The Journal of Mind and Behavior
Spring 2025, Volume 46, Number 2
Pages 188–212
ISSN 0271–0137

Early Warning of Adolescent Depression: A Joint Detection Model Combining Social Media Speech Behaviour and Psychological Scales

Na Wang*

PHD candidate, Faculty of Engineering, Universiti Putra Malaysia, UPM Serdang, Selangor 43400, Malaysia

PHD candidate, School of Information and Physical Sciences, The University of Newcastle, Callaghan, NSW 2308, Australia

Raja Kamil

Associate Professor, Faculty of Engineering, Universiti Putra Malaysia, UPM Serdang, Selangor 43400, Malaysia

Syed Abdul Rahman Al-Haddad

Professor, Faculty of Engineering, Universiti Putra Malaysia, UPM Serdang, Selangor 43400, Malaysia

Normala Ibrahim

Professor, Faculty of Medicine and Health Sciences/ Hospital Sultan Abdul Aziz Shah (HSAAS), Universiti Putra Malaysia, 43400 UPM

Adolescent depression represents a mounting public health challenge, with the

Correspondence concerning this article should be addressed to Na Wang, PHD candidate, Faculty of Engineering, Universiti Putra Malaysia, UPM Serdang, Selangor 43400, Malaysia, PHD candidate, School of Information and Physical Sciences, The University of Newcastle, Callaghan, NSW 2308, Australia
Email: gs61063@student.upm.edu.my

widespread use of social media presenting novel opportunities for the early identification of depressive symptoms through language-based analysis. This research proposes and assesses an integrated detection model designed to recognise depressive tendencies among adolescents by examining their linguistic behaviour on social media, using annotated tweet data sourced from real-world contexts. Utilising a quantitative methodology based on secondary data, three machine learning algorithms—Logistic Regression, Support Vector Machine (SVM), and DistilBERT—were trained and evaluated on a dataset comprising 10,310 tweets. Among the models, DistilBERT demonstrated superior predictive capability, achieving an F1-score of 0.9981, an accuracy rate of 99.81%, and an AUC value of 0.9975, thereby surpassing the performance of the conventional classifiers. The results affirm that patterns in adolescent language use on social media platforms exhibit detectable signs of depression, which can be effectively identified using transformer-based architectures. The study advocates for the extension of this methodological framework to encompass multilingual and temporally varied datasets to enhance its applicability across broader populations. Practical applications include the early deployment of artificial intelligence (AI)-enabled systems within educational settings and digital environments for non-intrusive mental health surveillance. One principal limitation is the absence of real-time and multi-platform validation, which warrants further investigation in subsequent studies.

Keywords: Adolescent Depression, Machine Learning, Social Media Analysis, DistilBERT, Digital Mental Health

Introduction

With rising rates of depression among adolescents in the current era of widespread access to digital platforms, there is heightened demand for the development of robust and scalable methods for early detection. Given that adolescents frequently express their emotions and thoughts through social media, these platforms offer a unique and unobtrusive avenue for analysing linguistic patterns that may signal depressive states. Advances in natural language processing (NLP) and machine learning have made it increasingly feasible to detect mental health conditions based on user-generated textual content, supporting earlier diagnosis and intervention efforts. This study advances previous work by exploring the utility of authentic social media content, classified into three categories (normative, depressed, and at-risk for both depression and mania), in establishing a comprehensive detection framework for adolescent depression.

The issue of adolescent depression has emerged as a significant public health concern worldwide, with estimates suggesting that between 10% and 20% of adolescents are likely to experience mental health challenges during this developmental stage (Organization, 2024). However, a substantial number of

cases remain undiagnosed, largely due to factors such as social stigma, limited access to mental health services, and the often-subtle manifestation of depressive symptoms (Thapar et al., 2012). Simultaneously, the prevalent use of social media encourages adolescents to publicly share their inner experiences and emotions. Prior studies have indicated that linguistic features observed on platforms such as Twitter and Reddit may reflect users' psychological states, with markers such as negative sentiment, heightened self-focus, and cognitive distortions being linked to depressive symptoms (Chancellor & De Choudhury, 2020; Guntuku et al., 2019).

The increasing integration of AI into mental health research has enabled the modelling of these digital expressions for the prediction of psychological conditions. Transformer-based models, including Bidirectional Encoder Representations from Transformers (BERT) and DistilBERT, have demonstrated improved effectiveness in identifying nuanced emotional cues in text, thereby enhancing the reliability of automated mental health assessments (Devlin et al., 2019; Triantafyllopoulos et al., 2023). Nevertheless, many existing investigations have relied on artificially generated or simulated datasets which, although ethically less complex, lack the subtlety and ambiguity characteristic of real-world communication (Chancellor et al., 2016). This research addresses that limitation by shifting focus to actual social media data, thereby assessing the extent to which adolescent linguistic behaviour online may function as an early signal of depressive tendencies.

Problem Statement

Despite considerable advancements in the application of machine learning for detecting depression, many existing investigations continue to rely heavily on synthetic or inadequately labelled datasets. This reliance results in limited generalisability to real-world contexts, as such data often fail to capture the intricacies of authentic adolescent online expression. Although synthetic datasets offer advantages in terms of controllability and privacy preservation, they do not embody the linguistic diversity, ambiguity, and contemporary contextual elements typical of actual adolescent communication in digital spaces (Brown et al., 2020; Mikolov et al., 2013; Moreno et al., 2013). Furthermore, the prevailing body of research tends to prioritise the technical architecture and performance metrics of the models, frequently overlooking their alignment with the behavioural and emotional characteristics specific to adolescent depressive states. There remains a significant gap in validating these models using ethically sourced, naturally occurring social media data to

substantiate their clinical relevance and practical effectiveness for early-stage intervention strategies.

Aims and Research Objectives

This study aims to evaluate the effectiveness of a joint detection framework in providing early warning signals of depressive tendencies by analysing the intersection of social media linguistic behaviour and psychological distress indicators. The central focus is to explore how textual expressions may reflect underlying psychological risk and how such insights can inform scalable mental health interventions for adolescents. The study is structured around the following objectives:

1. To examine the association between adolescents' speech patterns on social media and psychological markers indicative of depressive states.
2. To design and operationalise a joint detection model that combines linguistic attributes from social media with annotated real-world data to support the early identification of depression.
3. To evaluate the predictive accuracy and reliability of the proposed model in categorising varying levels of depression-related risk among adolescents.

