Early Depression Detection with Transformer Models: Analyzing the Relationship between Linguistic and Psychology-Based Features

Haya Halimeh
Paderborn University
halimeh@mail.upb.de

Matthew Caron
Paderborn University
matthew.caron@upb.de

Oliver Mueller
Paderborn University
oliver.mueller@upb.de

Abstract

Clinical depression is a serious mental disorder that poses challenges for both personal and public health. Millions of people struggle with depression each year, but for many, the disorder goes undiagnosed or untreated. Over the last decade, early depression detection on social media emerged as an interdisciplinary research field. However, there is still a gap in detecting hesitant, depression-susceptible individuals with minimal direct depressive signals at an early stage. We, therefore, take up this open point and leverage posts from Reddit to fill the addressed gap. Our results demonstrate the potential of contemporary Transformer architectures in yielding promising predictive capabilities for mental health research. Furthermore, we investigate the model's interpretability using a surrogate and a topic modeling approach. Based on our findings, we consider this work as a further step towards developing a better understanding of mental eHealth and hope that our results can support the development of future technologies.

Keywords: early depression detection, mental eHealth, transformers, surrogate models, LIWC

1. Introduction

Clinical depression is among the most prevalent disorders worldwide, burdening approximately 5% of adults throughout the world (World Health Organization, 2022). The more far-reaching implications of depression can be inferred from the reported socioeconomic consequences, where it is estimated to cost the global economy up to USD 1 trillion annually (The Lancet Global Health, 2020). Furthermore, it is the leading cause of disability and the most common psychiatric diagnosis associated with suicide (Yates et al., 2017). Nevertheless, depression often remains underdiagnosed and untreated (The Lancet Global Health, 2020). Reasons for this include limited medical resources, social stigma, or even discrimination (Orabi et al., 2018). A recent study that looked at mental health changes throughout the COVID-19 pandemic in the United States revealed, for instance, that the prevalence of elevated depressive symptoms increased from 27.8% in 2020 to 32.8% in 2021 (Ettman et al., 2022). This notable uptick in the number of people with mental health concerns has only exacerbated difficulties in meeting mental health care needs and overwhelmed health care workers (Van Wert et al., 2022). On the other hand, self-stigmatization has been shown to correlate with lower rates of help-seeking behavior and higher rates of social avoidance (Manos et al., 2009). Indeed, reports from the National Institute of Mental Health suggest that it takes an average of more than ten years for a person with a mental illness to seek help (National Institute of Mental Health, 2019); therefore, highlighting the need for other screening and support methods in psychological and social practice.

Over the past decade, however, the unprecedented growth of social media has made it clear that the way people communicate is changing. A direct example can be witnessed in the emergence of e-peer support communities, which claim to provide new channels that encourage free self-expression, and niches that enable peer social support for mental health (Manikonda & De Choudhury, 2017). The largely-increasing user-generated textual interactions in these communities occur in a rather uncontrolled and natural setting (Losada & Crestani, 2016), such that they can be used as a means to infer behavioral and psychological cues about a user's mental health status (Carey et al., 2018).

URI: https://hdl.handle.net/10125/103046
978-0-9981331-6-4

As a result, an interdisciplinary body of research driven by contemporary natural language processing (NLP) and Machine Learning (ML) is increasingly devoted to improving mental health. This development comes as no surprise, especially since a plethora of psychological research has established the link between language and psychological processes (Pennebaker et al., 2003).

Despite the inherent complexity of the underlying problem and the generally acknowledged need for further personal assessment, operationalizing the problem as an early depression detection task introduces an additional layer of intricacy. While some studies focus on identifying at-risk individuals as early as possible (Losada & Crestani, 2016), others aim to predict depression before it is even reported (De Choudhury et al., 2013). Nevertheless, most studies aimed at identifying depression-prone individuals based on their text interactions on social media do not retain depression-related discourse nor distinguish between interactions posted before and after joining a depression e-peer support community. In this work, we take a different approach and attempt to identify predisposed individuals before they even join an e-peer support depression community. Our motivation stems from the realization that user-generated text interactions on social media can be leveraged in conjunction with modern NLP techniques to detect, at an early stage, depression-prone individuals who would or should join such a community and, by this means, reach those who are still reluctant or not yet open about it.

To tackle the task at hand, we experiment with MentalBERT – i.e., a pretrained Transformer model for mental health research (Ji et al., 2022). Pretrained Transformers models are based on sophisticated architectures capable of capturing implicit and complex patterns in language by building on the idea of transfer learning – i.e., to exploit what has been learned in one task to improve another one (Pan & Yang, 2009; Qiu et al., 2020). However, MentalBert's lack of transparency and black-box characteristics makes it difficult to understand, which is an undesirable liability in a field as sensitive as mental health. Given this, the question arises as to the relationship between how modern task-oriented language models learn and linguistic indicators. To gain further insight into this area, we explore the link between the model's probabilistic predictions and the well-established gold standard LIWC – i.e., Linguistic Inquiry and Word Count (Pennebaker et al., 2015).

