Discover Mental Health

RESEARCH Open Access

Enhancing online well-being through transformer-based analysis of misinformation and mental health

Sarvesh Arora[1], Deepika Kumar[1], Sarthak Arora[1], Vallari Agrawal[1], Ritanjali Panigrahi[2*] and Md Abdullah Al Kafi[3]

*Correspondence:
Ritanjali Panigrahi
ritanjali.panigrahi@gmail.com
[1]Department of Computer Science and Engineering, Bharati Vidyapeeth's College of Engineering, New Delhi 110063, India
[2]IS & Analytics, Jindal Global Business School, O.P. Jindal Global University, Sonipat, Haryana 131001, India
[3]Department of Computer Science and Engineering, Daffodil International University, Birulia, Savar, Dhaka 1341, Bangladesh

Abstract

With the growing digitalization, social media has truly transformed the ways in which people connect, but not without consequences. Alongside enhanced connectivity, these platforms have become fertile ground for the rapid spread of misinformation, with troubling implications for mental health. While much of the existing research has focused on mitigating harmful content, emerging efforts are also recognizing the role of positive discourse, such as hope speech, in promoting digital well-being. This study introduces a hybrid transformer-based model, combining RoBERTa and LSTM architectures, to tackle three pressing challenges: identifying misinformation, gauging its psychological impact, and classifying related mental health disorders, particularly across diverse and low-resource language contexts. The model achieved impressive accuracy rates: 98.4% for misinformation detection, 87.8% for mental health assessment, and 77.3% for disorder classification. A significant link has been shown to substantiate the presence of a statistically significant association between exposure to deceptive information and mental health outcomes using statistical analysis (Pearson's Chi-Squared Test (with a p-value = 0.003871. These findings highlight the urgent need for effective strategies to both curb misinformation and encourage positive emotional engagement online. Future investigations are encouraged to expand this work by incorporating hope speech detection frameworks and addressing broader linguistic, demographic, and cultural perspectives.

Keywords Mental health, Disorder analysis, Deep learning, Fake news, Social media

1 Introduction

Human beings are inherently social, thriving on connection and companionship, which profoundly influence their health and well-being [1, 2]. In today's society, where social media has established itself as the foremost means of communication around the world, social interaction has never been more accessible. It not only acts as a platform for personal communication but is also a vast repository of information to enhance professional relations, all in the palm of one's hands [3]. However, as a major channel for the dissemination of news and information, social media has increasingly become a target for abuse and manipulation [4, 5]. The accessibility and immediacy of these platforms

have facilitated the rapid generation and spread of inaccurate information, commonly referred to as fake news—often driven by political and financial motives [6]. Due to the inherently open nature of social media, there is a lack of accountability for fact-checking, which allows anyone, including news organizations, to share content without being held responsible for its accuracy. Trying to verify the truth of a rumor or a terrifying story gives rise to feelings of rage, distrust, anxiety, and even depression among the public [7, 8]. Additionally, despite its intended purpose of fostering connection, excessive use of social media can lead to increased feelings of loneliness and isolation, exacerbating mental health issues like anxiety and depression [9].

As long-term, incapacitating consequences of depression and other common mental illnesses, neuropsychiatric disorders have contributed approximately 15% of the overall disease burden worldwide [10]. Literacy of mental health is relatively low and this supports the idea of raising its awareness. Figure 1 is the percentage of the world population who have a specific mental health disorder. As concerns around mental health grow globally, anxiety disorders stand out as one of the most prevalent and pressing challenges. Affecting approximately 275 million people, or 4% of the world's population, these disorders reflect a silent epidemic that transcends borders. Prevalence rates differ from country to country, ranging between 2.5% and 6.5%, influenced by factors such as socioeconomic conditions, healthcare access, and cultural attitudes toward mental health. Women bear a disproportionate burden, comprising around 62% of those affected, roughly 170 million individuals, compared to 105 million men. These figures underscore the urgent need for gender-sensitive mental health strategies and more nuanced global responses to anxiety-related conditions [11].

Amid growing global uncertainty, the urgent need to understand and manage the psychological effects of digital information environments has intensified [12]. While significant efforts have focused on detecting and mitigating harmful content such as misinformation, the next frontier lies in actively promoting positive emotional discourse, particularly through the identification and amplification of hope speech. Developing robust mechanisms to first identify risks, such as the mental health deterioration triggered by fake news, is a necessary precursor to building effective hope speech models.

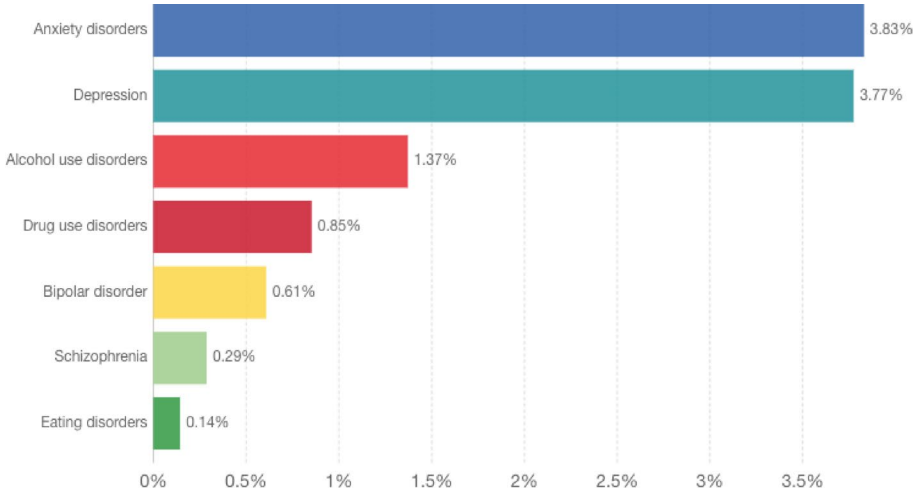


Fig. 1 Share of the global population with reported mental health disorders [11]

Various supervised classification algorithms and deep learning techniques has been implemented to detect fake news. Researchers have extensively worked to detect fake news using Recurrent Neural Network (RNN) and Long-Short Term Memory (LSTM) algorithms [13]. Models based on Support Vector Machines (SVM), Naive Bayes, Bidirectional Encoder Representations from Transformers (BERT) etc. have also been developed to classify posts as factually correct or incorrect [14]. To detect mental health disorders, Convolutional Neural Network (CNN), Feed Forward (FF) and SVM have been the leading tools for research [15]. Accordingly, this study introduces a three-step integrated framework that first addresses the challenges of misinformation and mental health assessment, laying the groundwork for future research in fostering positivity and resilience within online communities.

