177/14614448211027173

[59] Jonathan Haidt. 2013. Moral psychology for the twenty-first century. *Journal of Moral Education* 42, 3 (2013), 281–297.

[60] Jonathan Haidt and Jesse Graham. 2007. When Morality Opposes Justice: Conservatives Have Moral Intuitions that Liberals may not Recognize. *Social Justice Research* 20, 1 (June 2007), 98–116.

[61] Keith N Hampton, Harrison Rainie, Weixu Lu, Maria Dwyer, Inyoung Shin, and Kristen Purcell. 2014. *Social media and the 'spiral of silence'*. Pew Research Center Washington, DC, USA.

[62] Keith Hawton and Kathryn Williams. 2001. The connection between media and suicidal behavior warrants serious attention. (2001).

[63] Australia Department of Health and Aged Care Mental Health Branch. 1999. *Achieving the Balance: A Resource Kit for Australian Media Professionals for the Reporting and Portrayal of Suicide and Mental Illness*.

[64] Claire Henderson and Graham Thornicroft. 2009. Stigma and discrimination in mental illness: Time to Change. *The Lancet* 373, 9679 (2009), 1928–1930.

[65] Bernie Hogan. 2010. The presentation of self in the age of social media: Distinguishing performances and exhibitions online. *Bulletin of Science, Technology & Society* 30, 6 (2010), 377–386.

[66] Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, Gabriela Moreno, Christina Park, Tingyee E. Chang, Jenna Chin, Christian Leong, Jun Yen Leung, Arineh Mirinjian, and Morteza Dehghani. 2020. Moral Foundations Twitter Corpus: A Collection of 35k Tweets Annotated for Moral Sentiment. *Social Psychological and Personality Science* 11, 8 (2020), 1057–1071. https://doi.org/10.1177/1948550619876629 arXiv:https://doi.org/10.1177/1948550619876629

[67] Yulin Hswen, John A Naslund, John S Brownstein, and Jared B Hawkins. 2018. Online communication about depression and anxiety among twitter users with schizophrenia: preliminary findings to inform a digital phenotype using social media. *Psychiatric Quarterly* 89, 3 (2018), 569–580.

[68] RC Jiloha. 2020. COVID-19 and mental health. *Epidemiology International (E-ISSN: 2455-7048)* 5, 1 (2020), 7–9.

[69] Kristen Johnson and Dan Goldwasser. 2019. Modeling behavioral aspects of social media discourse for moral classification. In *Proc. Third Workshop on NLP-CSS*. 100–109.

[70] Megan-Jane Johnstone. 2001. Stigma, social justice and the rights of the mentally ill: Challenging the status quo. *Australian and New Zealand Journal of Mental Health Nursing* 10, 4 (2001), 200–209.

[71] Adam J. Joseph, Neeraj Tandon, Lawrence H. Yang, Ken Duckworth, John Torous, Larry J. Seidman, and Matcheri S. Keshavan. 2015. #Schizophrenia: Use and misuse on Twitter. *Schizophrenia Research* 165, 2-3 (July 2015), 111–115.

[72] Kyriaki Kalimeri, Mariano G. Beiró, Alessandra Urbinati, Andrea Bonanomi, Alessandro Rosina, and Ciro Cattuto. 2019. Human Values and Attitudes towards Vaccination in Social Media. In *Proc. WWW*. ACM, 248–254.

[73] Rishemjit Kaur and Kazutoshi Sasahara. 2016. Quantifying moral foundations from various topics on Twitter conversations. In *IEEE Big Data*.

[74] Bumsoo Kim, Eric Cooks, and Seong-Kyu Kim. 2021. Exploring incivility and moral foundations toward Asians in English-speaking tweets in hate crime-reporting cities during the COVID-19 pandemic. *Internet Research* (2021).

[75] Arthur Kleinman and Rachel Hall-Clifford. 2009. Stigma: a social, cultural and moral process. *Journal of Epidemiology & Community Health* 63, 6 (2009), 418–419. https://doi.org/10.1136/jech.2008.084277 arXiv:https://jech.bmj.com/content/63/6/418.full.pdf

[76] Anat Klin and Dafna Lemish. 2008. Mental Disorders Stigma in the Media: Review of Studies on Production, Content, and Influences. *Journal of Health Communication* 13, 5 (July 2008), 434–449.