In this work, we thus set out to, first, address the underlying problem of early depression detection using social media (specifically Reddit posts) and a domain-specific Transformer model. Second, we investigate the association between the model's estimates and the gold standard LIWC features. Our findings are three-fold: (1) we provide additional empirical evidence that language markers and emotional cues are strong predictors of mental health status, (2) we add to the existing body of mental health and AI research by leveraging state-of-the-art NLP and transfer learning technique to identify depressed individuals prior to their first participation to an e-peer support depression community on Reddit, and (3) we assess the interpretability of the pretrained Transformer model with the current LIWC gold standard and a topic modeling technique. With our work, we hope to contribute to improving both mental eHealth awareness and outcomes.

2. Related Work

2.1. Early depression detection

For decades, social and medical disciplines, such as psychiatry, psychology, and sociolinguistics, have combined efforts to understand and identify factors associated with depressive disorders. This multidisciplinary stream of research has addressed a wide range of factors, from somatic, such as decreased energy (Abdel-Khalek, 2004), to social, such as lack of social support (Brown et al., 1986). While this research is essential in improving the overall understanding of mental disorders, it builds primarily on small samples of individuals who are not necessarily representative of whole populations. Consequently, over the last years, the growing reach of social media platforms has led to an increasing interest in using social media as a tool for clinical analysis of depression (M. Park et al., 2012). Primarily, social media served as a source to detect clinically depressed individuals or explore discourse associated with them (see, for example, De Choudhury et al., 2013; Monselise & Yang, 2022).

When it comes to detecting depression on social media, with Twitter (Coppersmith et al., 2014) and Reddit (Losada et al., 2018) being the most prevalent, researchers have taken a variety of approaches and mostly established the ground truth for positive annotations through self-disclosure statements (Coppersmith et al., 2014), community participation (De Choudhury & De, 2014), or human assessment (Wang et al., 2013). With Reddit in focus, Losada and Crestani published in 2016, for example, a collection of textual interactions gathered from a random sample of self-declared depressed and non-depressed Reddit users. The collected corpus served as a basis for a novel shared task for depression detection on the web. Since then, multiple collaborative online risk screening tasks have annually taken place, covering both mental disorders and related symptomatology – e.g., self-injury[1].

The task of early depression detection is, henceforth, admittedly not new. However, a shortcoming across most studies is that they neither narrow the content to a limited extent nor distinguish the depressed group over time. Given the way most studies operationalize the task and the inclusion of depression-related discourse, it is reasonable to assume that algorithms trained on such datasets will concentrate primarily on depression-focused content. In fact, this line of work has heavily relied on feature engineering to detect depressive signals. The features varied accordingly from domain-specific vocabularies (De Choudhury et al., 2013), over to distinctive linguistic attributes (M. Park et al., 2012), to user-activity levels (Kotikalapudi et al., 2012). Yet, although the task of identifying depressive individuals is, as such, just as important, such models are expected to underperform when it comes to identifying individuals who are still hesitant or who suffer from self- or social stigma. In truth, Wolohan et al. (2018) made initial efforts to circumvent this problem by limiting positive class texts to non-depression-related subreddits. Still, the authors did not consider the temporal aspects and set the ground truth to posts on the subreddit *r/depression*. This approach could lead to including people who do not necessarily suffer from clinical depression but may just be curious or seeking advice on how to better cope with depressed loved ones. In addition, most studies examining depression on Reddit have primarily looked at a single subreddit – e.g., *r/depression* (Losada & Crestani, 2016) or *r/SuicideWatch* (Monselise & Yang, 2022). Such an approach comes with the disadvantage of limiting the analysis to one subreddit; as such, interactions posted elsewhere are not taken into account; therefore, potentially resulting in information loss. Thus and to the best of our knowledge, there exists a gap in detecting clinically depressed subjects prior to their initial participation in an e-peer support community.

In view of the above, we believe that if automated approaches were available to administrators of e-peer support communities, platforms could more efficiently identify at-risk but not yet active individuals sooner – i.e., individuals who are not yet actively participating in an e-peer support forum and may still be hesitant to talk about their own mental health issues – and, as a result, facilitate access to social capital and support. Hence, with this work, we aim to fill this gap by capturing initial participation in an e-peer support community for depression in general rather than a specific community. Members of the *r/GFD* for example – i.e., Gamers Fight Depression subreddit – may have posted first there and later in the *r/depression* subreddit.

