Algorithmic Teenagers' Depression Detection on Social Media and Automated Instant Engagement Using Therapy Bot Powered by Multimodal Deep Learning and Psychotherapy Intervention

Abstract

Social media is a large and growing feature of teen life across the world. While some research suggests that social media are partly to blame for growing rates of mental illness among teens, social media can also play a positive role in promoting teen mental health by giving teens new ways to socialize and feel part of a community. In this work, we propose a framework for developing system that can further enhance the upsides of social media use: a computational model that uses social media data to predict depression, as part of a detection-and-intervention loop that engages the user in positive conversations when dynamic indicators of depression present themselves in their social media activity. Our framework uses three pillars of multimodal Content, Behavioral, and Contextual data drawn from users' social media feeds in order to provide timely detection and intervention services via a chatbot. This multimodal architecture allows us to envision chat features that are precise and responsive to the behavior that triggers the detection. We present a review of the state of the art in depression detection systems, and then proceed to explain our system, which builds upon successes in deep learning-based detection systems, as well as placing these tools in the new dynamic setting of *online* depression detection that enables our chatbot to initiate therapeutic interactions with social media users.

In the age of social media, teenage mental health has become of increasing concern and also increasing prominence. The World Health Organization estimates that 322 million people worldwide suffer from depressive disorders (WHO, 2017), leading to sizable public health, well-being, and economic impacts. Furthermore, in countries with deep social media penetration, teen incidence of mental illness has seen a spike that coincides with the increasingly widespread use of social media. In the United States, government statistics indicate that the incidence of teen major depression has steadily increased since the early 2010s, with particularly acute incidence among teen girls, with 23% having experienced major depression in 2019 (The Federal Interagency Forum on Child and Family Statistics, 2021). The rise in teen mental disorder has led to widespread public debate about the possible role of social media use in driving this trend (Haidt, 2022), a view supported by a wealth of scientific research (Keles et al., 2020).

Yet, while incidence of depression has increased among teens, the rate at which teens with depression obtain mental health services has remained steady and relatively low (The Federal Interagency Forum on Child and Family Statistics, 2021), indicating a need for mechanisms that provide targeted interventions. Social media can be a contributing cause of mental illness, but it can also play a positive role by providing networks of detection and support for teens manifesting symptoms of depression online. One way to augment this positive aspect of social media in teens' lives is by integrating algorithmic detection and feedback loops that can intervene strategically when social media users are symptomatic, and route them to services that can help.

Facilitated by technology, the accessibility of mental health treatment has increased in recent years (WHO, 2020), but there remains a gap in the availability, cost, and timeliness of services. Telemedicine platforms like BetterHelp have made therapy widely available on demand, but accessibility to such services is limited by the cost of service, and people may not receive the services that they need at the optimal point of intervention. Considerations like this have led researchers, therapists, and therapy clients to explore the possibility of "robot therapy"—therapy powered by intelligent conversational agents that can promote mental health at moment it is needed, and at a low cost to the user (Wiseman, 2017). Digitally-delivered therapies can have particular impact in developing countries, where access to mental health services is especially sparse. According to the WHO, low-income countries have only 1 psychiatrist for every million people—a figure 9 times higher than the global average (Oladeji and Gureje, 2016), despite the fact that mental health crises in these countries may sometimes be even more severe.[1]

[1] Uddin et al. (2019), for example, find that schoolchildren

In this paper, we review the research on algorithmic approaches to detecting mental illness, as well as practical solutions that have been developed for remediating it, when detected, using social media-based interventions and conversational agents.

1 Review of the literature

1.1 Mental health and social media

Social media is a pervasive feature of modern life. It is estimated that nearly 3.5 billion people use social media worldwide (Karim et al., 2020), and its impacts on mental health have been widely studied (Valkenburg et al., 2022). While individual studies discover associations between social media and depression (Pantic et al., 2012; Jeri-Yabar et al., 2018), there is considerable scientific debate as to the causal effect of social media use on mental health. While the broad question of whether social media has a positive or negative impact on the mental health of users, studies have consistently shown that social media data provide reliable indicators of depression (see Section 1.2), providing a unique place where computational approaches can positively impact the mental health of social media users, providing useful diagnostic tools as well as intervention mechanisms.