[77] William H. Kruskal and W. Allen Wallis. 1952. Use of Ranks in One-Criterion Variance Analysis. *J. Amer. Statist. Assoc.* 47, 260 (1952), 583–621. http://www.jstor.org/stable/2280779

[78] Rohit Kulkarni. 2018. A Million News Headlines. (2018). https://doi.org/10.7910/DVN/SYBGZL

[79] Howard I Kushner. 1991. *American suicide: A psychocultural exploration*. Rutgers University Press.

[80] Haewoon Kwak, Jisun An, Elise Jing, and Yong-Yeol Ahn. 2021. FrameAxis: characterizing microframe bias and intensity with word embedding. *PeerJ Computer Science* 7 (jul 2021), e644. https://doi.org/10.7717/peerj-cs.644

[81] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media?. In *Proceedings of the 19th international conference on World wide web*. 591–600.

[82] David Lederer. 2006. *Madness, religion and the state in early modern Europe: a Bavarian beacon*. Cambridge University Press.

[83] Bruce G. Link and Jo C. Phelan. 2001. Conceptualizing Stigma. *Annual Review of Sociology* 27 (2001), 363–385. http://www.jstor.org/stable/2678626

[84] Eden Litt. 2012. Knock, knock. Who's there? The imagined audience. *Journal of broadcasting & electronic media* 56, 3 (2012), 330–345.

[85] Hankun Liu, Daojing He, and Sammy Chan. 2021. Fraudulent News Headline Detection with Attention Mechanism. *Computational Intelligence and Neuroscience* 2021 (March 2021), 6679661. https://doi.org/10.1155/2021/6679661

[86] Michael MacDonald. 1989. The medicalization of suicide in England: laymen, physicians, and cultural change, 1500-1870. *The Milbank Quarterly* (1989), 69–91.

[87] Winnie WS Mak, Cecilia YM Poon, Loraine YK Pun, and Shu Fai Cheung. 2007. Meta-analysis of stigma and mental health. *Social science & medicine* 65, 2 (2007), 245–261.

[88] Alice E Marwick and Danah Boyd. 2011. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New media & society* 13, 1 (2011), 114–133.

[89] Chandler McClellan, Mir M Ali, Ryan Mutter, Larry Kroutil, and Justin Landwehr. 2016. Using social media to monitor mental health discussions - evidence from Twitter. *Journal of the American Medical Informatics Association* 24, 3 (10 2016), 496–502. https://doi.org/10.1093/jamia/ocw133 arXiv:https://academic.oup.com/jamia/article-pdf/24/3/496/34149079/ocw133.pdf

[90] James McCorriston, David Jurgens, and Derek Ruths. 2015. Organizations are Users Too: Characterizing and Detecting the Presence of Organizations on Twitter. In *Proc. ICWSM*.

[91] Jonathan Michel Metzl, Arthur L Caplan, Joseph Turow, and Otto F Wahl. 2004. *Cultural sutures: Medicine and media*. Duke University Press.

[92] George A. Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM* 38, 11 (nov 1995), 39–41. https://doi.org/10.1145/219717.219748

[93] Brian L Mishara and David N Weisstub. 2016. The legal status of suicide: A global review. *Intl. journal of law and psychiatry* (2016), 54–74.

[94] Negar Mokhberian, Andrés Abeliuk, Patrick Cummings, and Kristina Lerman. 2020. Moral Framing and Ideological Bias of News. In *Social Informatics*, Samin Aref, Kalina Bontcheva, Marco Braghieri, Frank Dignum, Fosca Giannotti, Francesco Grisolia, and Dino Pedreschi (Eds.). Springer International Publishing, Cham, 206–219.

[95] Katarzyna Molek-Kozakowska. 2013. Towards a pragma-linguistic framework for the study of sensationalism in news headlines. *Discourse & Communication* 7, 2 (2013), 173–197. https://doi.org/10.1177/1750481312471668

[96] Logan Molyneux and Mark Coddington. 2020. Aggregation, clickbait and their effect on perceptions of journalistic credibility and quality. *Journalism Practice* 14, 4 (2020), 429–446.

[97] Robert R Morris, Kareem Kouddous, Rohan Kshirsagar, and Stephen M Schueller. 2018. Towards an artificially empathic conversational agent for mental health applications: system design and user perceptions. *Journal of medical Internet research* 20, 6 (2018), e10148.

[98] Susan L Morrow, Donna M Hawxhurst, AY Montes de Vega, Tamara M Abousleman, Carrie L Castañeda, RL Toporek, LH Gerstein, NA Fouad, G Roysircar, and T Israel. 2006. Toward a radical feminist multicultural therapy. *Handbook for social justice in counseling psychology: Leadership, vision, and action* (2006), 231–247.