The key contributions of this research are as follows:

1. Development of a hybrid transformer-based classifier to independently detect fake news, identify mental health implications, and analyze specific disorders.
2. Proposal of an integrated architecture that unifies the three models to analyze raw Reddit data for assessing the mental health effects of misleading content.
3. Evaluation of model performance against existing algorithms using standard metrics including precision, recall, F1-score, and accuracy.
4. Validation of interdependencies between the core features through statistical hypothesis testing.

The paper has the following structure: Sect. 2 includes an in-depth review of literature and current progress in the field of fake news and mental health detection. Section 3 includes the materials and methods used in the study. Section 4 elaborates the design of the proposed integrated model. Section 5 presents the results of the experiment and the analysis. Section 6 statistically analyzes the meaning of the findings. Implications to academia and industry are discussed in Sect. 7. Section 8 covers the shortcomings, while Sect. 9 summarizes the main findings and provides some conclusions and future research. Section 10 covers compliance with ethical standards.

2 Literature survey

Past research has recognized patterns and characteristics of fake news, which can be better detected and prevented. Machine learning can be used to filter real and fake news, and in particular light of polarization and information saturation, it can be useful to differentiate between them with the help of algorithms and data analysis. To tackle the growing challenge of fake news, researchers have turned to a combination of linguistic analysis and machine learning techniques. Using the Bag-of-Words method alongside the k-nearest neighbor (KNN) algorithm, one study applied text normalization and feature extraction to a dataset from the 2016 U.S. presidential election, achieving a striking 92% accuracy in identifying false social media content [16–18]. In a related effort, a model known as Social Article Fusion (SAF) was developed to detect fake news by analyzing both the linguistic features of the content and patterns of user engagement. Leveraging Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) architectures, the SAF model reached an accuracy rate of 74% [19, 20].

Studies have been done also using RNN models (named vanilla and GRU), LSTMs and Convolutional Neural Network (CNN) and a combination of CNN and GRU models.

GRU model based on RNN was the most accurate at 89 percent. It was however noted that the CNN model was 5 times faster than the RNN model [21–23]. Natural language processing methods, together with the Random forest (RF) algorithm, have been applied to social media posts in Morocco in order to identify fake news. The RF algorithm is superior to all the other models in all four measures. RF model is accurate by 79% (72 percent by SVM). The maximum memory rate of RF was 100% compared to 94% of Decision Tree (DT). The F1-score of RF = 83 per cent, and its precision is 85%. 37% of the total 2000 posts in the test set were false under this method [24–26]. Naive Bayes (NB), Maximum Entropy (ME) and SVM models achieved a precision of about 80 percent when using n-gram and bigram models, whereas ensemble and hybrid method-bases algorithm achieved a precision of 85 percent [27]. Using simple natural text processing including the methods of TF-IDF, bag of words, tokenization, and delete stop words, then classification with SVM and RF were applied, and it was concluded that the SVM method is the most reasonable one with the accuracy of 94% [28]. A Logistic Regression (LR) model was used to build a diffusion network using Twitter data by applying a classification framework and including a multi-layer structure of Twitter diffusion networks. The network was multi-layered, and several structural properties (network density, the number of strong/weak connected components and diameter) were calculated individually on each layer. In this manner, various tweets, retweets, mentions, replies, and quotes could be separated and analyzed. Two datasets of tweets posted in 2019 on Twitter, United States and Italy, were analyzed using this model with a maximum of 94% accuracy [29, 30]. Moreover, by training a proprietary semi-supervised model named ENDEMIC, using exogenous and endogenous post-related signals based on BERT and Bi-directional LSTM (Bi-LSTM) algorithms, researchers achieved 94% fraudulent news detection accuracy [31, 32].

Seddari et al. developed a method that merged linguistic and knowledge-based features, achieving a notable accuracy of 94.4%, which outperformed the individual accuracies of 89.4% and 81.2%. This result was obtained through an AutoAI experiment in IBM Watson Studio, where Random Forest (RF) was ultimately selected from four algorithms, including Logistic Regression (LR), Additional Trees Discriminant (ATD), and eXtreme Gradient Boosting (XGBoost) [33]. Sheng et al. leveraged pre-trained models like BERT within the News Environment Perception Framework (NEP) to detect macro and micro environmental signals linked to post popularity and novelty, achieving peak performance of 83.1% on the Chinese dataset and 72.8% on the English dataset [34].

Raza et al. addressed early fake news detection and label scarcity by combining news content and social contexts with a weak supervision labeling strategy, utilizing BART (Bidirectional and Autoregressive Transformers). This method significantly improved accuracy to 74.8% compared to traditional techniques [35]. Alyoubi et al. surpassed 95% accuracy using pre-trained context-based models like MARBERT and ARBERT combined with CNN and BiLSTM for Arabic Tweets, outperforming other embedding methods such as Keras, Word2Vec, and FastText. Advanced deep learning and transformer models notably enhance fake tweet detection accuracy over traditional methods [36]. Apuke and Omar compared the forecasts of sharing of fake news about using Structural equation modelling (SEM) in which the findings showed that both Skewness and kurtosis were significant. [37]. Baek et al. quested the reasons why people share links on Facebook through Logistic Regression and found that sharing links enlarges SNS