1.2 Depression detection

A wide array of studies have demonstrated that text and image data such as are available on social media platforms can be used to reliably predict depression. Approaches vary along the axes of the computational approaches used as well as the data that are used to yield predictions.

1.2.1 Text-based features

One of the most reliable indicators of a user's emotional state is the language that they use, and linguistic markers of emotion on social media have been found to significantly correlate with depression (Park et al., 2012) , making text-based sentiment analysis an important part of a social media-based depression detection loop. There exists a wealth of systems for transforming text data from social media posts into features reflective of emotional state (see Al-Tameemi et al. (2022) for a review). Generic sentiment analysis pipelines can provide a useful input into depression detection by highlighting a high rate of sad, angry, or negatively-weighted posts, which can enter into ensemble-model predictions along with other factors. Classical text-based sentiment analysis can be based on classical methods like BoW indicators or Tf-IDF vectors (which take into account the statistical frequencies of words in a corpus of documents) or pre-trained single-word distributed representations like Word2Vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2014) combined together within a post by averaging.

The current generation of text embedding models is based on recurrent neural network (LSTMs/-GRUs) and transformer architectures, which take into account syntactic and semantic arrangement in addition to the words themselves, allowing representations to distinguish between sentences like *My friend was sad so I made a bowl of soup :)* from *I was sad so my friend made a bowl of soup :)*, which contain exactly the same words but attribute the target emotion, *sad*, to different event participants.

Recently, state-of-the-art text encoders have been used effectively in depression detection studies. Ameer et al. (2022) present a text-based transfer learning approach based on pretrained text encoders (BERT, XLNet, and RoBERTa) trained to detect a battery of 5 mental ailments (including depression) and a control condition (no mental illness) on a dataset of 16,930 Reddit posts. As compared compared with N-gram based text features and several other deep learning architectures, the pretrained transformers consistently performed best, with RoBERTa achieving the best accuracy (F1=.83). One drawback of their training and evaluation set is ecological validity: the target classes are balanced, including the control condition, whereas most people do not experience mental health disorders in any given year. This can lead to a recall-bias in the real-world deployment of such systems, though this imbalance may in practice be mitigated by self-selection of depression-prone individuals into the user group, yielding user pools that are balanced in the deployment context.

Zogan et al. (2022) expand depression prediction beyond the level of single posts using a customized multi-stage architecture they title HAN: the Hierarchical Attention Network. The core architecture is made up of blocks of Gated Recur-

in the African region have the highest prevalence of suicidal ideation and planning among all global regions studied.

rent Units (GRUs) that embed posts sequentially in bidirectional fashion, with input representations initialized to GloVe vectors, fine-tuned on their dataset. The contextualized word embeddings output by the GRU are then multiplied by learned attention weights, and summed to yield the representation of a tweet. To aggregate tweets to user-level representations, a second GRU performs the same procedure on the ordered collection of individual tweet vectors. The text embedding module alone achieves 84.4% accuracy on depression detection on their custom dataset, and reaches 89.5% when combined with other modules.

1.2.2 Image data

Traditional sentiment analysis benchmarking is conducted on text data, facilitated by the widespread availability of large annotated training sets like the Amazon Product Database (He and McAuley, 2016), IMDB reviews (Maas et al., 2011), the Stanford Sentiment Treebank (Socher et al., 2013), , and including social media-specific datasets like Sentiment140 (Go et al., 2009). Major social media platforms like Instagram and TikTok, however, require users to structure all content around images or videos, with text playing an ancillary role. In order to incorporate these data, researchers have introduced the concept of image-sentiment analysis, and produced new datasets—based for instance on human-generated image tags (Gajarla and Gupta, 2015) or parallel corpora of images and sentiment-annotated text (Vadicamo et al., 2017)—making it possible to associate images with sentiment labels. Importantly, image data have been used successfully to predict user-level depression in various studies.