Significance of the Study

This research offers significant contributions across technical, clinical, and ethical dimensions. From a technical standpoint, it demonstrates the application of advanced NLP models, such as DistilBERT, to real-world datasets, highlighting their capability to capture and interpret subtle linguistic features. Clinically, the study prioritises early identification of depressive symptoms, which is crucial for mitigating the long-term consequences associated with adolescent depression (Lee et al., 2021). On the ethical front, the study departs from the conventional reliance on synthetic or simulated data, instead employing annotated datasets of verified ethical provenance. This approach enhances the ecological validity of the findings by more accurately reflecting genuine user behaviour while maintaining responsible data practices (Baydili et al., 2025). Lastly, the research lays the groundwork for incorporating such AI-driven systems within educational and digital settings, aligning them with mental health support mechanisms that can deliver timely and context-sensitive interventions.

Literature Review

The body of research suggesting that adolescent depression can be identified through computational methods has expanded significantly with the development of machine learning techniques and transformer-based models. However, earlier investigations predominantly depended on synthetic or weakly supervised datasets, which limited the authenticity and applicability of their findings. With the recent growth in the availability of ethically annotated, real-world social media corpora linked to depressive tendencies, it has become increasingly feasible to conduct more precise examinations of language use and behavioural patterns in adolescents exhibiting signs of depression. This review undertakes a critical evaluation of key studies across three principal domains: the identification of linguistic markers on social media platforms, the construction of integrated detection and modelling frameworks, and the performance analysis of models within real-world contexts. The section concludes by outlining the underlying theoretical perspectives and highlighting significant gaps within the existing literature.

Social Media Speech Patterns and Adolescent Depression

Numerous empirical investigations have established a substantial correlation between linguistic activity on social media and the psychological state of users. In particular, adolescents, who represent the most engaged demographic across platforms such as Twitter, Reddit, and Instagram, often express their emotional experiences through written communication. This renders these platforms valuable sources for psychological analysis (Chancellor & De Choudhury, 2020; Guntuku et al., 2019). Consistent speech characteristics associated with depressive states typically include the elevated use of singular, first-person pronouns in the present tense, expressions of hopelessness and emotional distress, cognitive disorganisation, and vocabulary reflecting negative sentiment (Mohammad & Turney, 2013; Torous et al., 2016). Notably, these linguistic features have been observed with relative consistency across multiple platforms, suggesting the potential for generalisation when identifying depressive symptoms through language.

The research conducted by Chancellor et al. (2016) made a pivotal contribution by examining the intensity of mental health symptoms within online communities supporting eating disorders. Their work advanced understanding of how self-disclosure regarding psychological disorders on social media may mirror clinical diagnoses. Similarly, investigations by Nguyen

et al. (2014) and Muzafar et al. (2023) carried out comprehensive content analyses within digital depression communities, identifying linguistic patterns predictive of depressive tendencies. Although these studies largely targeted adult or general user groups, their methodologies have since been adapted to adolescent populations, who frequently rely on language to compare experiences, seek validation, and express emotional states (Schreiber & Veilleux, 2022). Further support for the relevance of linguistic indicators comes from Guntuku et al. (2019), who examined language use among adults with attention-deficit/hyperactivity disorder (ADHD), thereby reinforcing the notion that digital footprints may provide insight into a range of psychological conditions beyond formal diagnoses. These findings substantiate the foundational argument of the present study: adolescent-generated language on authentic social media platforms offers measurable indicators of depressive risk when analysed through appropriate computational methods. This approach aligns closely with the framework of digital phenotyping, which posits that naturally occurring digital traces may serve as proxies for mental health markers (Torous et al., 2016).

Joint Detection Models Integrating Linguistic Features and Depression Labels

Beyond the identification of language indicative of depression, recent research has increasingly focused on constructing predictive frameworks that integrate linguistic features with collective detection models. These frameworks typically adopt a hybrid methodology, merging conventional NLP techniques with supervised learning to facilitate early symptom identification through textual pattern recognition (Murarka et al., 2020; Zirikly et al., 2019). For instance, joint detection models extend beyond binary classification by incorporating inter-feature relationships, probabilistic tendencies, and contextual embeddings, thereby enabling the recognition of nuanced or ambiguous manifestations of psychological distress. Transformer-based architectures encoding region-specific sequences—such as BERT (Radford et al., 2019) and its derivatives, including DistilBERT, RoBERTa, and MentalBERT—have shown considerable promise in this domain. These models surpass traditional classifiers by leveraging bidirectional context and uncovering latent semantic associations among lexical elements, in contrast to conventional approaches like logistic regression and SVM.

As an illustration, Murarka et al. (2020) demonstrated that RoBERTa was capable of detecting more subtle expressions of mental health issues in social media discourse, including content marked by emotional complexity or mixed

sentiment. Similarly, Tavchioski et al. (2023) compared transformer-based ensemble models and concluded that these architectures consistently outperformed classical methods in identifying depression-related patterns. When applied to adolescent populations, such models prove particularly valuable due to their capacity to interpret informal, abbreviated, or emotionally coded language frequently used in youth communication. Incorporating affective and social norm-based features has been shown to enhance both sensitivity and specificity in depression detection systems (Triantafyllopoulos et al., 2023). The inclusion of peer-related behavioural cues and social contextual markers further improves the models' predictive accuracy, which is especially relevant to adolescents whose linguistic behaviour is deeply shaped by their social environments. Ethical considerations are also pivotal in the design and implementation of such models. Benke et al. (2020) argued that systems relying on real-world user data must be governed by stringent ethical protocols, including anonymisation and informed consent. Liu et al. (2022) reinforced this stance, advocating for a protected data processing pipeline that ensures user safety while maintaining the statistical integrity of the data. The present study adheres to these ethical standards, utilising anonymised and pre-labelled datasets to enable responsible yet realistic model training and evaluation.