roles [38]. It is important to identify and treat mental health issues early. Mentally ill persons can find considerable relief by early detection, accurate diagnosis, and proper treatment [39]. Mental disorders, including depression, anxiety, bipolar, borderline personality disorder (BPD), schizophrenia and autism have been identified by creating six independent binary classification models on social media data by using XGBoost and CNN which achieved maximum accuracies of 94.91% and 96.96 respectively on autism [40]. A two-CNN multimodal model was trained to model fMRI and sMRI modalities to predict attention-deficit/hyperactivity disorder (ADHD) with 66.04 and 65.86 model accuracy respectively [41]. Machine learning models have been created by scientists to identify and classify social media-based mental health content in 11 disorder themes. The models were trained on a mix of Convolutional Neural Networks (CNN), Feed Forward networks (FF), linear classifiers and Support Vector Machines (SVM) and reached high accuracy rates of 91.08, 90.79, 85.84, and 86, respectively [15]. On this basis, a new hybrid methodology was proposed, combining a factor graph model with CNN to examine not only the content of tweets but also the social interactions of users. This technique was found to be useful in detecting stress, with an accuracy rate of 91.55 percent [42]. A Multimodal Deep Denoising Autoencoder (MultiDDAE) and a Paragraph Vector (PV) are combined and injected into Multitask Deep Neural Network (DNN) to identify bipolar disorder and depression with 71.7 and 83.9 accuracies respectively [43]. A combination of different classification models has also been created to identify mental illnesses that has reached an accuracy of 89% [44]. Suicidality is another significant feature of psychologically affected users that has been identified in Twitter users using SVM to identify 80 percent of strongly concerning tweets correctly [45]. Unsupervised algorithms, such as clustering to detect at-risk behaviours, have been used to classify users, based on the extent to which their behaviours changed, with an accuracy of 76.12% [46]. Clinical psychologists have been stupended to fetch data to identify a blitz of mental disorders among children utilizing eight learning methods—Averaged One-Dependence Estimator (AODE), Multilayer Perceptron (MP), Radial Basis Function (RBF) Network, IB1, KStar, Multiclass classifier (MC), Functional Trees (FT) and Logical Analysis of Data (LAD) tree, with MP, MC and LAD performing notably superior to all the others with an approximate 85% accuracy [47]. It has also presented a Mental Health Diagnostic Expert System that assists psychologists by using Rule-Based Reasoning, Fuzzy Logic and Fuzzy-Genetic Algorithms to diagnose patients and prescribe appropriate treatment policies [48].

Zarate et al. employed natural language processing techniques to analyze 233,000 tweets from 605 users, achieving high prediction accuracy (Naïve Bayes 81.1%, Random Forests 79.8%, LASSO-regression 79.4%) in identifying self-reported anxiety diagnoses on Twitter, and identified four distinct user profiles characterized by sentiment and behavioral patterns [49]. Rawat et al. explored deep learning approaches for predicting mental states from Twitter data, utilizing models including Simple Recurrent Neural Networks, Long Short-Term Memory (LSTM), Bidirectional LSTM, and BERT with a Multilayer Perceptron (MLP). The study found that LSTM achieved 99.27% accuracy on the full dataset, while BERT combined with MLP excelled with 98.49% accuracy on a balanced dataset [50]. Al Banna et al. (2023) developed a hybrid deep learning model combining LSTM and Convolutional Neural Networks, achieving 99.42% accuracy in identifying depressive tweets from 571,000 Twitter posts, revealing a notable

rise in depressive content during the COVID-19 pandemic [51]. Kokane et al. (2024) proposed a methodology using NLP Transformers to detect mental illness, specifically depression, by analyzing social media posts from Twitter and Reddit. Four NLP Transformers—BERT, XLNet, RoBERTa, and DistilBERT—were applied to two datasets, with DistilBERT achieving the highest accuracy, 91% on the Twitter dataset and 84% on the Reddit dataset, demonstrating its effectiveness in identifying symptoms of depression [52]. Gupta et al. (2023) conducted a comparative analysis of six NLP models—BERT, RoBERTa, DistilBERT, ALBERT, Electra, and XLNet—for detecting depression on Twitter. The study found that XLNet, DistilBERT, and RoBERTa achieved the highest performance, with accuracies exceeding 99% [53]. Deshpande et al. (2024) introduced the LocalTweets dataset and presented a novel framework for population-level mental health (MH) surveillance. The LocalHealth approach, involving sampling, encoding, aggregation, and prediction, utilizes the LocalTweets data to predict MH outcomes for different neighborhoods. Among the evaluated models—RoBERTa-base, RoBERTa-large, Twitter-RoBERTa-base, PHS-BERT, and GPT-3.5—GPT-3.5 achieved the highest performance with an F1-score of 0.7429 and 79.78% accuracy, while RoBERTa-base also demonstrated strong performance in data-limited settings [54]. While extensive research has been conducted separately in the areas of fake news detection and mental health disorder analysis, few studies have attempted to bridge the two; leaving a critical gap in understanding how misinformation and mental well-being intersect. Moreover, there is no literature that has drawn a correlation between fake news and mental health and how the former may influence the latter. The purpose of this work is to fill these gaps and conclude that mental health disorders are a direct consequence of fake news.

3 Materials & methods

This section provides information regarding the datasets that are used to develop and train the proposed hybrid transformer model. A short history of the pre-processing methods has also been described.

3.1 Dataset description

This research uses three sets of pre-labelled datasets for the purpose of training and testing fake news detection, severity of the mental health problem and the analysis of disorders individually [55–57]. The datasets are as follows:

3.1.1 Fake news dataset

The dataset used to train the fake news detection model on social media includes '6,420 posts' categorized as 'real' or 'fake' [55]. It is sourced from the 'COVID-19 Fake News Detection in English' task released under the 'Constraint@AAAI-2021 shared evaluation campaign' and comprises COVID-19–related content collected from multiple social media platforms, including 'Twitter', 'Facebook', and 'Instagram'. Annotations are provided by the task organizers following standardized labelling guidelines. The dataset is limited to English-language posts and reflects the pandemic-specific information environment, which may introduce topical and temporal biases.