Approaches based on heuristics. One category of approaches is based on coarse heuristics applied to shallow image features. Reece and Danforth (2017) use image-wide averages of pixel-level Hue, Saturation, and Value properties—whicfh intuitively help distinguish dark and colorless from lively and colorful images—to predict depression from Instagram profiles. A coarse measure of sociality was derived from a count of the number of faces detected in each image. Behavioral metadata such as the number of likes each image received were also included, as well as human-annotated emotive ratings for a subset of images. A composite model that includes all of these features significantly predicted depression (F1=.647), outperforming benchmark diagnostic rates from a general physician. Katthula report a similar result from heuristic image analysis in a more informal analysis.[2]

A richer set of aggregate image-level features is used by Guntuku et al. (2019) to predict depression from profile images, including aesthetic ratings of image lighting, motion blur, depth of field, vividness of color, and other elements, showing significant correlations between many of these factors and depression/anxiety. However, since they pursued a correlational line of analysis and do not provide accuracy or F1 values, this approach is difficult to compare with other methods.

Image-to-text protocols. A number of studies pursue sentiment analysis on images using pipelines that transform images into linguistic data that are descriptive of the image. The resulting linguistic tags can then be processed using NLP tools for sentiment analysis. In a multimodal depression detection system, (Safa et al., 2022) run images through a queryable API that returns a set of word-level tags that are descriptive of each image, yielding a “Bag-of-Visual-Words” (BoVW), which are then processed as independent features. Accuracy for the image-to-text model alone tops out at .69 accuracy on users who self-reported DSM-5 depression symptoms in their posts. Fused with other feature sets including text, their model achieves accuracy upwards of 90%.

Deep learning-based approaches. A final category of image analysis pipelines makes use of state of the art approaches based on deep learning architectures. (Qian et al., 2019) use text-image feature fusion to analyze sentiment on Twitter posts containing both images and text. Visual features were contributed by AlexNet, a convolutional neural network (CNN) deep architecture. Visual features were paired with embeddings of text. Notably, they were able to achieve 72% accuracy using just the visual features after fine-tuning the pre-trained AlexNet on a very small dataset (1,269) of Twitter images labeled with sentiment, demonstrating the power of transfer learning in settings where data must be collected. A model fusing visual and textual features achieves 80% accuracy on single posts.

Huang et al. (2019) employ a similar paradigm to predict user-level depression tendency from Instagram images, with additional feature fusion

[2] https://github.com/VipanchiKatthula/Instagram-Depression-Detection

from behavioral observables. Image features were derived from a CNN pre-trained on ImageNet to yield compact 64-dimensional image embeddings, with the vectors for a fixed number L of images concatenated into a $L \times 64$-d vector representing the user's images. In the full model, the image features are concatenated with text features and behavioral indicators (e.g. time of posts, number of likes, etc.), with a post-processing network that pools the features and outputs softmax probabilities. Image data alone is sufficient to obtain .774 F1 with relatively high precision (.811), with the fusion model performing at F1=.823 (precision=.888) on a custom dataset.

Our assessment of the state of the art. Heuristic methods based on coarse image features have proven useful in a number of studies, but it is worth noting that such image statistics can be trivially derived from most deep learning architectures. The mean pixel hue for a given color, for example, is the output of an image-wide convolutional filter with weight $1/(H \times W)$ on the appropriately-weighted RGB components of the image. Hence, the heuristic features are learnable in a DL framework, and given the success of models fine-tuned on small depression-detection datasets, we suspect that such models are not overly parametrized for this task. For the same reason, image-to-text preprocessing seems unnecessary from the point of view of achieving high-performance models. The state of the art in this field is difficult to assess due to the heterogeneity of methods and, especially, of evaluation datasets. However, image embedding-based systems consistently perform well in the literature, and do better than methods that do not take the end-to-end fine-tuning approach.

One advantage of text-to-image preprocessing pipelines is interpretability: the numerical features output by CNN-based models are difficult to associate with the sort of emotive content targeted by our system, whereas image-tags like *dark* or *alleyway* vs *bright* or *beach* have a clear link to the poster's emotive state. This interpretability gap can be dealt with by using vision-language (VL) models that connect image embeddings to linguistic representations. Transformer-based VL models like CLIP (Alec Radford, 2021) that are trained to map image embeddings into a shared space with language embeddings can provide an interpretable basis for sentiment detection on images without prior mapping to the language modality, balancing the raw predictive power of DL-based numerical image features with the ability to probe the model for keywords that explain the model's decision on a given image.