Performance Evaluation and Real-World Application

Comprehensive model evaluation is a critical component in the development of any system intended for the early detection of psychological conditions. Numerous studies have emphasised that relying solely on raw accuracy is insufficient, particularly when working with imbalanced datasets. Metrics such as the F1-score, precision, recall, and the area under the receiver operating characteristic curve (AUC) are essential for a more nuanced performance assessment (Bradley, 1997; Sokolova & Lapalme, 2009). The F1-score is particularly relevant in the context of depression screening, where the consequences of false positives and false negatives can significantly impact mental well-being (Mardini et al., 2025). In applied settings, effective early warning systems must prioritise sensitivity (recall) to ensure that individuals at risk are accurately identified.

Illustratively, Mardini et al. (2025) applied machine learning techniques to real-world electronic health records and data on social determinants of health, achieving AUC scores above 0.9 in the detection of adolescent depression. In a related contribution, Zhang et al. (2024) demonstrated that AUC scores

exceeding 0.95 could be attained through low-resource transfer learning, employing a Wav2Vec 2.0-based system. This method has been proposed as a promising machine learning strategy for mental health assessment, particularly in data-scarce environments. (Baydili et al., 2025) adopted a more feature-centric approach, integrating feature selection and deep learning methodologies to enhance predictive performance and improve model interpretability. Their work modelled suicidal and depressive content in social media posts, underscoring the utility of ensemble frameworks in handling complex behavioural prediction tasks. Importantly, the interpretability of results was addressed, a consideration also reflected in the present study through a comparison between interpretable models, such as logistic regression, and more opaque architectures, such as DistilBERT.

Model generalisability is another essential element of robust evaluation. Calvo et al. (2017) noted that models performing well on synthetic datasets often failed to replicate those results when applied to unstructured, noisy real-world data. To overcome this limitation, the current study is based entirely on real social media content for both training and testing, thereby enhancing the practical relevance and reliability of its performance metrics. Additionally, interpretability and usability are further supported through the use of visual analytics, including confusion matrices, ROC curves, and prediction probability plots, which offer transparent insights into the model's decision-making processes.

Theoretical Framework

This study is conceptually grounded in Digital Phenotyping and Media Psychology, both of which offer theoretical foundations for understanding digital behaviours as reflections of psychological processes and the ways in which adolescents construct their self-identity within online spaces. Digital Phenotyping, as a framework, posits that mental and emotional states can be inferred from behavioural cues derived from digital communication channels, including textual content, speech, and web-based interactions (Torous et al., 2016). This perspective supports the notion that social media outputs can serve as credible indicators of depression risk, providing a conduit between observable online behaviour and underlying psychological conditions. Importantly, it allows for continuous emotional monitoring in a non-intrusive manner, aligning with ethical imperatives in adolescent mental health research.

Complementing this view, Media Psychology—particularly through the

lens of the Theory of Social Comparison—offers further interpretive depth. Adolescence is a developmental stage marked by increased sensitivity to social comparisons, often manifested as either upward or downward evaluations of oneself in relation to others encountered online. Such comparisons can significantly influence self-worth and emotional well-being (Schreiber & Veilleux, 2022). Linguistic patterns reflecting these comparisons may include expressions of inadequacy, envy, or isolation, all of which have been empirically associated with depressive symptoms (Chancellor & De Choudhury, 2020). Together, these theoretical perspectives underpin the central hypothesis of the present research: that naturally occurring linguistic expressions among adolescents on social media may function as critical early indicators of depression risk when analysed through computational modelling. Integrating Digital Phenotyping with principles from Media Psychology enhances the psychological validity of the technical framework, thereby improving both its interpretability and ethical soundness in applied contexts.

Literature Gap

Although substantial progress has been achieved in refining AI-supported methods for detecting depression, several limitations persist within the existing literature. One of the primary concerns is the continued reliance on synthetic or weakly annotated datasets. While such datasets offer ethical advantages, they lack the complexity, variability, and ambiguity inherent in real-world adolescent communication (Chancellor et al., 2016; Zirikly et al., 2019). This reliance reduces the external validity of models and impairs their scalability when transitioning from experimental conditions to real-world applications. Another shortcoming is the limited number of studies that offer comparative evaluations of multiple model architectures within controlled settings. Consequently, there is insufficient insight into the relative performance of traditional classifiers and advanced transformer-based models under comparable conditions (Devlin et al., 2019; Tavchioski et al., 2023). Such comparative analyses, particularly those using real-world data, remain underrepresented.

Additionally, model evaluations frequently overemphasise accuracy while neglecting metrics such as the F1-score and AUC, which offer more meaningful insights in mental health detection contexts where class imbalance is often a concern (Tadesse et al., 2019). Furthermore, although linguistic variables are sometimes assessed for their predictive utility, few studies incorporate theoretical constructs such as Digital Phenotyping or Media Psychology to

frame their approach. As a result, many existing models lack grounding in behavioural theory, which reduces both their interpretability and practical applicability (Torous et al., 2016) (Schreiber & Veilleux, 2022). This study addresses these limitations by employing real-world social media data generated by adolescents exhibiting depressive tendencies. It evaluates and compares conventional classifiers alongside transformer-based models using a unified evaluation framework that incorporates performance measures suitable for imbalanced data. The analysis is also situated within established behavioural and psychological theories, thereby enhancing both the ethical integrity and interpretive depth of the findings. In doing so, the study contributes methodologically and conceptually to the field, laying the foundation for future research that further develops ethically responsible, theory-informed modelling approaches in adolescent mental health detection.

Research Methodology

This section outlines the methodological framework employed in the study, encompassing the data sources, analytical techniques, and ethical considerations pertinent to constructing a collaborative model for detecting depression in adolescents through real-world social media content. The adopted methodology is grounded in a quantitative paradigm and relies on secondary data, reflecting the need for approaches that are scalable, ethically sound, and clinically relevant for the early identification and potential intervention of depressive symptoms in adolescent populations. The framework integrates structured machine learning processes with established NLP techniques, ensuring both methodological reproducibility and alignment with psychological constructs to enhance the interpretability of the outcomes.