3.1.2 Mental health implication dataset

The dataset used to train the mental health implication model is divided into three labels: 0, where posts demonstrate no mental health problems; 1, where mental health is mentioned as an expression of irritation or without serious intent; and 2, where posts indicate self-expressed mental-health difficulties [56]. These labels reflect self-reported mental-health implications in social media text and do not correspond to clinically diagnosed conditions. The dataset consists of 2,460 English-language tweets collected via Twitter's official API and manually annotated by two authors with cross-verification to ensure labeling consistency. The dataset is publicly available online from Manipal Institute of Technology and class-balanced, but it is limited to a single platform and time period, which may affect generalizability.

3.1.3 Mental health disorder analysis dataset

The dataset used to train the mental health disorder analysis model consists of 58,116 filtered data points from original dataset and categorizes posts by mental-health disorder implication into anxiety, BPD, bi-polar, depression, schizophrenia, and other mental diseases [57]. The dataset comprises publicly available Reddit posts reflecting self-expressed mental-health discussions, rather than clinically validated diagnoses. Some posts contain limited contextual information, and the presence of low-quality or troll content may affect model learning. Reddit was selected as the data source due to its discussion-driven structure, which enables users to express detailed reactions and opinions on news-related content. The platform's topic-focused communities support rich textual interactions that are well suited for analyzing misinformation and mental health–related expressions. In addition, the availability of publicly accessible data through official APIs allows ethical and reproducible data collection. Additionally, reliance on a single platform and self-reported content may introduce biases and limit generalizability. Due to the large size of the original dataset, a subset was used to ensure manageable processing while maintaining representativeness, Stratified sampling was employed to preserve the original class distribution across all disorder categories, thereby minimizing sampling bias and ensuring reliable model training. The same has been depicted in Table 1:

In order to understand the linguistic pattern of each dataset better and derive meaningful information, the most frequently used words (without stopwords) related to each

Table 1 Distribution of each dataset

Dataset	Total count	Label count	Link
Fake news dataset	6420	Fake: 3060 Real: 3360	https://www.kaggle.com/datasets/elvinagammed/covid19-fake-news-dataset-nlp
Mental health implication dataset	2460	No mental disorder: 820 Expression of annoyance: 820 Mental disorder present: 820	Item—ADAM-SDMH: A DAtaset from Manipal for Severity Detection in Tweets related to Mental Health—figshare—Figshare
Disorder Analysis dataset (subset of original)	58,116	Anxiety: 16,153 BPD: 21,280 Bi-polar: 3541 Depression: 12,150 Schizophrenia: 1180 Other: 3812	Mental Disorders Identification (Reddit)

label in the datasets are shown in Fig. 1. Also, the characteristics of each dataset are presented in Table 1.

Figure 2 offers visual insights into the dominant vocabulary across the three core analytical tasks: fake news detection, mental health implications, and disorder analysis. In subfigure (a), the word cloud for fake news detection reveals commonly used terms that signal misinformation, often characterized by sensational or emotionally charged language. Subfigure (b) highlights vocabulary associated with mental health expression, capturing words that reflect stress, anxiety, and emotional distress. Meanwhile, subfigure (c) presents terms linked to specific mental health disorders, showcasing patterns that the model learns to recognize during classification. These word clouds not only illustrate key linguistic features but also emphasize the distinct semantic landscapes each model must navigate Table 2.

3.2 DATA pre-processing

The information in both datasets lacked coherence and contained redundant components like hashtags, URLs, emoticons, and other irrelevant information that did not add any value to the meaning of the text. Consequently, the pre-processing of the data was performed in advance, and further cleaning was based on the text form. This included elimination of stop words, URLs, emojis, special characters, hashtags, web links and other extraneous items. The text was then cleaned and tokenized by dividing it into separate tokens.

4 Proposed methodology

4.1 Hybrid RoBERTa-LSTM classifier

In this work, we present a RoBERTa-LSTM classifier, which combines the contextual power of the already trained RoBERTa model [58] with the sequential learning ability of a Long Short-Term Memory (LSTM) neural network [59], as depicted in Fig. 3. This model is used to categorize text inputs into various categories.

The proposed model was trained using a fixed set of hyperparameters to ensure stable and reproducible performance. Training was conducted with a predefined learning rate and batch size selected based on preliminary experimentation to balance convergence speed and generalization. The model was optimized over a fixed number of epochs using an adaptive optimization algorithm to minimize the loss function effectively. To ensure consistency across multiple runs and eliminate variability caused by random initialization, random seeds were fixed during data splitting, weight initialization, and training procedures.

Each of these two model components serves a distinct role in the text classification process, and their combination allows for a comprehensive analysis of the input data.

1)RoBERTa: Robustly Optimized BERT Approach

RoBERTa (Robustly optimized BERT approach) is a transformer-based model that builds upon the foundations of the Bidirectional Encoder Representations from Transformers (BERT) model. It is designed to process and understand natural language by capturing the context in which words appear. Unlike traditional models that consider words in isolation, RoBERTa interprets each word in relation to the surrounding words, which allows it to understand the subtle nuances and complexities of language.