1.2.3 Incorporating video features

. On most platforms—with the exception of TikTok—text and images provide the majority of content. Still, industry surveys suggest that a large share (around 15%) of both Facebook and Instagram posts come in the form of video[3]. Several approaches can be used to enable feature extraction from video. Fast object-detectors like Faster-RCNN (Ren et al., 2015) can quickly iterate through and tag video frames in order to generate data on the objects included in the video, yielding, for example, measures of sociality based on depicted interactions with people, animals, and possessions for example. An innovative approach to video-based depression diagnosis is proposed in (Lee and Park, 2022), in the context of an interactive chatbot that detects depression-associated facial expressions used as a basis for interventions. The authors propose to use Fast-RCNN to detect detect people in video frames and perform analysis of facial expressions in an interactive session with a chatbot therapist, allowing for online classification of the user's emotional state and helping identify markers of depression. This application specifically depends on videos of the target user, and could be used in a social media setting when coupled with prior detection of video participants (since many social media posts are of the user themselves). However, this method remains experimental even in the context of explicitly diagnostic sessions where the user's facial expressions are elicited by the therapy chatbot. Further work is needed to assess the efficacy of the approach.

1.2.4 Behavioral factors

Beyond text and image data, behavioral metadata about the user's time and frequency of posting can provide useful indicators. DSM5 criteria like insomnia and fatigue may be detected from the user's time (e.g. late-night posting) and frequency (e.g. reduction in the amount of posts) of social media engagement. Measures such as the user's number of followers and likes can indicate vary-

[3] https://www.digitalinformationworld.com/2021/05/images-videos-or-links-study-shows-most.html

ing rates of social engagement and the presence or absence of a community of support—factors that correlate with depression (Elmer and Stadtfeld, 2020).

Multiple studies have incorporated behavioral factors into online depression detection. De Choudhury et al. (2013) conducted a study of social media behavior and depression, testing a variety of statistical models of diurnal activity (posting times), volume (number of posts per day), interactivity (posting vs @reply behavior) and sociality (followers/followees), finding each of these factors to be statistically significant. Shen et al. (2017) incorporate the user's number of tweets and social interactions, as well as the posting time distribution, in a multimodal ensemble model that achieves .85 F1 when combining text, behavioral, and image features. A model built from behavioral features alone achieved about F1=.67, indicating an independent contribution to overall accuracy from this component.

1.3 Depression detection in a dynamic setting

The main methodological gap in the depression detection techniques reviewed here is that depression is identified statically at the level of the individual user, rather than dynamically in the context of user behavior across time. Of course, the same variables that apply at the user level can also help us identify acute signs of depression from social media posts, and since the goal is to promote friendly intervention rather than clinical diagnosis, a certain amount of detection error is tolerable. On a technical level, each of the systems used for user-level detection can be adapted to the dynamic prediction setting by windowing the user's social media timeline and normalizing to statistical baselines for each variable set, specific to each user, in order to ensure that interventions occur at relevant times.

1.4 Conversational agent-based therapies

Chatbot-based therapy is a growing area of research and clinical practice. As distinct from in-person engagement with a therapist, AI-powered conversational agents can deliver low-cost therapies to patients on demand, dramatically expanding the accessibility of mental health services. Existing commercial systems like *Wysa* and *Woebot* deliver mental health-focused conversations to users via Android and iOS applications, incorporating clinical paradigms like Cognitive Behavioral Therapy (CBT) and mindfulness (Tewari et al., 2021).

AI-based therapy systems have been shown to be effective in multiple studies. (Liu et al., 2022) conducted a randomized control trial of a custom chatbot based on Cognitive Behavioral Therapy (CBT) principles, finding that patients who engaged with their chat system had significantly improved scores on the PHQ-9 and GAD-7—standard psychological questionnaires for assessing depression and anxiety. While their chatbot was primarily based on multiple-choice conversation flows, other researchers have developed systems that allow for more flexible engagement. MS et al. (2022) developed a chatbot proposal that uses BERT-based sentiment analysis on user-input text to modulate the chatbot's output in a way that is responsive to the user's emotional state. Although more difficult to parametrize so that their output is modulated by model-derived inputs, Sequence-to-Sequence models with deep learning architectures like Google's LaMDA (Thoppilan et al., 2022) can provide an additional layer of flexibility by enabling open-ended and truly interactive conversation with the user (Nuruzzaman and Hussain, 2018).