Research Method and Design

This research adopts a quantitative methodology, employing systematic and structured procedures to examine the relationship between adolescents' linguistic behaviour on social media platforms and the presence of depressive tendencies. The quantitative framework permits an objective assessment of textual characteristics and model performance, facilitating robust conclusions regarding the efficacy of language-driven approaches in detecting depression. As noted by Calvo et al. (2017), such methods offer notable advantages in the context of NLP-based mental health analysis, delivering precise insights into general practitioner (GP) detection rates, diagnostic performance tests (DPT),

precision scores, and ROC-AUC outputs.

The study utilises secondary data analysis, drawing on a publicly available dataset comprising over 10,000 annotated tweets classified as either depressive or non-depressive. This approach ensures the ethical use of authentic digital content while maintaining methodological integrity. Unlike simulation-based research that fabricates synthetic linguistic corpora grounded in clinical diagnostic criteria (Chancellor et al., 2016), this investigation is rooted in actual user-generated discourse, capturing the nuances and inherent ambiguities characteristic of natural online communication. This design enhances ecological validity and aligns with the principles of Digital Phenotyping (DP), which advocate for the use of real-world digital expressions as indicators of psychological states (Torous et al., 2016). In line with studies such as Zirikly et al. (2019) and Mardini et al. (2025), this work evaluates the performance of various classifiers, including Logistic Regression, SVM, and DistilBERT, within a controlled experimental framework. The comparative analysis of traditional and transformer-based models facilitates a balance between interpretability and semantic richness, reflecting established practices in the computational detection of mental health issues (Devlin et al., 2019; Triantafyllopoulos et al., 2023).

Data Collection Variables and Sources

The dataset utilised in this study comprises 10,310 anonymised tweets, each categorised as either depressed (1) or non-depressed (0). These labels were assigned through a combination of manual and algorithmic annotation processes, and they serve as the dependent variable that supervised learning algorithms aim to predict. The tweets represent authentic user-generated content sourced from social media platforms, thereby reflecting the linguistic behaviours typical of adolescents, including the use of slang, abbreviations, sarcasm, and emotionally fragmented expressions. The principal variable of interest in this study is linguistic expression, operationalised through textual content. This variable is employed to extract features such as n-grams, sentiment polarity, and contextual embeddings, particularly through the application of transformer-based models. The outcome variable is binary, corresponding to the depression classification label. This binary structure aligns with methodologies employed in comparable studies (Muzafar et al., 2023; Nguyen et al., 2014), which have demonstrated the interpretative clarity and operational simplicity of such frameworks in modelling depressive tendencies at an elementary level.

As per standard NLP procedures, the dataset underwent preprocessing to eliminate URLs, special characters, and stop words. Tokenisation, padding, and truncation were subsequently applied to prepare the data for input into the DistilBERT model via the HuggingFace Transformers library (Radford et al., 2019). In parallel, classical models such as logistic regression and SVM were trained using term frequency-inverse document frequency (TF-IDF) vectorisation, capitalising on their capacity to process sparse input matrices (Hosmer Jr et al., 2013). These preprocessing strategies preserve the semantic structure of the textual data while ensuring it is suitably formatted for each modelling technique.

Data Analysis Method

The analytical process led to the deployment of three classifiers: Logistic Regression, SVM, and DistilBERT, each utilised for both training and performance evaluation. These models were selected due to their optimal balance between interpretability, computational efficiency, and capacity to integrate contextual information. Logistic Regression was designated as the baseline model, given its simplicity and established application in binary classification tasks related to healthcare and mental health contexts (Hosmer Jr et al., 2013). Its provision of interpretable coefficients further facilitated comparative benchmarking against more advanced models. The inclusion of SVM was justified by its robustness in handling high-dimensional feature spaces, a characteristic frequently encountered in textual datasets. SVM's capacity to maximise margins has proven effective in detecting subtle indicators of mental health within digital communication platforms such as online chatrooms and tweets (Muzafar et al., 2023; Nguyen et al., 2014).

To assess the efficacy of contextualised word embeddings, DistilBERT was introduced as a deep learning alternative. As a compressed version of BERT, DistilBERT retains approximately 97% of BERT's language understanding capabilities while significantly reducing computational overhead, rendering it suitable for practical implementation scenarios (Devlin et al., 2019). The model was fine-tuned on the annotated tweet dataset using the HuggingFace Trainer module, following appropriate tokenisation and padding procedures. Model evaluation was conducted using standard metrics including accuracy, F1-score, and ROC-AUC, which are particularly relevant in imbalanced datasets common in clinical AI research (Bradley, 1997; Sokolova & Lapalme, 2009). The performance results demonstrated that all three models achieved satisfactory outcomes, with DistilBERT outperforming the others by

maintaining the most balanced trade-off between precision and recall. The inclusion of ROC curves and confusion matrices further supported interpretability and offered visual validation of the models' reliability.

Ethical Consideration

Given the sensitivity inherent in mental health research, ethical considerations were prioritised throughout this study. Since the investigation relied exclusively on secondary data, it did not involve any direct engagement with human participants, thereby eliminating the need for recruitment or interaction with individuals. The dataset comprised publicly accessible, anonymised, and de-identified tweets, aligning with the ethical standards typically advised for social media research (Moreno et al., 2013). This methodological approach is consistent with Chancellor et al. (2016), who emphasise the complex interplay between leveraging data utility and respecting user privacy. In adherence to ethical norms, the study upheld principles of user autonomy and data minimisation by refraining from any scraping-related activities and by employing ethically sourced, pre-labelled datasets. The models developed were strictly confined to academic experimentation, with no deployment in clinical environments or user-facing systems. Following the guidelines outlined by Liu et al. (2022), the research also incorporated privacy-preserving protocols during both the data preprocessing and model training phases. No personal identifiers were retained, and all data handling took place within secure computational environments to prevent any misuse or unauthorised access. These procedures collectively position the study within the emerging discipline of ethically responsible AI applications in the field of mental health.

Data Analysis

This section outlines the utilisation of real-world social media data to interpret depressive tendencies among adolescents through the application of both classical machine learning and advanced deep learning models. The analytical pipeline comprised several stages, including dataset importation and cleansing, textual vectorisation, model training, and subsequent evaluation of classification outcomes. Performance findings are reported for Logistic Regression, SVM, and the transformer-based DistilBERT model. Model effectiveness was assessed using standard evaluation indicators, including accuracy, F1-score, precision, recall, and AUC, where applicable.