2.2. Transfer learning

In itself, transfer learning is a process that describes the transfer of knowledge acquired by solving a previous task to help solve a different but related one (Pan & Yang, 2009). As such, transfer learning aims to resemble human learning and has, on that account, gained significant momentum in deep learning applications and revolutionized the field of natural language understanding.

In NLP, sequential transfer learning is the most extensively used approach. This approach typically consists of two phases: (1) pretraining and developing a skillful source model on vast amounts of data, and (2) adapting the initial model to a usually smaller but more targeted text corpus to accomplish a given task. According to several works, most contemporary approaches use language modeling for the pretraining phase (Qiu et al., 2020), where the model tries to figure out the next token depending on the previous ones. This has the advantage that the model tries to learn the source language's comprehensive semantic and syntactic representations. In the adaptation phase, the pretrained language model can then be further fine-tuned to a specific domain where the model learns patterns specific to the task at hand. Currently, most transfer learning techniques in NLP use so-called Transformer architectures (Vaswani et al., 2017), which in turn have been praised by Google (Nayak, 2019).

The merit of using transfer learning with state-of-the-art Transformer architectures is the ability to perform otherwise complicated and expensive NLP tasks with more efficient resources and high predictive performance (Caron et al., 2022). Hence, in this work, we experiment with MentalBERT – i.e., a Transformer-based language model pretrained explicitly on a diversity of mental health-related posts collected from Reddit to ease the automatic detection of mental disorders online (Ji et al., 2022).

2.3. LIWC linguistic features

There is, thus far, a broad consensus in scientific research about the association between language and psychological phenomena (Tausczik & Pennebaker, 2010). A large body of work has consistently demonstrated that people's language can reveal rich insights into their psychological states, including their emotions and thought patterns (Pennebaker et al., 2015). As such, in mental health research, language has

[1] For further reference, see https://erisk.irlab.org/2017/index.html

been repeatedly used as a critical factor in detecting depressive signals (Pennebaker et al., 2003). For instance, Beck's cognitive theory of depression (1979), which dates back to the last century, states that depressed people tend to perceive themselves and their surroundings more negatively, that they are more likely to express themselves in negative terms, and that they use first-person pronouns more often.

To ease capturing people's psychological states in written texts, Pennebaker et al. (2015) developed the so-called LIWC program[2]. LIWC is a psychometric analysis software with over 100 built-in dictionaries created to identify people's psychological and social states in texts. The performance of LIWC has been validated in countless studies in the field of NLP and mental illnesses (see, for example, Ali et al., 2015; X. Chen et al., 2018; Wolohan et al., 2018).

Following this line of thought, it is reasonable to assume that the language depression-prone individuals use may reflect patterns that existed before they first posted in a depression-related community on Reddit. At the same time, building on our argument about transfer learning in the last section begs the question of whether there is a connection between the way modern task-oriented language models learn and the linguistic categories of LIWC. More specifically, it is interesting to investigate if there is a relationship between the predictions made by the fine-tuned MentalBERT model and the subjects' written texts on which the model bases its predictions.

3. Experiments

3.1. Setting

As discussed above, in this paper, we set out to address the problem of early depression detection using social media while also investigating (1) if advanced NLP models, such as MentalBERT (Ji et al., 2022), can detect depression-prone individuals before the first post in an e-peer support community and (2) what the nature of the relationship between the probabilistic estimates of the fine-tuned MentalBERT for depression-prone individuals and their texts is? To this end, we begin by creating a cohort of depression-prone and non-depressed subjects before proceeding with our experiments with a surrogate model.

3.2. Dataset

The text corpus used in this work is based on posts collected on the widely used communication platform Reddit. Reddit was chosen because it enables the creation of communities, so-called subreddits, that cover various topics, including medical problems such as depression. Members typically have a long history of submitted posts that can be leveraged for different NLP tasks[3]. The submissions include posts – i.e., standalone submissions – and comments. The following subsections illustrate the creation steps for the negative and positive classes.

3.2.1. Depression-prone group (positive). While there are different approaches as to how the ground truth labels can be established, we rely, for the reasons explained in section 2.1, on self-reported diagnoses following a methodology akin to the one employed by Losada and Crestani (2016). Accordingly, the diagnosed group is created by retrieving subjects from the subreddit *r/depression* with explicit mention of diagnosis – e.g., "I was diagnosed with [...]". For further validation, every retrieved submission was manually reviewed, and subjects with expressions such as – e.g., "I feel depressed" – were not included.

Since our goal is to detect depressed individuals ex-ante – that is, before they started posting in an e-peer support community – we identified the most popular e-peer support communities for depression based on community memberships and participation in these subreddits by individuals from our data sample. The resulting sixteen subreddits[4] are listed in Table 1.