In terms of setting design parameters for a therapeutic conversational agent, (Grové, 2021) pioneered a collaborative methodology for co-development of a teen mental health chatbot that incorporated input from stakeholders in the development/design phrase. Youth were actively recruited into qualitative interviews that elicited their views on what factors would make an effective and engaging chatbot—one with a style and vocabulary tuned to communicating with the target demographic (use of slang and emojis), and with a "guide-type" personality that is "inspiring and charismatic, fun, friendly, empathetic, humorous". The engineered system itself—"Ash"—was based on keyword/phrase detection used to initiate dialogue scripts co-written by researchers and research participants. This methodology provides a template for designing an algorithmic social media intervention system that not only accurately detects the need to engage the user in an interaction, but promotes a positive experience with the chatbot.

2 Proposed architecture

Our approach combines an online depression detection system with a series of targeted interventions triggered by detection outputs.

2.1 High-level description of the system

Figure 1 presents the overall architecture. On the left hand side, we first extract multimodal data from the user's social media content, with the data from all modalities fed into a series of analysis modules that featurize the data into Content, Behavioral, and Contextual indicators.

2.2 Text features

In the language modality, the text associated with the user's Twitter, Instagram, and TikTok posts/-comments/hashtags is combined to extract content features using state-of-the-art NLP tools, yielding sentiment ratings for each post. Models suitable for the initial feature extraction include BERT, RoBERTa, and ALBERT, each of which is based on the state-of-the-art transformer architecture, and which have been evaluated on social media text (with fine-tuning on the social media text data, which differ stylistically from a typical pre-training distribution for language models) (Singla and N., 2020). Core to this approach is transfer learning: fine-tuned models based on pre-training with masked language modeling (MLM) allow accurate sentiment models to be learned from small labeled samples, such as the Corona Tweets dataset (Lamsal, 2020). Single-post F1 values for transfer-learning based sentiment analysis models top out at around .85 in Singla and N. (2020)'s survey, but in our approach, noise in this segment of the labeling is smoothed out by combining sentiment ratings for multiple adjacent posts in order to yield final predictions.

In addition to making use of text analysis of the *user's* posts, our system also provides an analysis of interactions with their social network via replies. The goal is to identify potentially negative interactions with others, and flag negative posters with interventions by the chatbot module, which can recommend ending interactions with or unfollowing users generating potentially harmful content. Similarly, the chatbot can identify positive users/interactions and suggest further engagement, e.g. "*I'm sorry to hear that you're feeling sad. Would you like to talk about it? // You seem to be getting along with X. Do you know them well? Why don't you see if they have time to chat with you?*

2.3 Visual features

In the visual modality, we combine the user's posted images and frame samples from videos. Crucially, since Instagram and TikTok posts frequently include user-generated in-image captions, we add text detection and recognition (OCR) in pre-processing, with the results fed to the language processors. Instagram posts in particular frequently include long segments of user text as well as emojis, making this a crucial step.

Our visual feature analysis modules are based primarily on end-to-end deep learning networks. While some multimodal approaches like those detailed in (Safa et al., 2022) and (Guntuku et al., 2019) use image-to-language protocols to label images with tags reflecting the content of the image, which are then processed in the text modality, deep learning approaches provide overall better performance, and at lower cost in terms of feature engineering. Feature extraction via deep neural networks have shown accuracy up to 73% on predicting emotive human-generated Flickr tags (Gajarla and Gupta, 2015), or upwards of 90% in binary-choice (positive/negative sentiment) tasks (Alghalibi et al., 2020). Combining such systems with other factors via ensembling promises yet further improvements to the detection system.

2.4 Behavioral/Contextual features

In addition to the machine-learning-powered content analysis, Behavioral and Contextual data provide the final two pillars of our new system. Behavioral features like the time and frequency of posting have proven useful in predicting depression from social media activity (see Section 1.2.4), and we use them to provide statistical measures of the user's level of engagement in their social media community, as well as estimating any variation in these measures. Since some users may have naturally higher posting rates than others, we take a baselining approach to feature-generation: the features input from the behavioral module are derived by windowing the user's current activity (e.g. number of posts in the last K days) and comparing it to the user's baseline routine and its natural variability. Statistical models of the behavioral observables are mapped into vectors comparing the user's windowed behavior to their base levels, and