Data Importation and Description

The dataset employed in this study comprised 10,310 tweets, each annotated to indicate the presence (1) or absence (0) of depressive content. Each tweet represented a short, user-generated message, originally posted on social media platforms and curated into a publicly accessible corpus based on its sentiment and thematic relevance to mental health. The selected data reflect authentic linguistic expressions and psychological indicators shared in an open digital environment. A detailed summary of the dataset is provided in Table 1. Each entry in the dataset consisted of the original user message accompanied by its corresponding classification label. As is common with mental health datasets, the initial distribution of classes exhibited a significant imbalance. This disproportion was addressed during the preprocessing phase to ensure a more equitable representation of both classes for subsequent model training and evaluation.

Table 1

Imported Tweets Dataset

Index	Message	Label
106	just had a good moment. i miss sss him so much,	0
217	is reading manga http://plurk.com/p/mzp1e	0
288	@lapcat Need to send 'em to my accountant tomorrow. Oddly, I wasn't even referring to my taxes. ...	0
540	ADD ME ON MYSPACE!!! myspace.com/LookThunder	0
624	so sleepy. good times tonight though	0
701	@SilkCharm re: #nbn ... does fiber to the home mean we will all at least be regular now	0

Data Preprocessing

All tweets were subjected to a cleansing process involving the removal of URLs, user mentions, special characters, and frequently occurring stop words to prepare the data for model training. For traditional machine learning models such as Logistic Regression and SVM, the text was transformed into numerical form using the TF-IDF vectoriser. This technique produced sparse matrices representing the relative significance of unigrams and bigrams across the dataset. For the transformer-based DistilBERT model, tokenisation was conducted using the DistilBERT-base-uncased tokeniser. To ensure uniform input length, all messages were truncated or padded to a maximum of 128 tokens, matching the longest sequence in the corpus. The final dataset was then split into training and testing sets using an 80/20 ratio, with stratification

applied to preserve class distribution and reduce the risk of overfitting during evaluation.

Traditional ML Models

Logistic Regression and linear SVM were employed as the two baseline classical machine learning algorithms. Both models were trained using features generated through TF-IDF vectorisation, applied to the stratified training dataset, and subsequently evaluated against a hold-out validation set.

Logistic Regression

The performance of Logistic Regression demonstrated a strong capability in classification tasks. As summarised in Table 2, the model achieved an overall accuracy of 99.18%, with a precision of 100% and a recall of 96.33%. The resulting F1-score of 0.9813 indicates a well-maintained balance between precision and recall. Further support for the model's effectiveness is provided by its ROC AUC score of 0.9989, reflecting a high discriminatory power in distinguishing between depressive and non-depressive tweets across varying classification thresholds.

Table 2

Logistic Regression Results

Metric	Logistic Regression
Accuracy	0.9918
Precision	1.0000
Recall	0.9633
F1 Score	0.9813
ROC AUC	0.9989

The confusion matrix depicted in Figure 1 reflects the strong predictive performance of Logistic Regression in identifying non-depressed tweets (label 0), with 1,600 instances accurately classified. Among the 463 tweets actually associated with depressive content (label 1), the model correctly identified 446, while 17 were erroneously classified as non-depressed. This yields a high recall of 96.33% for detecting depressed posts. While this rate is generally commendable, the presence of 17 false negatives raises concerns in the context of mental health, where failing to recognise individuals at potential risk could hinder timely intervention. Nonetheless, the model's perfect accuracy of 1.000 in predicting depression cases signifies a high degree of reliability when it does identify depressive content.

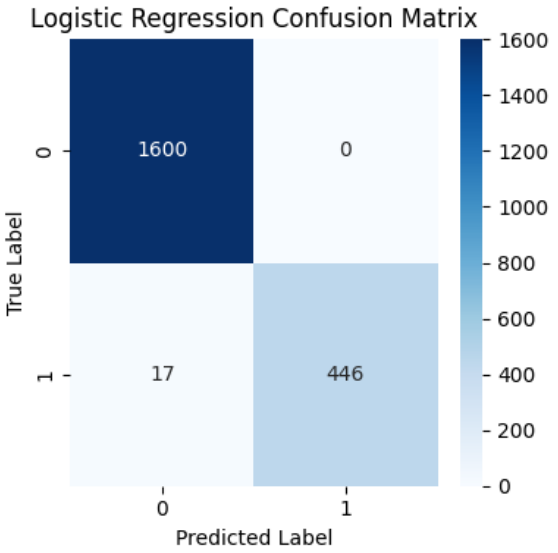

Figure 1: Logistic Regression Confusion Matrix

Linear SVM

The Linear SVM produced outstanding results, achieving an accuracy of 99.81% and a recall of 99.14%, as indicated in Table 3. The model also attained a perfect precision score of 1.00. Its F1-score, recorded at 0.9957, demonstrates a superior balance between precision and recall compared to Logistic Regression. Although the ROC AUC metric was not computed due to the non-probabilistic nature of SVM, the confusion matrix nonetheless confirms the model's excellent classification capabilities.

Table 3

SVM Results

Metric	Linear SVM
Accuracy	0.9981
Precision	1.0000
Recall	0.9914
F1 Score	0.9957
ROC AUC	N/A

The linear SVM demonstrated superior performance compared to Logistic Regression. It correctly classified all 1,600 non-depressed instances and accurately identified 459 out of 463 actual depressed tweets, resulting in only four false negatives. This yields a recall of 99.14%, which represents a notable

improvement over Logistic Regression, especially in reducing missed detections. With no false positives and perfect precision, the SVM model exhibits exceptional reliability and strong generalisation capability. These outcomes are consistent with earlier findings by Nguyen et al. (2014) and Muzafar et al. (2023), who reported the effectiveness of SVMs in handling sparse and high-dimensional text data, particularly in mental health prediction tasks. While both models produced commendable results, SVM showed slightly superior generalisation on the test set. However, a shared limitation of both models lies in their inability to capture contextual relationships within sentences, reducing their effectiveness in interpreting ambiguous or complex emotional expressions.