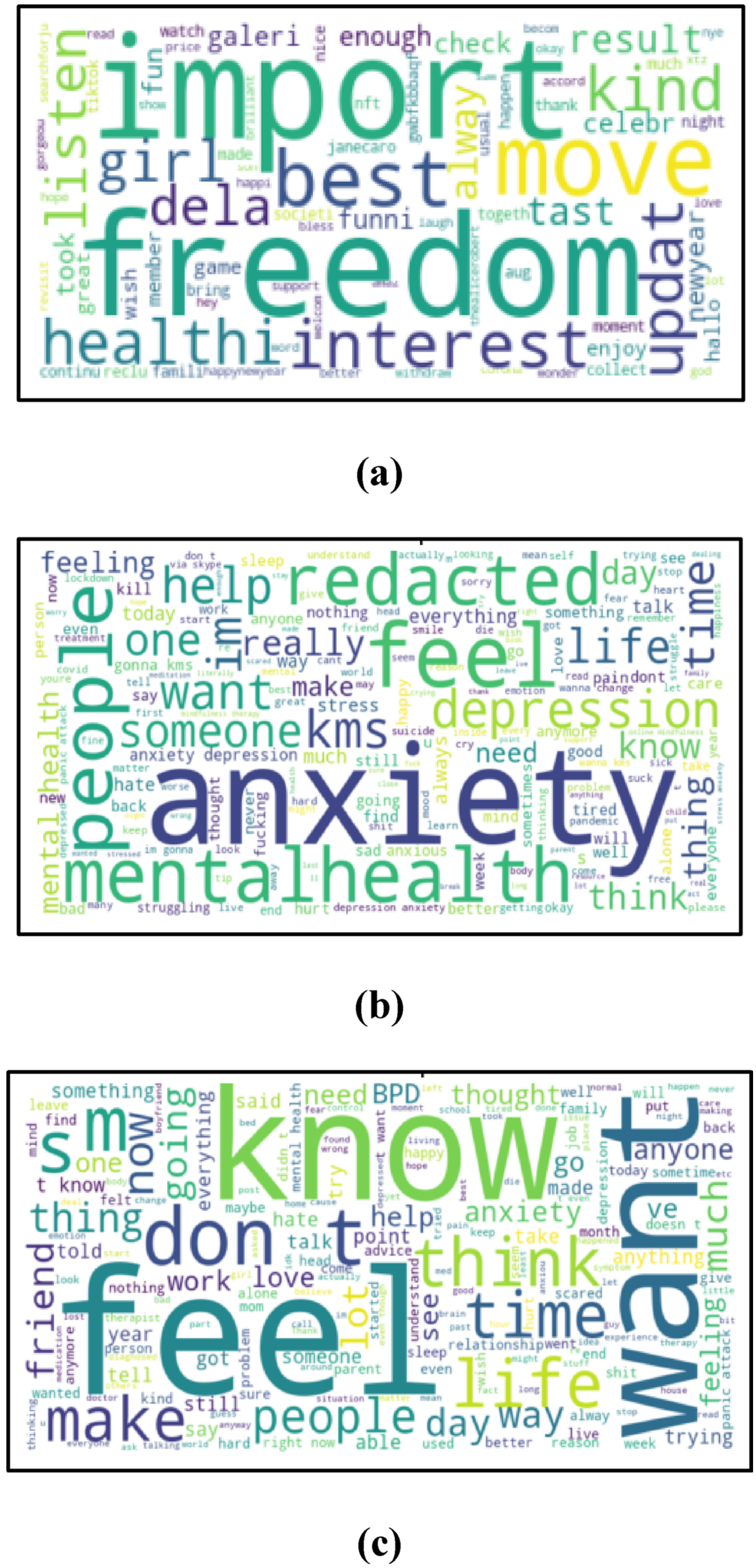


Fig. 2 Word clouds representing the vocabulary for **a** Fake News Detection, **b** Mental Health Implication and **c** Disorder Analysis

Table 2 Evaluation Metric Table

$Precision = \frac{TPos}{(TPos+FPos)}$	(1)
$Recall = \frac{TPos}{(TPos+FNeg)}$	(2)
$F-Measure = \frac{(2*Precision*Recall)}{(Precision+Recall)}$	(3)
$Accuracy = \frac{NumberofCorrectPredictions}{TotalNumberofPredictions}$	(4)