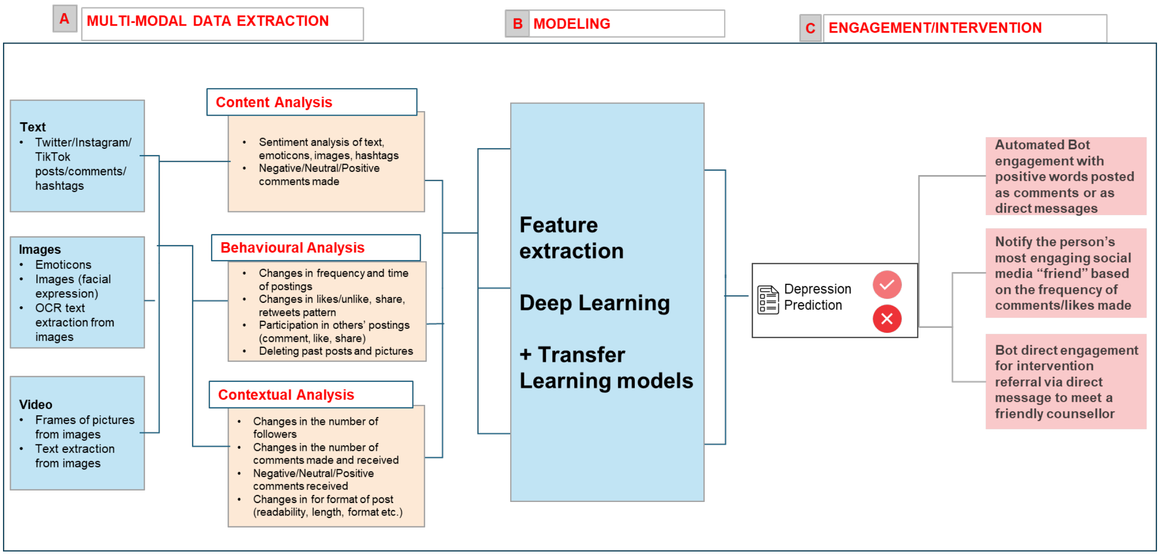
A
MULTI-MODAL DATA EXTRACTION
B
MODELING
C
ENGAGEMENT/INTERVENTION
Text
Twitter/Instagram/ TikTok posts/comments/ hashtags
Images
Emoticons
Images (facial expression)
OCR text extraction from images
Video
Frames of pictures from images
Text extraction from images
Content Analysis
Sentiment analysis of text, emoticons, images, hashtags
Negative/Neutral/Positive comments made
Behavioural Analysis
Changes in frequency and time of postings
Changes in likes/unlike, share, retweets pattern
Participation in others' postings (comment, like, share)
Deleting past posts and pictures
Contextual Analysis
Changes in the number of followers
Changes in the number of comments made and received
Negative/Neutral/Positive comments received
Changes in for format of post (readability, length, format etc.)
Feature extraction
Deep Learning
+ Transfer Learning models
Depression Prediction
Automated Bot engagement with positive words posted as comments or as direct messages
Notify the person's most engaging social media "friend" based on the frequency of comments/likes made
Bot direct engagement for intervention referral via direct message to meet a friendly counsellor

Figure 1: Flow diagram for the proposed system. See section 2.1 for discussion.

triggering chatbot responses when anomalies are detected.

The Contextual pillar includes analysis of aggregate data associated with the user profile: number of followers, amount of engagement with the user's profile (positive/negative comments received), as well as coarse statistics about the user's own posts—the length of posts, or changes in format (for example, decrease in amount of text and increase in amount of image posts). In this set of modules as well, we compare the current values to the baseline values to derive inputs to the inference ensemble.