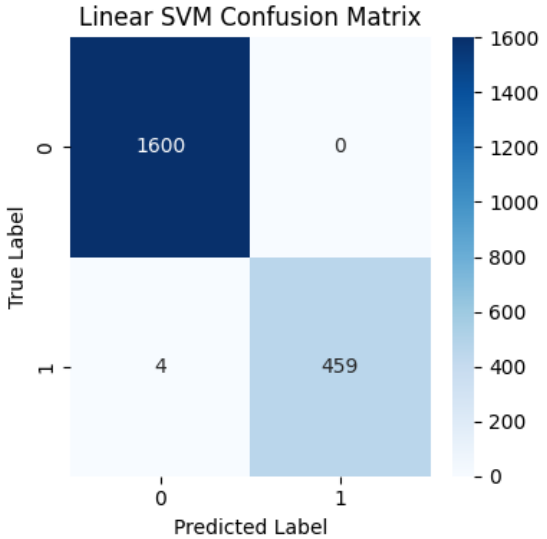

Figure 2: Linear SVM Confusion Matrix

DistilBERT Model for Depression Classification

To evaluate the efficacy of deep learning and contextual embeddings, DistilBERT was employed for binary classification. Following established practices in transformer-based NLP (Devlin et al., 2019), the model was trained on 80% of the pre-processed dataset, with the remaining 20% reserved for testing. The training process encompassed two epochs with a batch size of 8 for training and 16 for evaluation. The total training duration was approximately 13,334 seconds, underscoring the considerable computational demands associated with transformer architectures, even in their distilled variants.

Nonetheless, the model achieved rapid convergence and demonstrated strong predictive accuracy and reliability. In comparison to the classical models, DistilBERT demonstrated superior performance, particularly through marginal improvements in the F1-score and the balance between precision and recall. The model achieved an evaluation loss of 0.0115 and an AUC of 0.9975, indicating a high level of classification accuracy and confidence in distinguishing between the two classes. Moreover, its capacity to capture contextual nuances, including sarcasm and idiomatic expressions, highlights a significant advantage over traditional vectorisation approaches, which often fail to recognise such subtleties.

Table 4

DistilBERT Evaluation Results

Metric	DistilBERT
Eval Loss	0.0115
Eval Accuracy	0.9981
Eval F1 Score	0.9981
Eval Runtime (s)	13334.40
AUC Score	0.9975

The DistilBERT model outperformed both classical models, accurately classifying 1,604 non-depressed and 455 depressed tweets, with only four false negatives and no false positives. This confusion matrix corresponds to a recall of 98.91% and an F1-score of 0.9981, consistent with the earlier reported metrics. The marginal improvement over SVM, particularly in terms of predictive confidence and generalisability, may be attributed to the contextual understanding provided by transformer-based architectures. Such models are well-suited for identifying deeper semantic and emotional cues within textual data, as emphasised by Radford et al. (2019), Triantafyllopoulos et al. (2023), and Tavchioski et al. (2023), making them particularly effective in recognising subtle indicators of depression in adolescents' social media content. Overall, the findings affirm that all three models exhibited robust performance on real-world data. However, the transformer-based model demonstrated the most comprehensive and adaptable results, particularly in processing emotionally nuanced and complex expressions frequently encountered in adolescent discourse. This underscores its suitability for integration into early detection systems, where interpretability and sensitivity to linguistic subtleties are paramount.

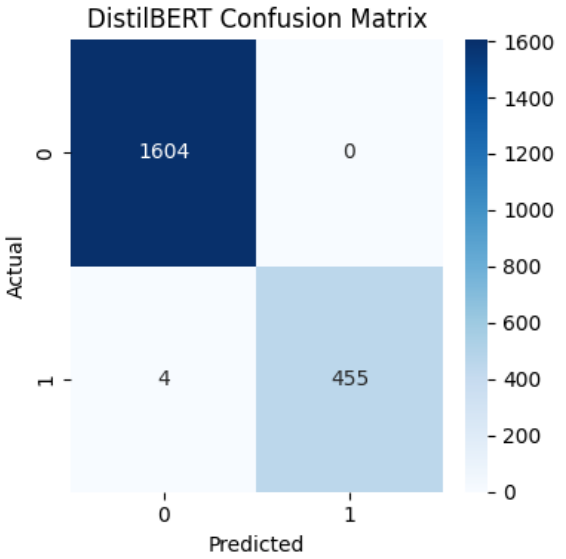

Figure 3: DistilBERT Confusion Matrix

Discussion

This section presents a discussion of the findings in relation to the three research questions. Drawing upon real-world social media data, both traditional machine learning and deep learning approaches were employed to evaluate the predictive capacity of adolescent linguistic behaviour in identifying depressive tendencies. The results are contextualised within the framework of existing literature and theoretical perspectives, thereby elucidating the broader implications of these findings for digital mental health screening initiatives.

Interpreting Adolescent Speech Patterns and Psychological Tendencies

The study aimed to investigate the relationship between adolescent speech patterns on social media and psychological indicators of depression. Analysis of a dataset comprising over 10,000 tweets revealed distinct linguistic cues associated with both depressive and non-depressive states. Terms reflecting hopelessness, loneliness, and negative emotions were particularly prevalent in the posts categorised as depressive, aligning with prior research on the online expression of psychological distress through emotional language (Chancellor & De Choudhury, 2020; Guntuku et al., 2019). The strong performance of the classification models, particularly their consistently high recall scores,

substantiates their ability to accurately detect depressive instances, thereby reinforcing this association. These findings lend further support to the concept of digital phenotyping, which posits that digital footprints such as language use may serve as proxies for psychological conditions (Torous et al., 2016). The recurring patterns observed in adolescent tweets affirm the hypothesis that online speech behaviours may reflect latent mental health challenges among youth, suggesting the feasibility of employing language-based systems for early detection and monitoring.