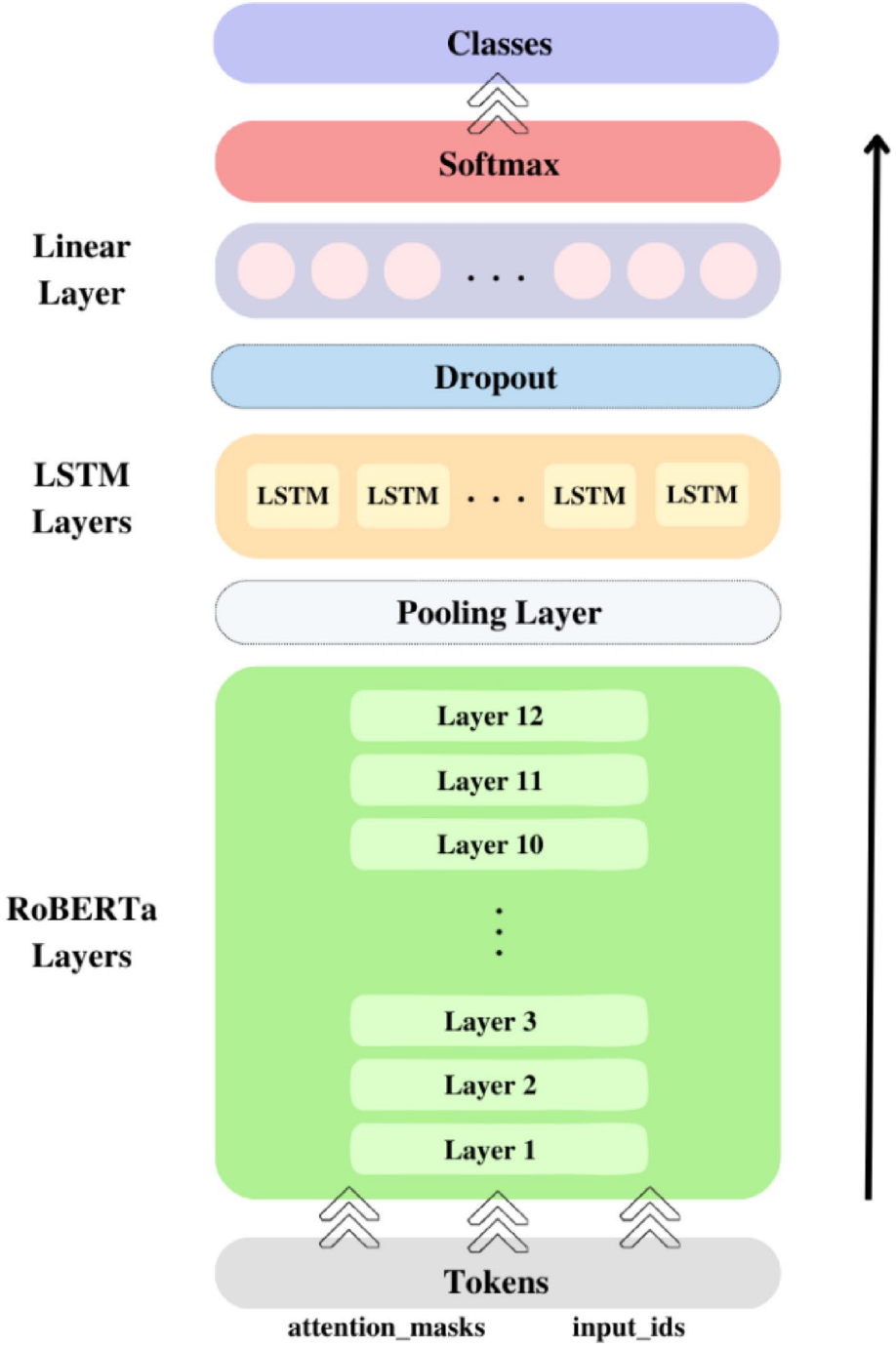


Fig. 3 RoBERTa-LSTM Architecture

RoBERTa is pre-trained on large text corpora using a technique called self-supervised learning, where the model learns to predict missing words in a sentence. This pre-training enables RoBERTa to develop a deep understanding of linguistic structures, semantics, and the relationships between words. When fine-tuned on specific tasks, such as text classification, RoBERTa can leverage this pre-existing knowledge to extract meaningful features from the input text.

During input processing, RoBERTa tokenizes the text, breaking it down into smaller units known as tokens. These tokens are then converted into numerical vectors called embeddings, which represent the meaning of each word in the context of the entire sentence. Additionally, RoBERTa generates an attention mask that helps the model focus on the relevant parts of the input, ensuring that padding tokens (used to standardize input lengths) do not interfere with the model's processing.

One of the key strengths of RoBERTa is its ability to generate contextual embeddings—dense vector representations that capture the meaning of words based on their context. This is particularly important in natural language processing (NLP) tasks, as the meaning of a word can vary significantly depending on the context in which it is used.

2)LSTM: Long Short-Term Memory

Long Short-Term Memory (LSTM) is a specialized type of Recurrent Neural Network (RNN) designed to handle sequential data. In many NLP tasks, the order of words or tokens in a sentence is crucial for understanding the meaning. LSTM is particularly effective in modeling these sequential dependencies, making it well-suited for tasks where the temporal order of information is important.

> *LSTM processes sequences of tokens one at a time, in contrast to transformer models like RoBERTa, which process entire sequences in parallel. This sequential processing allows LSTM to capture the relationships between tokens across the entire sequence, including long-range dependencies that may span multiple words or even sentences.*

The LSTM architecture is unique due to its memory mechanism, which consists of three types of gates: input, forget, and output gates. These gates control the flow of information through the network, determining which information should be retained, which should be forgotten, and which should be passed to the next time step. This mechanism allows LSTM to maintain a balance between remembering important information and discarding irrelevant data, enabling the model to capture long-term dependencies effectively.

Rationale for Model Choice:

Hybrid models combining transformer-based language models with recurrent networks have been shown to effectively capture both deep contextual semantics and sequential dependencies in text classification tasks. [60, 61]

RoBERTa was selected for its robust pretraining, providing high-quality contextual embeddings suitable for social media text. While models such as DeBERTa, or DistilRoBERTa offer efficiency, RoBERTa provides a favorable trade-off between accuracy and computational cost. The LSTM layer was chosen over RNNs or GRUs for its ability to capture long-term dependencies and complex sequential patterns via dedicated gating mechanisms [62]. This hybrid design enables the model to combine contextual understanding from RoBERTa with sequential relational patterns from LSTM, improving performance on tasks with nuanced and informal language.

RoBERTa-LSTM Classifier is created on the basis of PyTorch library and is made up of multiple layers. The pre-trained RoBERTa model is the first layer that is initialised with pre-trained weights and is fine-tuned in the process of training. The architecture of the proposed RoBERTa-LSTM classifier is designed to accept input sequences in two essential parts input ids and attention mask. They are inputted into the RoBERTa model which produces contextual embeddings of every token in the input. These token-level embeddings are then condensed into one vector in a subsequent pooling layer that represents the nature of the whole input sequence.

This pooled representation is passed into a stack of LSTM layers designed to model long-term dependencies within the text. With multiple hidden units and layered depth, the LSTM component captures complex relational patterns among input tokens. To mitigate overfitting, a dropout layer is incorporated, randomly deactivating neurons during training to encourage the model to learn generalizable features rather than memorizing the dataset.

For the classification task, the final LSTM output is passed through a fully connected linear layer that maps it to the desired number of output classes. A softmax activation function is then applied to convert raw logits into class probabilities, offering interpretable insights into the model's prediction confidence.

Overall, The RoBERTa-LSTM hybrid classifier effectively combines the contextual embeddings generated by RoBERTa with the sequential processing capabilities of LSTM. This integration allows the model to capture both contextual understanding and temporal relationships between tokens, enhancing its performance in text classification tasks. The model is trained through supervised learning, with the optimization of weights achieved by minimizing the loss function during the training process.

4.2 Integrated model architecture for the mental health implications of fake news posts

This paper will examine the psychological health consequences of fake news in a three-step learning model. It should include: (1) distinguishing between news posts to figure out whether they represent real or fake news, (2) deciding whether or not the posts concerned display worsening mental health, and (3) categorizing the disorders suggested by the posts. Although each of the mentioned models has significant research potential, when combined into a single system, they will allow us to take a deeper look at the negative effects of social media through a psychological lens.