2.5 Ensembling

Each of the Text/Image, Behavioral, and Contextual computational modules provides the data used to generate a final prediction by combining information from each subsystem. To generate dynamic online inference about the user's depression, we concatenate the vector representations of each feature set and compute depression predictions for each user at a given interval t. Let $\boldsymbol{T}_t, \boldsymbol{B}_t, \boldsymbol{C}_t$ denote the vectors of Content, Behavioral, and Contextual features respectively at time t, and let $\mathcal{M}$ be a binary prediction model that outputs a probability that the user is depressed. The ensembling model $\mathcal{M}$ may initially be set to logistic regression, which yielded good results for Shen et al. (2017). The depression prediction for the user at t may be modeled as:

$$\mathcal{M}\left[\boldsymbol{T}_t \oplus \boldsymbol{B}_t \oplus \boldsymbol{C}_t\right] \in [0, 1] \qquad (1)$$

2.6 Engaging the user via the chatbot

Since features are relativized to the user's baselines, we may choose a global threshold that triggers a chatbot intervention. The main variable triggering an interaction with the chatbot is detection that the user is potentially depressed, in which case the chatbot requests a check-in with the user. In addition to the single prediction, the system may also make use of the values of individual features in order to guide the conversation:

Example check-in initiators

1. **Generic** *I looked at your feed and you seem to be feeling badly these days. Would you like to talk about it?*

2. **Content** *You've been posting lots of dark photos recently. I wanted to see if everything's OK!*

3. **Behavioral** *I haven't seen you as much on Instagram these days. How are you feeling? Let's check in!*

4. **Contextual** *I noticed some people are being really mean on your feed. I would love to share some tips on how to build some more positivity in your life!*

The design of the chatbot itself takes into account the design considerations highlighted in (Grové, 2021) to promote engagement by the targeted youth demographic, using slang, communication styles, and scripts tuned to the local context of deployment (youth in different countries use different slang terms, for example, and an effective chatbot will need to provide interactions properly localized to the user's social context. Therefore, we build our conversational agent by actively seeking out collaboration with youth in the places where it is deployed.

References

Israa Khalaf Salman Al-Tameemi, Mohammad-Reza Feizi-Derakhshi, Saeed Pashazadeh, and Mohammad Asadpour. 2022. A comprehensive review of visual-textual sentiment analysis from social media networks.

Chris Hallacy Aditya Ramesh Gabriel Goh Sandhini Agarwal Girish Sastry Amanda Askell Pamela Mishkin Jack Clark Gretchen Krueger Ilya Sutskever Alec Radford, Jong Wook Kim. 2021. Learning transferable visual models from natural language supervision. *ICML 38*.

Maha Alghalibi, Adil Al-Azzawi, and Kai Lawonn. 2020. Deep attention learning mechanisms for social media sentiment image revelation. *International Journal of Computer and Communication Engineering*, 9.

Iqra Ameer, Muhammad Arif, Grigori Sidorov, Helena Gòmez-Adorno, and Alexander Gelbukh. 2022. Mental illness classification on social media texts using deep learning and transfer learning.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the 7th Internatioal AAAI Conference on Weblogs and Social Media*.

Timon Elmer and Christoph Stadtfeld. 2020. Depressive symptoms are associated with social isolation in face-to-face interaction networks. *Nature Scientific Reports*, 10.

Vasavi Gajarla and Aditi Gupta. 2015. Emotion detection and sentiment analysis of images.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision.

Christine Grové. 2021. Co-developing a Mental Health and Wellbeing Chatbot With and for Young People. *Frontiers in Psychiatry*, 11.

Sharath Chandra Guntuku, Daniel Preotiuc-Pietro, Johannes C Eichstaedt, and Lyle H Ungar. 2019. What twitter profile and posted images reveal about depression and anxiety. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 236–246.

Jonathan Haidt. 2022. Testimony before the senate judiciary committee, subcommittee on technology, privacy, and the law.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. *World Wide Web Conference Committee*.

Yu Huang, Chieh-Feng Chiang, and Arbee L. P. Chen. 2019. Predicting depression tendency based on image, text and behavior data from instagram. In *Proceedings of the 8th International Conference on Data Science, Technology and Applications*.

Antoine Jeri-Yabar, Alejandra Sanchez-Carbonel, Karen Tito, Jimena Ramirez-delCastillo, Alessandra Torres-Alcantara, Daniela Denegri, and Yhuri Carreazo. 2018. Association between social media use (twitter, instagram, facebook) and depressive symptoms: Are twitter users at higher risk? *International Journal of Social Psychiatry*, 65.

Fazid Karim, Azeezat A Oyewande, Lamis F Abdalla, Reem Chaudhry Ehsanullah, and Safeera Khan. 2020. Social media use and its connection to mental health: A systematic review. *Cureus*, 12.