[99] Stephen J Morse. 1999. Craziness and criminal responsibility. *Behavioral sciences & the law* 17, 2 (1999), 147–164.

[100] Fred Morstatter, Liang Wu, Uraz Yavanoglu, Stephen R. Corman, and Huan Liu. 2018. Identifying Framing Bias in Online News. *Trans. Soc. Comput.* 1, 2, Article 5 (jun 2018), 18 pages. https://doi.org/10.1145/3204948

[101] Christin L Munsch, Liberty Barnes, and Zachary D Kline. 2020. Who's to Blame? Partisanship, Responsibility, and Support for Mental Health Treatment. *Socius* 6 (2020), 2378023120921652.

[102] J. A. Naslund, K. A. Aschbrenner, L. A. Marsch, and S. J. Bartels. 2016. The future of mental health care: peer-to-peer support and social media. *Epidemiology and Psychiatric Sciences* (April 2016).

[103] John A Naslund, Kelly A Aschbrenner, Lisa A Marsch, and Stephen J Bartels. 2016. The future of mental health care: peer-to-peer support and social media. *Epidemiology and psychiatric sciences* 25, 2 (2016), 113–122.

[104] Eve Ng. 2020. No grand pronouncements here...: Reflections on cancel culture and digital media participation. *Television & New Media* 21, 6 (2020), 621–627.

[105] Alina Pavlova and Pauwke Berkers. 2020. "Mental Health" as Defined by Twitter: Frames, Emotions, Stigma. *Health Communication* (Dec. 2020).

[106] Sachin R Pendse, Daniel Nkemelu, Nicola J Bidwell, Sushrut Jadhav, Soumitra Pathare, Munmun De Choudhury, and Neha Kumar. 2022. From Treatment to Healing: Envisioning a Decolonial Digital Mental Health. In *CHI*.

[107] James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. *Linguistic Inquiry and Word Count*. Lawerence Erlbaum Associates, Mahwah, NJ.

[108] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1532–1543. https://doi.org/10.3115/v1/D14-1162

[109] Shruti Phadke and Tanushree Mitra. 2020. Many Faced Hate: A Cross Platform Study of Content Framing and Information Sharing by Online Hate Groups. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376456

[110] J Hunter Priniski, Negar Mokhberian, Bahareh Harandizadeh, Fred Morstatter, Kristina Lerman, Hongjing Lu, and P Jeffrey Brantingham. 2021. Mapping moral valence of tweets following the killing of George Floyd. *arXiv preprint arXiv:2104.09578* (2021).

[111] Tao Qi, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. 2021. *Personalized News Recommendation with Knowledge-Aware Interactive Matching*. Association for Computing Machinery, 61–70. https://doi.org/10.1145/3404835.3462861

[112] Nicola J. Reavley and Pamela D. Pilkington. 2014. Use of Twitter to monitor attitudes toward depression and schizophrenia: an exploratory study. *PeerJ* 2 (2014), e647. https://doi.org/10.7717/peerj.647

[113] Markus Reiter-Haas, Simone Kopeinik, and Elisabeth Lex. 2021. Studying Moral-based Differences in the Framing of Political Tweets. *Proceedings of the International AAAI Conference on Web and Social Media* 15, 1 (May 2021), 1085–1089. https://ojs.aaai.org/index.php/ICWSM/article/view/18135

[114] Barbara Ricci and Lisa Dixon. 2015. What can we do about stigma? *Psychiatric Services* 66, 10 (2015), 1009–1009.

[115] Patrick Robinson, Daniel Turk, Sagar Jilka, and Matteo Cella. 2019. Measuring attitudes towards mental health using social media: investigating stigma and trivialisation. *Social Psychiatry and Psychiatric Epidemiology* 54, 1 (Jan. 2019), 51–58. https://doi.org/10.1007/s00127-018-1571-5

[116] Shamik Roy and Dan Goldwasser. 2021. Analysis of nuanced stances and sentiment towards entities of US politicians through the lens of moral foundation theory. In *Proc. SocialNLP@NAACL*.

[117] Koustuv Saha, John Torous, Eric D Caine, Munmun De Choudhury, et al. 2020. Psychosocial effects of the COVID-19 pandemic: large-scale quasi-experimental study on social media. *Journal of medical internet research* 22, 11 (2020), e22600.