Table 1. Depression-related subreddits

r/depression	r/depressed	r/selfhelp
r/AnxietyDepression	r/depression_help	r/GFD
r/SuicideWatch	r/mentalhealthmemes	r/mmfb
r/overcoming	r/getting_over_it	r/itgetsbetter
r/HaveHope	r/helpmecope	r/offmychest
r/mentalillness	r/mentalhealth	

Following the work by De Choudhury et al. (2013), we extracted, for each subject, all posts submitted in the twelve months prior to a subject's first post in one of the identified subreddits. The rationale behind the twelve-month time window is that capturing large chronologies may weaken the model's predictive power, as diagnosed individuals may have developed depression at later stages, and their early contents may have been more neutral. For further validation, authors with no previous posting history prior to their first post in one of the online support communities were excluded. For the same reason, authors with less than

[2] https://www.liwc.app/

[3] The data was obtained using https://github.com/pushshift/api.

[4] For formal descriptions, visit https://www.reddit.com.

four submissions – i.e., posts and comments – were also excluded. This threshold was used by A. Park, Conway, et al. (2017) to determine so-called lurkers – i.e., individuals who are not yet regular contributors. Lastly, all submissions were chronologically arranged and assigned a positive label.

3.2.2. Control group (negative). The control cohort was sampled by retrieving random subjects from various subreddits. The submitted texts of each subject in the control cohort were equally chronologically ordered and assigned a negative label. Subjects having posted at least once in one of the identified subreddits in table 1 were removed to mitigate the possibility of having depressed individuals in the control group. In addition, a check to rule out collision between the two collections was performed.

Forming the control class in this way – i.e., without restriction to an annual interval – introduces a possible bias in the number of texts among subjects in both cohorts. At the same time, this avoids the problem of biased language – i.e., linguistic changes with changing circumstances and events. Thus, restricting the control group to a one-year interval could bias the learning procedure and the resulting findings. On this account, contributions from the control group were restricted to the same time interval as those from the depression-prone group. All contributions from both cohorts occurred between January 1, 2015, and December 31, 2021. To alleviate the former bias, both classes are equally weighted in training, and the metric used for evaluation is the F_1 macro since it treats all classes equally regardless of their support values.

Table 2. Descriptive statistics

Aspect	Depression-Prone	Control
Number of Subjects	362	487
Total Submissions	52.939	698.635
Avg. Tokens / Text	27	24

Finally, for the sake of generalization and robustness, we conducted five different experiments, each with a random allocation of subjects to a train, validation, and a test set, with their total retrieved text chronologies (with 80%, 10%, 10% split of Subjects). Table 2 provides the main descriptive statistics of both cohorts.

3.3. Transformer-based approach

To tackle the task at hand, we opted for the uncased version of the MentalBERT model (Ji et al., 2022). In essence, MentalBERT is a language model based on the BERT-base architecture proposed by Devlin et al. (2019) and pretrained in the same fashion as BERT (Bidirectional Encoder Representations from Transformers) – i.e., arguably one of the most widely used Transformer models. The model incorporates several bidirectional encoder layers that take raw unprocessed text data as input and convert it into contextual vectors.

Pretrained Transformer models usually do not require extensive preprocessing. Hence, we limited our preprocessing to removing URLs as well as lowercasing and truncating the texts to a limit of 140 tokens (95% of the documents in our corpus are shorter or equal to this maximum length). The shorter texts were subsequently padded to the same length and eventually fed to the model to train. Since the texts of the control subjects are over-represented compared to those of the positive subjects, we adjusted the class weights in the loss function inversely proportional to class frequencies before fine-tuning the model to the downstream task. Accordingly, we treat each class equally, independent of its volume.

The model was then trained for a total of 3 epochs, using the hyperparameters recommended by the BERT authors – i.e., $batch\ size = 16$, $learning\ rate = 2e-5$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $weight\ decay = 0.01$ (Devlin et al., 2019). While there are several strategies for determining which metric to use when optimizing a classification task, we follow a standard procedure for imbalanced learning and determine the optimal threshold (instead of the default 50%) for discriminating between the two classes by optimizing the macro score F_1 on the validation set. We then aggregate the predictions for each individual in the test set and evaluate the resulting average probability using the optimized threshold. Accordingly, any subject in the test set with an average probability above this value is classified as vulnerable to depression, while probabilities below the threshold result in a non-depressive prediction.