No current dataset, however, brings together all three components. The datasets that are currently being used to train the individual models have limitations like lack of global representation and are subject to biasing based on age, regional backgrounds and cultural viewpoints.

This study used the Reddit API to retrieve 8,000 informational posts in subreddits like news, info, information and world news using Python through PRAW package to present unbiased and unfiltered information about social media. The raw data was strictly cleaned and pre-processed in ways that had been done to current data before.

The purged dataset was initially fed through a trained fake news detection neural net based on a RoBERTa-LSTM classifier, which returned a real or fake post. As per the aims of the study, the fake posts were picked to be analyzed further. The research used the answers to these posts to assess the presence of mental health issues in users who submitted these posts and retrieved them through the Reddit API.

The responses related to the filtered fake posts were subsequently processed with a trained mental health implication model. In this model, the replies were categorized in terms of underlying mental health outcomes. Lastly, responses that were indicative of mental health conditions were analyzed using a disorder analysis model that categorized them into particular mental health disorders, including anxiety, borderline personality disorder (BPD), bipolar disorder, depression, and other disorders via RoBERTa-LSTM classifier [63]. Figure 4 shows the integrated architecture of this framework.

5 Result and analysis

5.1 Experimental setup

The system used in this study had the total disk storage capacity of 78.2 GB, 12.7 GB of system RAM and the NVIDIA-SMI GPU memory of 15 GB. The models employed in this paper were created by the Python package PRAW to access Reddit posts in real time via the Reddit API. A Reddit developer account provided access tokens and consumer keys that were used to access the API.

5.2 Evaluation metrics

Precision, recall, F1-score, and accuracy were used to evaluate the proposed method and the algorithms associated with it. These measures were computed using the values of True Positives (TPos), True Negatives (TNeg), False Positives (FPos) and False Negatives (FNeg) as given by Eqs. 1–4. These metrics are normally used in tasks associated with classifications and they measure the degree to which the target variable is correctly predicted by the model.

5.3 Results

This research has incorporated a three-step model of learning: first, the posts on Reddit are categorized as factual or deceptive; second, mental health implications are determined based on the responses retrieved in response to the deceptive posts; and third, mental health disorders are labeled. Although performance of the proposed

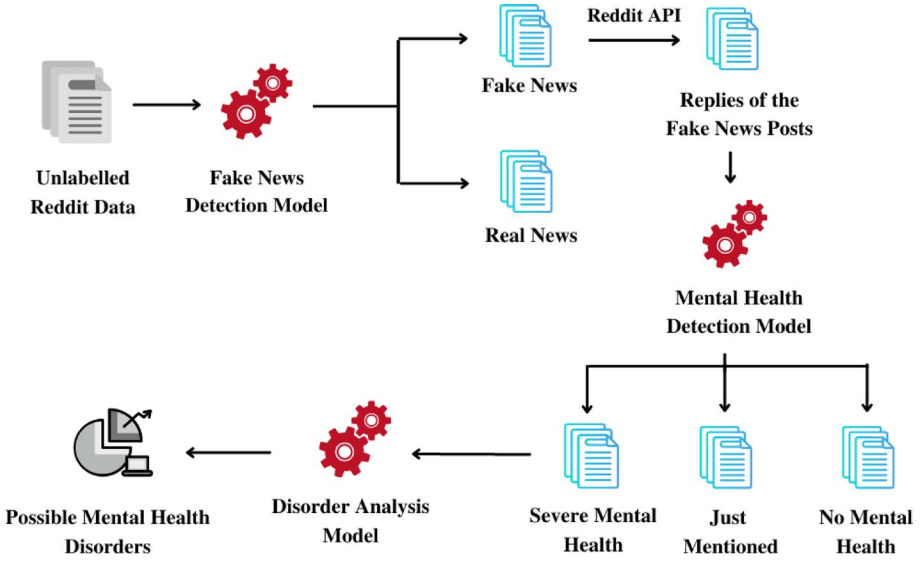


Fig. 4 Proposed Integrated Model Architecture

Table 3 Comparative Analysis of Fake News Detection Model

Model	Precision	Recall	F1-score	Accuracy (%)
LSTM	0.887	0.885	0.885	88.5
BERT	0.951	0.950	0.950	95.2
RoBERTa	0.964	0.964	0.964	96.4
RoBERTa-LSTM	0.984	0.984	0.984	98.4