Implementing a Joint Detection Model with Real-World Labelled Data

The subsequent focus of the study involved establishing how a combined detection model could be formulated using authentic linguistic patterns and annotated indicators of depression. By employing actual user-generated content, the research addressed concerns of ethical and ecological validity typically associated with artificially constructed corpora (Chancellor et al., 2016; Ndikumana et al., 2025). The utilisation of naturally occurring language enabled a more accurate simulation of real-world conditions, allowing the models to learn from the inherent complexities of adolescent speech, including informal expressions and context-dependent nuances. Among the approaches applied, DistilBERT demonstrated particular effectiveness, achieving an F1-score of 0.9981 and an AUC of 0.9975—marginally surpassing the performance of Logistic Regression and SVM, both of which also exhibited strong results. These outcomes are consistent with emerging literature on the application of transformer-based models in mental health classification tasks (Devlin et al., 2019; Tavchioski et al., 2023; Triantafyllopoulos et al., 2023). The evidence reinforces the notion that integrated detection systems, when supported by advanced NLP methods and well-annotated datasets, can produce scalable, context-aware insights for mental health assessment.

Assessing Accuracy and Robustness in Depression Classification

The objective of the study extended to evaluating the reliability and predictive accuracy of all implemented models. Logistic Regression, SVM, and DistilBERT each demonstrated exceptional classification performance, achieving accuracy rates exceeding 99%. Notably, DistilBERT exhibited strong robustness when processing the intricacies of naturally occurring language. Its ability to accommodate diverse syntactic and semantic variations found in real-world text highlights its relevance for mental health applications, as also suggested by Murarka et al. (2020) and Mardini et al. (2025). A comprehensive

set of evaluation metrics—including F1-score, AUC, precision, and recall—was employed to measure model performance, particularly important in cases involving class imbalance, as endorsed by Tadesse et al. (2019). These results suggest that, when paired with rigorous preprocessing and systematic evaluation strategies, AI-based models hold promise for delivering clinically relevant outcomes in the context of early-stage mental health detection. In summary, the discussion validates the primary hypothesis of this research, affirming that adolescent language on social media platforms can be effectively leveraged to detect depressive tendencies within ethical, scalable, and technically sound machine learning frameworks.

Conclusion

This research has demonstrated the viability of employing social media data in real-world settings to detect depressive tendencies among adolescents through a combined detection model that integrates linguistic behaviour with supervised learning techniques. Departing from simulation-based methodologies, the study utilised a dataset comprising 10,310 annotated tweets related to depression and evaluated the comparative performance of three machine learning algorithms: Logistic Regression, SVM, and DistilBERT. Among these, DistilBERT delivered the most robust overall performance, although all three models achieved notably high levels of accuracy and F1-scores. The analysis confirmed that adolescent-generated content on social media exhibits discernible linguistic markers directly associated with depressive states. These findings align with the broader scholarly consensus that a significant link exists between online linguistic expression and mental health status. By adopting a quantitative lens to analyse speech patterns, the study substantiated the concept of digital phenotyping, which posits that textual data sourced from the internet can serve as a proxy for psychological assessment.

The integration of both traditional and transformer-based models established a robust benchmark for evaluating classification performance, highlighting trade-offs between interpretability and contextual sensitivity. Logistic Regression and SVM served as effective baselines, while DistilBERT excelled due to its ability to detect subtle semantic and emotional cues. The study addressed prior concerns about ecological validity and ethical risk by using anonymised, ethically sourced social media data, thereby enhancing generalisability and demonstrating the viability of privacy-conscious systems

for early intervention in educational, clinical, and digital wellbeing settings. All research objectives were met: linguistic patterns and model outputs confirmed a strong link between adolescent language and depressive tendencies; the joint detection model proved both feasible and effective; and evaluation metrics such as F1-score and ROC-AUC validated model reliability in handling subtle, real-world digital expressions. As in previous research, these metrics are essential for assessing AI's clinical potential. Overall, the findings support the ethical and accurate use of machine learning and social media data for adolescent mental health assessment. However, further research is needed to improve scalability, adapt to linguistic and contextual variations, and ensure cross-population transferability. Still, this study represents a significant step toward responsible, AI-supported mental health monitoring.

Recommendations

Future investigations should aim to enhance the current model by integrating multilingual datasets and accounting for linguistic diversity, including the use of slang, sarcasm, and regionally specific dialects prevalent among adolescent populations. Moreover, it is important to evaluate these models within live, real-time environments and across various social media platforms to determine their flexibility and reliability when subjected to changing user behaviours. To ensure clinical relevance and ethical implementation, collaboration with mental health practitioners will be critical. Such partnerships can help align these systems with established therapeutic practices, reinforcing the role of technology as a support mechanism for human-led interventions rather than a replacement. Additionally, the development of ensemble approaches that combine linguistic cues with metadata-based features offers a promising direction for improving the predictive accuracy and robustness of depression detection systems.

Practical Implications

The findings of this study hold direct implications for the development of systems aimed at monitoring adolescent mental health. By passively analysing linguistic indicators of psychological distress, such systems can support timely interventions. Integration of these technologies within educational settings, mental health support organisations, and digital health platforms presents a practical opportunity for broader application. Given the high levels of sensitivity and accuracy demonstrated, particularly by the DistilBERT model,

these tools could be embedded within non-intrusive, assistive frameworks to aid counsellors and educators in identifying at-risk individuals. Furthermore, the deployment of such models on digital platforms may enable the automatic recommendation of mental health resources to users exhibiting signs of emotional difficulty, thereby enhancing user well-being and fostering a more socially responsible digital environment.

Limitations

Despite its contributions, the present study is not without limitations. Firstly, the dataset used was confined to a single language and social media platform, which may restrict the broader applicability of the findings across different linguistic and demographic contexts. Additionally, the models were evaluated within a controlled, offline setting, and have not yet been exposed to real-time, unstructured environments where user behaviour and data quality can vary significantly. Another constraint lies in the use of DistilBERT, which, while effective, may have diminished nuanced emotional content without offering the level of interpretability required in sensitive mental health assessments. Furthermore, the study did not consider the temporal dimension of depression, such as variations in symptoms over time, which is critical for accurate clinical diagnosis and effective intervention planning.