3.4. Baseline models

For our baseline, we draw on the depression detection literature and experiment with commonly used models with LIWC features. Specifically, we experiment with a Logistic Regression model (Cox, 1958), a Random Forest model (Ho, n.d.) and an Extreme Gradient Boosting model (T. Chen & Guestrin, 2016) (see, for example, X. Chen et al., 2018; Trotzek et al., 2017). The models are trained with default hyperparameters[5], and the classification threshold is set

[5] https://scikit-learn.org/stable/

for the subjects in the same fashion as in MentalBERT. Because of the study's objectives, we consider the selected models with LIWC features a reasonable choice for the baseline.

4. Results

4.1. Predictive results

Table 3 summarizes the performance of the deployed models in terms of Precision, Recall, and F_1 macro scores averaged over the five runs. The obtained predictive performance shows that all models perform well on the test set, thereby providing additional empirical evidence that language is a strong predictor of mental health status. More interestingly, the results indicate that depressed and non-depressed individuals express themselves differently online, even before they join an e-peer depression support group or manifest their depressive concerns there. As can be seen, MentalBERT outperforms its counterparts in all metrics. This interesting finding suggests that a Transformer-based approach, such as MentalBERT, could be sensitive to different linguistic signals and capture more information than the employed LIWC-based approaches. We further investigate this point in the following subsection.

Table 3. Predictive results

Model	Precision	Recall	F_1
LIWC + Logistic Regression	0.74	0.71	0.69
LIWC + Random Forest	0.77	0.74	0.73
LIWC + XGBoost	0.78	0.77	0.76
MentalBERT (Ji et al., 2022)	**0.82**	**0.81**	**0.82**

Altogether, the results imply that the application of state-of-the-art NLP techniques has promising potential to detect depressed individuals in online settings and prior to their first posting to an e-peer depression community. In addition, they add to the existing body of research on LIWC features and demonstrate their suitability for NLP tasks in online mental health research.

4.2. The relationship between linguistic and psychology-based signals

In the last section, we experimented with domain-specific transfer learning and the gold standard LIWC to solve an early detection classification task. Because Transformer models are sophisticated architectures, it is not surprising that they surpassed the dictionary-based approaches. At the same time, given that the LIWC-based approaches achieved competitive results on the given test data, they proved to be helpful in modeling important textual cues for the early detection task. With this in mind, it is interesting to explore whether there is a relationship between what MentalBERT learns and what the LIWC features capture in order to decipher the black-box nature of the model and, as a result, better understand its results.

Thus, we follow the analogy of a surrogate model in which the outcome of interest – i.e., the probability estimates – is modeled by another, more easily understood model. By interpreting the surrogate model, we can draw conclusions about the black box model (Molnar, 2022). We began the analysis by building the LIWC measures for the test data at the aggregate level. The measures were consequently computed for every submission and averaged for each author. We follow this strategy because we operationalized the task at the individual level rather than the document level. Given that the whole process is, in a cross-validation manner, repeated five times with different random splits, some authors may appear more than once in the resulting test sets. This is remedied by averaging the probability values for each subject as a function of the frequency of their occurrence in the test data sets. Then, to quantitatively assess the relationship between the LIWC characteristics and the averaged MentalBERT probability estimates, we fit an ordinary least squares (OLS) regression model with the LIWC features as the independent variables, and the averaged MentalBERT probability estimates as the dependent variables.

Next, to choose the statistically significant LIWC features (ρ value threshold set to 5%), we proceed with a Sequential Backward Selection (SBS) (Pudil et al., 1994). Initially, the model is fitted with all LIWC features as predictors. Then, the feature space is iteratively deduced into subspace features by removing the predictor with the highest ρ value and refitting the model with the remaining subspace features. The selection is repeated until all remaining explanatory variables are statistically significant against the predefined ρ value (see above). As the LIWC categories are closely related – e.g., negative emotions and anger – multicollinearity becomes a concern. Therefore, we determined the variance inflation factor (VIF) of all selected features and excluded those with a VIF factor above the rule-of-thumb threshold – i.e., 10 – to rule out any multicollinearity concerns (Gujarati, 2021). Finally, several commonly used statistical tests were conducted, including a Jarque-Bera normality test, a Durbin-Watson test for autocorrelation, and a White-Test for the homoscedasticity assumption. All obtained statistics were in favor of the OLS assumption. More interestingly, the R-squared value is about 0.72.