Betul Keles, Niall McCrae, and Annmarie Grealish. 2020. A systematic review: the influence of social media on depression, anxiety and psychological distress in adolescents. *International Journal of Adolescence and Youth*, 25:79–93.

Rabindra Lamsal. 2020. Coronavirus (COVID-19) Tweets Dataset.

Young-Shin Lee and Wong-Hung Park. 2022. Diagnosis of depressive disorder model on facial expression based on fast r-cnn. *Diagnostics*, 12.

Hao Liu, Huaming Peng, Xingyu Song, Chenzi Xu, and Meng Zhang. 2022. Using ai chatbots to provide self-help depression interventions for university students: A randomized trial of effectiveness. *Internet Interventions*, 27:100495.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *ICLR*.

Muneshwara MS, Swetha MS, Poorvi Rohidekar, and Pranove AB. 2022. Implementation of therapy bot for potential users with depression during covid-19 using sentiment analysis. *Journal of Positive School Psychology*, 6:7816–7826.

Mohammad Nuruzzaman and Omar Khadeer Hussain. 2018. A survey on chatbot implementation in customer service industry through deep neural networks. In *Proceedings of IEEE 15th International Conference on e-Business*.

Bibilola D. Oladeji and Oye Gureje. 2016. Brain drain: a challenge to global mental health. *BJPsych International*, 13:61–63.

Igor Pantic, Aleksandar Damjanovic, Jovana Todorovic, Dubravka Topalovic, Dragana Bojovic-Jovic, Sinisa Ristic, and Senka Pantic. 2012. Association between online social networking and depression in high school students: behavioral physiology viewpoint. *Psychiatria Danubina*, 24:90–93.

Minsu Park, Chiyoung Cha, and Meeyoung Cha. 2012. Depressive moods of users portrayed in twitter. In *Proceedings of ACM SIGKDD Workshop on Healthcare Informatics*.

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)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Chen Qian, Edoardo Ragusa, Iti Chaturvedi, E. Cambria, and Rodolfo Zunino. 2019. Text-image sentiment analysis.

Andrew G Reece and Christopher M Danforth. 2017. Instagram photos reveal predictive markers of depression. *EJP Data Science*, 6.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Ramin Safa, Peyman Bayat, and Leila Moghtader. 2022. Automatic detection of depression symptoms in twitter using multimodal analysis. *The Journal of Supercomputing*.

Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Fent, Cunjun Zhang, Tianriu Hu, Tat-Seng Chua, and Wenwu Zhu. 2017. Depression detection via harvesting social media: A multimodal dictionary learning solution. In *IJCAI-17*.

Saurav Singla and Ramachandra N. 2020. Analysis of transformer based pre-trained nlp models. *Journal of Computer Sciences and Engineering*, 8.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Abha Tewari, Amit Chhabria, Ajay Singh Khalsa, Sanket Chaudhary, and Harshita Kanal. 2021. A survey of mental health chatbots using NLP. In *Proceedings of the International Conference on Innovative Computing & Communication*.

The Federal Interagency Forum on Child and Family Statistics. 2021. America's children: Key national indicators of well-being.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed H. Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications. *CoRR*, abs/2201.08239.

Riaz Uddin, Nicola W Burton, Myfanwy Maple, Shanchita R Khan, and Asaduzzaman Khan. 2019. Suicidal ideation, suicide planning, and suicide attempts among adolescents in 59 low-income and middle-income countries: a population-based study. *The Lancet Child Adolescent Health*, 3(4):223–233.

Lucia Vadicamo, Fabio Carrara, Andrea Cimino, Stefano Cresci, Felice Dell'Orletta, Fabrizio Falchi, and Maurizio Tesconi. 2017. Cross-media learning for image sentiment analysis in the wild. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 308–317.

Patti M. Valkenburg, Adrian Meier, and Ine Beyens. 2022. Social media use and its impact on adolescent mental health: An umbrella review of the evidence. *Current Opinion in Psychology*, 44:58–68.

WHO. 2017. Depression and other common mental disorders. *WHO Global Health Estimates*.

WHO. 2020. Mental Health Atlas.

Eva Wiseman. 2017. Is robot therapy the future? *The Guardian*.

Hamad Zogan, Imran Razzak, Xianzhi Wang, Shoaib Jameel, and Guandong Xu. 2022. Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. *World Wide Web*, 25:281–304.