References

Baydili, İ., Tasci, B., & Tasci, G. (2025). Deep learning-based detection of depression and suicidal tendencies in social media data with feature selection. *Behavioral Sciences*, *15*(3), 352. https://doi.org/10.3390/bs15030352

Benke, I., Feine, J., Venable, J. R., & Maedche, A. (2020). On implementing ethical principles in design science research. *AIS Transactions on Human-Computer Interaction*, *12*(4), 206-227. https://doi.org/10.17705/1thci.00136

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, *30*(7), 1145-1159. https://doi.org/10.1016/S0031-3203(96)00142-2

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901. http://dx.doi.org/10.48550/arXiv.2005.14165

Calvo, R. A., Milne, D. N., Hussain, M. S., & Christensen, H. (2017). Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, *23*(5), 649-685. https://doi.org/10.1017/S1351324916000383

Chancellor, S., & De Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*, *3*(1), 43. https://doi.org/10.1038/s41746-020-0233-7

Chancellor, S., Lin, Z., Goodman, E. L., Zerwas, S., & De Choudhury, M. (2016). Quantifying and

predicting mental illness severity in online pro-eating disorder communities. Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing, http://dx.doi.org/10.1145/2818048.2819973

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, volume 1 (long and short papers), https://doi.org/10.18653/v1/N19-1423

Guntuku, S. C., Ramsay, J. R., Merchant, R. M., & Ungar, L. H. (2019). Language of ADHD in adults on social media. *Journal of attention disorders*, *23*(12), 1475-1485. https://doi.org/10.1177/1087054717738083

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons. https://doi.org/10.1002/9781118548387

Lee, E. E., Torous, J., De Choudhury, M., Depp, C. A., Graham, S. A., Kim, H.-C., Paulus, M. P., Krystal, J. H., & Jeste, D. V. (2021). Artificial intelligence for mental health care: clinical applications, barriers, facilitators, and artificial wisdom. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *6*(9), 856-864. https://doi.org/10.1016/j.bpsc.2021.02.001

Liu, F., Cheng, Z., Chen, H., Wei, Y., Nie, L., & Kankanhalli, M. (2022). Privacy-preserving synthetic data generation for recommendation systems. Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, https://doi.org/10.1145/3477495.3532044

Mardini, M. T., Khalil, G. E., Bai, C., DivaKaran, A. M., & Ray, J. M. (2025). Identifying Adolescent Depression and Anxiety Through Real-World Data and Social Determinants of Health: Machine Learning Model Development and Validation. *JMIR Mental Health*, *12*, e66665. https://doi.org/10.2196/66665

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*. https://www.researchgate.net/publication/257882504

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, *29*(3), 436-465. https://doi.org/10.1111/j.1467-8640.2012.00460.x

Moreno, M. A., Goniu, N., Moreno, P. S., & Diekema, D. (2013). Ethics of social media research: Common concerns and practical considerations. *Cyberpsychology, behavior, and social networking*, *16*(9), 708-713. https://doi.org/10.1089/cyber.2012.0334

Murarka, A., Radhakrishnan, B., & Ravichandran, S. (2020). Detection and classification of mental illnesses on social media using roberta. *arXiv preprint arXiv:2011.11226*. https://doi.org/10.48550/arXiv.2011.11226

Muzafar, D., Khan, F. Y., & Qayoom, M. (2023). Machine learning algorithms for depression detection and their comparison. *arXiv preprint arXiv:2301.03222*. https://doi.org/10.48550/arXiv.2301.03222

Ndikumana, F., Izabayo, J., Kalisa, J., Nemerimana, M., Nyabyenda, E. C., Muzungu, S. H., Komezusenge, I., Uwase, M., Ndagijimana, S., & Twizere, C. (2025). Machine learning-based predictive modelling of mental health in Rwandan Youth. *Scientific Reports*, *15*(1), 16032. https://doi.org/10.1038/s41598-025-00519-z

Nguyen, T., Phung, D., Dao, B., Venkatesh, S., & Berk, M. (2014). Affective and content analysis of online depression communities. *IEEE transactions on affective computing*, *5*(3), 217-226. https://doi.org/10.1109/TAFFC.2014.2315623

Organization, W. H. (2024). Mental health of adolescents. https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9. https://api.semanticscholar.org/CorpusID:160025533

Schreiber, R. E., & Veilleux, J. C. (2022). Perceived invalidation of emotion uniquely predicts affective distress: Implications for the role of interpersonal factors in emotional experience. *Personality and Individual Differences*, *184*, 111191. https://doi.org/10.1016/j.paid.2021.111191

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information processing & management*, *45*(4), 427-437. https://doi.org/10.1016/j.ipm.2009.03.002

Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2019). Detection of depression-related posts in reddit social media forum. *Ieee Access*, *7*, 44883-44893. https://doi.org/10.1109/ACCESS.2019.2909180

Tavchioski, I., Robnik-Šikonja, M., & Pollak, S. (2023). Detection of depression on social networks using transformers and ensembles. *arXiv preprint arXiv:2305.05325*. https://doi.org/10.48550/arXiv.2305.05325

Thapar, A., Collishaw, S., Pine, D. S., & Thapar, A. K. (2012). Depression in adolescence. *The lancet*, *379*(9820), 1056-1067. https://doi.org/10.1016/S0140-6736(11)60871-4

Torous, J., Kiang, M. V., Lorme, J., & Onnela, J.-P. (2016). New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. *JMIR Mental Health*, *3*(2), e5165. https://doi.org/10.2196/mental.5165

Triantafyllopoulos, I., Paraskevopoulos, G., & Potamianos, A. (2023). Depression detection in social media posts using affective and social norm features. *arXiv preprint arXiv:2303.14279*. https://doi.org/10.48550/arXiv.2303.14279

Zhang, X., Zhang, X., Chen, W., Li, C., & Yu, C. (2024). Improving speech depression detection using transfer learning with wav2vec 2.0 in low-resource environments. *Scientific Reports*, *14*(1), 9543. https://doi.org/10.1038/s41598-024-60278-1

Zirikly, A., Resnik, P., Uzuner, O., & Hollingshead, K. (2019). CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. Proceedings of the sixth workshop on computational linguistics and clinical psychology, https://doi.org/10.18653/v1/W19-3003