Table 4. Surrogate model

	Dependent variable
	MentalBERT (Probabilities / σ)
Constant	0.199*** (0.024)
First singular Pronoun (I)	0.017*** (0.001)
Personal Pronoun (you)	0.016*** (0.003)
Personal Pronouns (she/he)	0.014*** (0.005)
Word Count	0.001*** (0.000)
Authenticity	-0.001** (0.000)
Swear	0.016*** (0.004)
Negative Tone	0.012*** (0.002)
Positive Tone	0.006*** (0.001)
Substances	0.030*** (0.010)
Anxiety	0.040*** (0.013)
Positive Emotion	-0.010*** (0.003)
Power	-0.008*** (0.002)
Lack	0.016*** (0.003)
Nonfluencies	0.045*** (0.013)
Culture	0.002*** (0.001)
Auditory	0.008*** (0.002)
Curiosity	-0.012** (0.005)
Ethnicity	-0.015** (0.006)
Female	0.016*** (0.004)
Friend	0.012*** (0.003)
Religion	-0.016*** (0.004)
Reward	0.015*** (0.004)
Sexual	0.016*** (0.004)
Social Referents	-0.010*** (0.002)
Punctuation	-0.001*** (0.000)
Comma Usage	0.003** (0.001)
Conjunctions	0.012*** (0.002)
Observations	345
R^2	0.719
Adjusted R^2	0.695
Residual Std. Error	0.072(df = 317)
F Statistic	29.985*** (df = 27.0; 317.0)
Note:	*p<0.1; **p<0.05; ***p<0.01

As can be seen in Table 4, the regression results indicate that there is an association between the selected LIWC features and MentalBERT's probability predictions. In fact, the relatively high value of R-squared suggests that these features can explain around 72% of the variation in MentalBERT's estimates. Some of these features, associated examples (Boyd et al., 2022), and their relationship to the model's predictions are briefly discussed below:

- **Standard linguistic dimensions:**
  - *First Person Singular*: The first person singular "I" is among the most prominent features used in online mental health research (see, for example, Pennebaker et al., 2015) and a more frequent use is linked to increased attention towards the self and is known to correlate with depression (Rude et al., 2004). The positive and significant coefficient suggests that the more often an individual uses this pronoun c.p., the more confident the MentalBERT model is at classifying a given individual as depressed.
  - *Words Count*: Word count has been linked to talkativeness (Tausczik & Pennebaker, 2010). We find a significant positive link between the number of words and a positive prediction.
- **Psychological processes:**
  - *Anxiety and Positive Emotion*: LIWC emotional measures have been repeatedly used to estimate perceived emotions in textual communication (X. Chen et al., 2018). The coefficients of the predictors Anxiety and Positive Emotion are significant and in line with prior work (M. Park et al., 2012).
  - *Affect*: Similarly, the model links a higher degree of affect in speech (*Negative/Positive Tone*) to a higher likelihood of depression (Wolohan et al., 2018).
  - *Swear Words*: Literature on depression indicates that people suffering from depression more frequently use swear words (A. Park, Conway, et al., 2017). The coefficient is equally positive and significant.
  - *Friends & Female*: Interestingly, a more frequent use of words referring to females and friends – e.g. girlfriend – also correlate to higher model certainty for a positive case.
  - *Power & Curiosity*: The negative coefficients of power and curiosity indicate that the more such words a person uses c.p., the less confidence MentalBERT has in a positive classification.
- **Expanded dictionary:**
  - *Substances*: The more frequent use of substance words – e.g. beer or drunk – is associated with a more confident decision about a positive prediction.
  - *Lack*: Deficiency refers to the state of having a deficiency – e.g., do not have or being hungry. The coefficient shows a positive correlation between a positive estimate and a more frequent use of expressions describing a deficiency state.
- **Authenticity:** Authenticity measures perceived spontaneity and absence of self-filtering in texts (Newman et al., 2003). MentalBERT associates higher levels of authenticity with a negative prediction, contrasting with earlier work (Wolohan et al., 2018).

Overall, most of the features exposed in the regression model are in accordance with previous empirical research. Surprisingly, however, authenticity appears to be negatively correlated with the model estimate of depression likelihood. This could be related to the nature of topics discussed on Reddit or/and the subjects' characteristics. Nevertheless, although the LIWC features account for substantial variation in MentalBERT's estimates (72%), the 28% gap in the R-squared metric suggests that the Transformer model is actually capturing additional information not modeled by the LIWC features. We pick up on this point and extend the analysis by applying advanced topic modeling techniques. More concretely, since we are interested in gaining further insight into what MentalBERT is learning differently, we analyze the most common topics among the depression-susceptible individuals that are correctly classified as such by the MentalBERT model but incorrectly labeled by the best-performing LIWC-based approach. To proceed, we use BERTopic – i.e., a state-of-the-art topic modeling technique that uses pretrained Transformers to build document embeddings, groups them into dense clusters, and applies a class-based TF-IDF to create topic representations (Grootendorst, 2022).