[118] Christian Schwarzenegger. 2020. Personal epistemologies of the media: Selective criticality, pragmatic trust, and competence–confidence in navigating media repertoires in the digital age. *New Media & Society* 22, 2 (2020), 361–377.

[119] Usman Shahid, Barbara Di Eugenio, Andrew Rojecki, and Elena Zheleva. 2020. Detecting and understanding moral biases in news. In *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*. Association for Computational Linguistics, Online, 120–125. https://doi.org/10.18653/v1/2020.nuse-1.15

[120] Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2021. Towards Facilitating Empathic Conversations in Online Mental Health Support: A Reinforcement Learning Approach. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 194–205. https://doi.org/10.1145/3442381.3450097

[121] Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *Proceedings of the Web Conference 2021*. 194–205.

[122] Eva Sharma and Munmun De Choudhury. 2018. Mental health support and its relationship to linguistic accommodation in online communities. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.

[123] Chaitanya Shinkhede. 2019. Digital Frailty: Proliferation of Clickbait, Beguiled Readers, and Questioning the Morality of Online Journalism. *marketing* 6 (2019).

[124] Richard A. Shweder, Nancy C. Much, Manamohan Mahapatra, and Lawrence Park. 1997. The "big three" of morality (autonomy, community, divinity) and the "big three" explanations of suffering.

[125] Jacob Smith and Jonathan Spiegler. 2020. Explaining gun deaths: Gun control, mental illness, and policymaking in the American states. *Policy studies journal* 48, 1 (2020), 235–256.

[126] Sashank Sridhar and Sowmya Sanagavarapu. 2021. Content Based News Recommendation Engine using Hybrid BiLSTM-ANN Feature Modelling. In *2021 Joint 10th ICIEV and 2021 5th icIVPR*. 1–8.

[127] Dustin S. Stoltz and Marshall A. Taylor. 2021. Cultural cartography with word embeddings. *Poetics* 88 (Oct 2021), 101567.

[128] Michael Strupp-Levitsky, Sharareh Noorbaloochi, Andrew Shipley, and John T Jost. 2020. Moral "foundations" as the product of motivated social cognition: Empathy and other psychological underpinnings of ideological divergence in "individualizing" and "binding" concerns. *PloS one* 15, 11 (2020), e0241144.

[129] Heather Stuart. 2008. Fighting the stigma caused by mental disorders: past perspectives, present activities, and future directions. *World Psychiatry* 7, 3 (2008), 185.

[130] Anne Marie Stupinski. 2020. Measuring Mental Health Stigma on Twitter. (2020).

[131] Steven A Sumner, Moira Burke, and Farshad Kooti. 2020. Adherence to suicide reporting guidelines by news shared on a social networking platform. *PNAS* 117, 28 (2020), 16267–16272.

[132] Thomas S Szasz. 1961. The myth of mental illness. (1961).

[133] Tahmina Tanny and Chowdhury Abdullah Al-Hossienie. 2019. Trust in government: Factors affecting public trust and distrust. *Jahangirnagar Journal of Administrative Studies, Department of Public Administration* 12 (2019), 52.

[134] Phillip T Tatum, Silvia Sara Canetto, and Michael D Slater. 2010. Suicide coverage in US newspapers following the publication of the media guidelines. *Suicide and Life-Threatening Behavior* 40, 5 (2010), 524–534.

[135] Shubhra Tewari, Renos Zabounidis, Ammina Kothari, Reynold Bailey, and Cecilia Ovesdotter Alm. 2021. Perceptions of Human and Machine-Generated Articles. *Digital Threats* 2, 2, Article 12 (apr 2021), 16 pages. https://doi.org/10.1145/3428158

[136] The Carter Center. 2015. Journalism Resource Guide on Behavioral Health. https://www.cartercenter.org/resources/pdfs/health/mental_health/2015-journalism-resource-guide-on-behavioral-health.pdf

[137] Felicity Thomas, Lorraine Hansford, Joseph Ford, Katrina Wyatt, Rosemarie McCabe, and Richard Byng. 2018. Moral narratives and mental health: rethinking understandings of distress and healthcare support in contexts of austerity and welfare reform. *Palgrave communications* 4, 1 (2018), 1–8.

[138] Andrew Thompson. 2020. https://components.one/datasets/all-the-news-2-news-articles-dataset/

[139] Charles B Truax and Robert Carkhuff. 2007. *Toward effective counseling and psychotherapy: Training and practice*. Transaction Publishers.

[140] Melissa Tully, Emily K Vraga, and Anne-Bennett Smithson. 2020. News media literacy, perceptions of bias, and interpretation of news. *Journalism* 21, 2 (2020), 209–226.

[141] Jessica Vitak. 2012. The impact of context collapse and privacy on social network site disclosures. *Journal of broadcasting & electronic media* 56, 4 (2012), 451–470.

[142] Ivo Vlaev, Dominic King, Paul Dolan, and Ara Darzi. 2016. The theory and practice of "nudging": changing health behaviors. *Public Administration Review* 76, 4 (2016), 550–561.

[143] Otto F. Wahl. 1995. *Media madness: Public images of mental illness.* Rutgers University Press. Pages: xiv, 220.

[144] Otto F. Wahl. 2003. News Media Portrayal of Mental Illness: Implications for Public Policy. *American Behavioral Scientist* (Aug. 2003).

[145] Weirui Wang. 2019. Stigma and Counter-Stigma Frames, Cues, and Exemplification: Comparing News Coverage of Depression in the English- and Spanish-Language Media in the U.S. *Health Communication* 34, 2 (Jan. 2019), 172–179. https://doi.org/10.1080/10410236.2017.1399505

[146] Jeffrey Rodgers Watt. 2004. *From sin to insanity: Suicide in early modern Europe.* Cornell University Press.

[147] WHO. 2019. Mental disorders. https://www.who.int/news-room/fact-sheets/detail/mental-disorders

[148] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2021. Empowering news recommendation with pre-trained language models. In *Proc. SIGIR.*

[149] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou. 2020. MIND: A Large-scale Dataset for News Recommendation. In *Proc. ACL.* 3597–3606.

[150] Eva Yampolsky and Howard I Kushner. 2020. Morality, mental illness and the prevention of suicide. *Social Epistemology* 34, 6 (2020), 533–543.

[151] Diyi Yang, Zheng Yao, Joseph Seering, and Robert Kraut. 2019. The channel matters: Self-disclosure, reciprocity and social support in online cancer support groups. In *Proceedings of the 2019 chi conference on human factors in computing systems.* 1–15.

[152] Lawrence H Yang, Fang-pei Chen, Kathleen Janel Sia, Jonathan Lam, Katherine Lam, Hong Ngo, Sing Lee, Arthur Kleinman, and Byron Good. 2014. "What matters most:" a cultural mechanism moderating structural vulnerability and moral experience of mental illness stigma. *Social science & medicine* 103 (2014), 84–93.

[153] Lawrence Hsin Yang, Arthur Kleinman, Bruce G Link, Jo C Phelan, Sing Lee, and Byron Good. 2007. Culture and stigma: Adding moral experience to stigma theory. *Social science & medicine* 64, 7 (2007), 1524–1535.

[154] Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level Convolutional Networks for Text Classification. In *NIPS.*

[155] Qinfeng Zhu and Marko M Skoric. 2021. From context collapse to "safe spaces": Selective avoidance through tie dissolution on social media. *Mass Communication and Society* 24, 6 (2021), 892–917.

APPENDIX

A VALIDATION OF THE PROPOSED APPROACH

To what extent does our BERT-based framework capture the underlying moral foundations of a tweet or a news article? Following prior work [94, 113], we validated our method using the publicly available Moral Foundations Twitter Corpus (MFTC) [66]. MFTC consists of 31,108 tweets that cover seven discourse topics: All Lives Matter (ALM), Black Lives Matter (BLM), the Baltimore protests, the 2016 Presidential election, hate speech, Hurricane Sandy, and #MeToo. These tweets were hand-annotated by at least 3 trained annotators for the 10 categories of moral sentiments (virtue and vice dimensions considered separately for the five foundations.)

We verified the proposed BERT-based method by comparing the moral foundation category labels assigned via our framework against the ground truth annotations, for tweets in MFTC. Using the approach described in Section 4.3, the extracted sentence level BERT embedding for each tweet present in MFTC was scored against the five moral foundation embedding vectors using the cosine similarity metric, resulting in a five element feature vector. For instance, comparing a tweet against the Care/Harm moral foundation generated a cosine similarity score such that, a score greater (lesser) than 0 (the selected threshold value) is indicative of alignment with Care (Harm). Finally, each tweet was assigned the moral sentiment category that received the highest absolute score.