Table 5 shows the most commonly discussed topics and their top keywords. It can be easily observed that the topics cover neutral rather than depression-focused content, but at the same time, the topics relate to current and general life matters. *Topic 1* touches on issues related to relationships and social aspects. *Topic 2*, on the other hand, revolves around ethnic-related concerns. *Topic 3* contains words referring merely to physical appearance and may reflect thoughts about self-perception. While *Topic 4* seems to deal with the discussion of current problems, *Topic 5* includes swear words. *Topic 6* describes time-related aspects referring to symptoms – e.g., "tired" or "sleep" – and finally, *Topic 7* includes positive words that could express mutual appreciation of help among members.

Table 5. Most commonly discussed topics

Topic	Top Keywords
Topic 1	women, men, want, sex, attracted, confidence, guys
Topic 2	black, white, racist, racism, country, police, history
Topic 3	hair, eyes, ugly, acne, skin, cute, nose, cutie, smile
Topic 4	things, reason, talk, specific, problem, news, authority
Topic 5	swear words
Topic 6	sleep, woke, bed, hours, wake, tired, minutes, time
Topic 7	thank, thanks, helped, welcome, correcting, fantastic

5. Conclusion

In summary, this paper demonstrates the potential of harnessing social media to improve mental eHealth awareness. First, we used Reddit to build a text corpus with a cohort of clinically depressed and non-depressed subjects and aimed to identify depression-prone individuals with minimal direct depressive signals at an early stage. The considerably high predictive accuracy achieved by our transformer-based approach yields promising results for the early detection of depression and indicates differences in word use between both cohorts of subjects. Second, we leveraged a surrogate and a topic modeling technique to assess the interpretability of the model.

Nonetheless, the empirical results reported herein should be considered in light of some limitations. First, the ground truth annotations in the corpus are established on self-disclosure expressions. This means we can not exclude the possibility of having depressed individuals in the non-depressed class and non-depressed people in the depressed one. Still, we believe this to be the exception, not the norm and argue that other traditional screening methods, such as surveys, are noisy for the same reason (Losada & Crestani, 2016). Similarly, predictive algorithmic screening, surveys, and human assessments could be employed as complementary approaches to yield better screening outcomes for both practical and academic goals. A professional clinical evaluation remains, however, necessary for a final diagnosis. Second, because of the nature of the data on Reddit, no demographic or clinical information about the subjects was available and hence disregarded from the study. In addition, for the sake of generalization, the proposed approach should be tested and evaluated on different depression-related datasets.

We hope that providing automated approaches capable of identifying at-risk individuals earlier will help e-peer communities to more efficiently channel access to social capital and reach the affected sooner. Future efforts could be directed towards exploring incidental behaviors related to mental health, such as help-seeking behavior, and understanding what factors affect levels of perceived helplines. By the same token, further work, could, for instance, examine the effectiveness of e-peer support communities for depression using advanced data-driven approaches. Besides, this study does not discuss or address privacy concerns, but simply provides a mechanism for early detection. If individuals at risk for depression do not wish to be identified, considerations about privacy violation become even more critical. This aspect deviates from the objective of this study but is important for future research.

Ultimately, based on these findings, we view our study as a further step toward developing a better understanding of mental eHealth and hope our results can support the development and application of future technologies in this field.

References

Abdel-Khalek, A. M. (2004). Can somatic symptoms predict depression? *Social Behavior and Personality: an international journal*, *32*(7).

Ali, K., Farrer, L., Gulliver, A., & Griffiths, K. M. (2015). Online Peer-to-Peer Support for Young People With Mental Health Problems: A Systematic Review. *JMIR Mental Health*, *2*(2).

Beck, A. T. (1979). *Cognitive therapy of depression*. Guilford Press.

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). The Development and Psychometric Properties of LIWC-22. *Austin, TX: University of Texas at Austin*.

Brown, G. W., Andrews, B., Harris, T., Adler, Z., & Bridge, L. (1986). Social support, self-esteem and depression. *Psychological medicine*, *16*(4).

Carey, J. L., Carreiro, S., Chapman, B., Nader, N., Chai, P. R., Pagoto, S., & Jake-Schoffman, D. E. (2018). SoMe and Self Harm: The use of social media in depressed and suicidal youth. *Proceedings of the 51st Hawaii International Conference on System Science*, *2018*.

Caron, M., Bäumer, F. S., & Müller, O. (2022). Towards Automated Moderation: Enabling Toxic Language Detection with Transfer Learning and Attention-Based Models. *Proceedings of the 55th Hawaii International Conference on System Science*.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Chen, X., Sykora, M., Jackson, T., Elayan, S., & Munir, F. (2018). Tweeting Your Mental Health: an Exploration of Different Classifiers and Features with Emotional Signals in Identifying Mental Health Conditions. *Proceedings of the 51st Hawaii International Conference on System Science*.

Coppersmith, G., Dredze, M., & Harman, C. (2014). Quantifying Mental Health Signals in Twitter. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*.

Cox, D. R. (1958). The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, *20*(2).

De Choudhury, M., & De, S. (2014). Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity. *Eighth International AAAI Conference on Weblogs and Social Media*.

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting Depression via Social Media. *Seventh International AAAI Conference on Weblogs and Social Media*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Ettman, C. K., Cohen, G. H., Abdalla, S. M., Sampson, L., Trinquart, L., Castrucci, B. C., Bork, R. H., Clark, M. A., Wilson, I., Vivier, P. M., et al. (2022). Persistent depressive symptoms during covid-19: A national, population-representative, longitudinal study of us adults. *The Lancet Regional Health-Americas*, *5*.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Gujarati, D. N. (2021). *Essentials of Econometrics*. SAGE Publications.

Ho, T. K. (n.d.). Random Decision Forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*.

Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., & Cambria, E. (2022). MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*.

Kotikalapudi, R., Chellappan, S., Montgomery, F., Wunsch, D., & Lutzen, K. (2012). Associating Depressive Symptoms in College Students with Internet Usage Using Real Internet Data. *IEEE Technology and Society Magazine*.

Losada, D. E., & Crestani, F. (2016). A Test Collection for Research on Depression and Language Use. *International Conference of the Cross-Language Evaluation Forum for European Languages*.

Losada, D. E., Crestani, F., & Parapar, J. (2018). Overview of eRisk: Early Risk Prediction on the Interne. *International Conference of the Cross-Language Evaluation Forum for European Languages*.

Manikonda, L., & De Choudhury, M. (2017). Modeling and Understanding Visual Attributes of Mental Health Disclosures in Social Media. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*.

Manos, R. C., Rusch, L. C., Kanter, J. W., & Clifford, L. M. (2009). Depression Self-Stigma as a Mediator of the Relationship Between Depression Severity and Avoidance. *Journal of Social and Clinical Psychology*, *28*(9).

Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*.

Monselise, M., & Yang, C. C. (2022). "I'm always in so much pain and no one will understand" - detecting patterns in suicidal ideation on reddit. *Companion Proceedings of the Web Conference 2022*.

National Institute of Mental Health. (2019). Depression. https://www.nimh.nih.gov/health/topics/depression

Nayak, P. (2019). Understanding searches better than ever before. https://blog.google/products/search/search-language-understanding-bert/

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying Words: Predicting Deception from Linguistic Styles. *Personality and social psychology bulletin*, *29*(5).

Orabi, A. H., Buddhitha, P., Orabi, M. H., & Inkpen, D. (2018). Deep Learning for Depression Detection of Twitter Users. *Proceedings ofs the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*.

Pan, S. J., & Yang, Q. (2009). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10).

Park, A., Conway, M., et al. (2017). Longitudinal changes in psychological states in online health community members: Understanding the long-term effects of participating in an online depression community. *Journal of medical Internet research*, *19*(3).

Park, M., Cha, C., & Cha, M. (2012). Depressive moods of users portrayed in Twitter. *Proceedings of the 18th ACM International Conference on Knowledge Discovery and Data Mining*.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The Development and Psychometric Properties of LIWC2015.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, *54*(1).

Pudil, P., Novovičová, J., & Kittler, J. (1994). Floating search methods in feature selection. *Pattern recognition letters*, *15*(11).

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). *Science China Technological Sciences*, *63*(10).

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

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*.

The Lancet Global Health. (2020). Mental health matters. *The Lancet Global Health*, *8*(11).

Trotzek, M., Koitka, S., & Friedrich, C. M. (2017). Linguistic Metadata Augmented Classifiers at the CLEF 2017 Task for Early Detection of Depression. *CLEF*.

Van Wert, M. J., Gandhi, S., Gupta, I., Singh, A., Eid, S. M., Haroon Burhanullah, M., Michtalik, H., & Malik, M. (2022). Healthcare Worker Mental Health After the Initial Peak of the COVID-19 Pandemic: a US Medical Center Cross-Sectional Survey. *Journal of general internal medicine*, *37*(5).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*.

Wang, X., Zhang, C., Ji, Y., Sun, L., Wu, L., & Bao, Z. (2013). A Depression Detection Model Based on Sentiment Analysis in Micro-blog Social Network. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*.

Wolohan, J., Hiraga, M., Mukherjee, A., Sayyed, Z. A., & Millard, M. (2018). Detecting Linguistic Traces of Depression in Topic-Restricted Text: Attending to Self-Stigmatized Depression with NLP. *Proceedings of the First International Workshop on Language Cognition and Computational Models*.

World Health Organization. (2022). https://www.who.int/health-topics/depression

Yates, A., Cohan, A., & Goharian, N. (2017). Depression and Self-Harm Risk Assessment in Online Forums. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.