Furthermore, to assess the reliability of BERT-based embeddings we compared the performance of our framework with an existing work [94] – that adopts Moral Foundation Theory to study partisanship in news media. [94] uses GloVe-based [108] word embeddings to generate representation of the five moral foundations, and scores tweets by calculating a weighted (by word count) average of cosine similarity between each individual word's GloVe embedding in the tweet and the five vector representations for the five moral foundations. Table A1 summarizes the performance of our method and [94] for assigning the moral foundation categories to tweets in MFTC. We report results for each moral foundation separately, combining the virtue and vice facets. From Table A1 it can be seen that our method outperforms the GloVe-based framework, emphasizing the importance of capturing relationships between words using a contextualized sequence encoder like BERT, for a better representation of moral framing in text. Lastly, we also validated the choice of 0 as a threshold value. Experimenting with other thresholds between -1 and 1 resulted in decreased performance on MFTC, making 0 an empirically sound choice.

Table A1: Performance of our framework on MFTC compared to the GloVe embedding framework [94] across all the five moral foundations.

Moral Foundation	**Precision**		**Recall**		**F1-score**	
	GloVe	Ours	GloVe	Ours	GloVe	Ours
Care/Harm	0.68	**0.75**	0.71	**0.77**	0.70	**0.76**
Fairness/Cheating	0.66	**0.78**	0.73	**0.79**	0.69	**0.78**
Loyalty/Betrayal	0.72	**0.83**	0.74	**0.86**	0.73	**0.84**
Authority/Subversion	0.81	**0.87**	0.82	**0.86**	0.81	**0.86**
Sanctity/Degradation	0.84	**0.91**	0.86	**0.94**	0.85	**0.92**

B REGULAR EXPRESSION QUERIES TO FILTER PERSONAL SELF-DISCLOSURES

Table A2 provides the regular expression queries used to filter personal self-disclosures.

C APPROVAL/STIGMA DICTIONARY

Table A3 lists the keywords present in our Approval and Stigma dictionaries.

Table A2: Regular expression queries used to filter tweets containing personal self-disclosures of a mental health condition.

Regular expression query	Keyword
i (am | was | have been) diagnosed with i (think i) have diagnosed me with	(anxiety | depression | mental illness | bpd | ptsd | schizophrenia | mental disorder | social anxiety | anorexia | bulimia | binge eating disorder | eating disorder)
i (am | was | have been)	suicidal
i (am thinking | have thought | though) about	suicide
i (attempted | considered)	suicide
i (am | was | have been)	(schizophrenic | depressed | anorexic | bulimic | bipolar)
i (used to)	(self-harm | self harm)
i feel	(depressed | bipolar)
i (had | am having) a/an	(panic attack | anxiety attack)

Table A3: Keywords in the Approval/Stigma dictionaries.

Dimension	Keywords
Approval	respect, support, endorse, endorsement, sanction, valid, validation, accept, accepting, accredit, confirm, agree, agreement, compliance, cooperation, receipt, accedence, affirm affirmation, recognition, award, concurrence, admire, admiration, account, applause, favor, praise, regard, nod, accept, acceptation, acceptance, appreciate, matter, approbation, recommendation, acclaim, esteem, encourage, commendation, confirmation, confirming, laudable, permit, blessing, approve, approving, approval, countenance, accord, favour, repute, endorsed, vouched, backed, allow, allowed, supportive, accordance, admission, assent, consent, honor, accolade, glory, glorify, glorification, understandable, acknowledge, acknowledgement, okay, yes, condone, recommend, credit, commend, content, recognize, boost, promote, accommodate, take-in, receiving, welcome, value
Stigma	humiliation, disgrace, reject, rejection, rejected, outlaw, taboo, scorn, exclusion, despair, refuse, refusal, dishonor, dislike, disapproval, disfavor, disapprobation, dissatisfaction, displeasure, distrust, displeased, discontent, criticism, discouragement, shame, discredit, spot, blemish, defect, flaw, sin, guilty, fault, faulty, derogation, derogatory, object, objection, flawed, admonishment, admonition, doom, abstaining, discord, hostile, hostility, turn-down, abnegation, exclude, exclusion, unnatural, unjustifiable, shallow, embarrassment, embarrassing, reckless, defective, refrain, restrict, tarnish, stain, scar, censure, condemn, condemnation, stigmatizing, stigmatization, malign, maligning, prejudice, ignore, ignorance, discriminate, discrimination, reproach, disappointment, shut-out, cut-out, bar, neglect, disrepute, evil, weak, despise, indifference, shameful