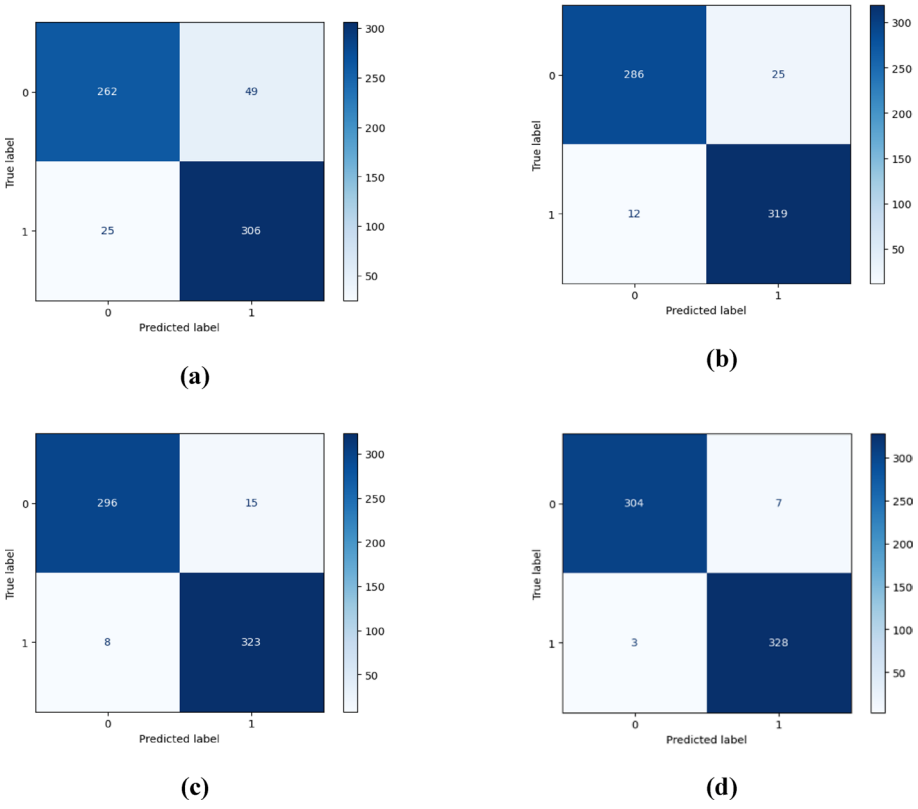


Fig. 5 Confusion Matrix of **a** LSTM, **b** BERT, **c** RoBERTa, and **d** RoBERTa-LSTM; for Fake News Detection

RoBERTa-LSTM transformer model has been measured in comparison to a number of algorithms that use accuracy, precision, f-1 score and recall, the accuracy score has been chosen as the primary evaluation measure. The analysis and findings of the proposed methodology are discussed in detail in the subsections below.

5.3.1 Hybrid approach for fake news detection

In order to assess the performance of the proposed RoBERTa-LSTM model to identify fake news posts, the models were trained on a structured dataset of 3,060 fake posts and 3,360 real posts. Compared analysis was made with the standalone RoBERTa and LSTM models. The hybrid RoBERTa-LSTM model has better performance, higher precision, recall, and F1-score as shown in Table 3. This is also evident in its testing accuracy of 98.4% accuracy. The confusion matrix visualizations of each model are shown in Fig. 5 and indicate the classification performance of each model. The RoBERTa-LSTM model has observed 7 False Positives (1.09% of total data points) and 3 False Negatives (0.46%

Table 4 Comparative Analysis of Mental Health Implication Model

Model	Precision	Recall	F1-score	Accuracy (%)
LSTM	0.809	0.809	0.808	80.8
BERT	0.834	0.833	0.833	83.3
RoBERTa	0.846	0.845	0.845	84.5
RoBERTa-LSTM	0.880	0.878	0.877	87.8

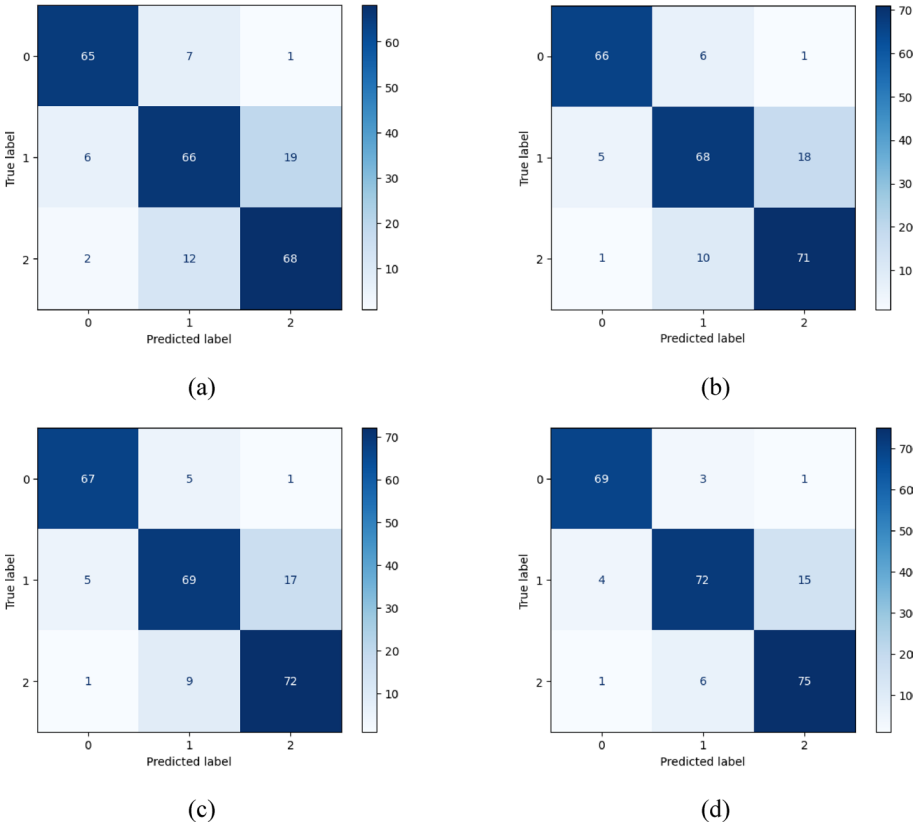


Fig. 6 Confusion Matrix of **a** LSTM, **b** BERT, **c** RoBERTa, and **d** RoBERTa-LSTM; for Mental Health Implication

of total data points), whose low values pose no threat in the real world and therefore do not act as driving tools for an incorrect prediction of fake news.

5.3.2 Mental health implication model

The same methodology as the one used in the previous model was used when classifying posts in terms of their mental health implications. The resultant comparative analysis is given in Table 4. The proposed RoBERTa-LSTM model reached a testing accuracy of 87.8 per cent, which was higher than all other models. Other measures such as precision, recall and F1-score are also described. The confusion matrices of the three models are shown in Fig. 6 which clearly show that RoBERTa-LSTM has highest True Positive and True Negative values among the four models:

5.3.3 Disorder analysis model

Once again, the proposed RoBERTa-LSTM model has been used to classify different mental health disorders with a testing accuracy of 77.3%. Other performance metrics

Table 5 Comparative Analysis of Mental Health Disorder Analysis Models

Model	Precision	Recall	F1-Score	Accuracy (%)
LSTM	0.717	0.724	0.719	72.4
BERT	0.735	0.744	0.738	74.4
RoBERTa	0.749	0.757	0.751	75.7
RoBERTa-LSTM	0.766	0.773	0.768	77.3

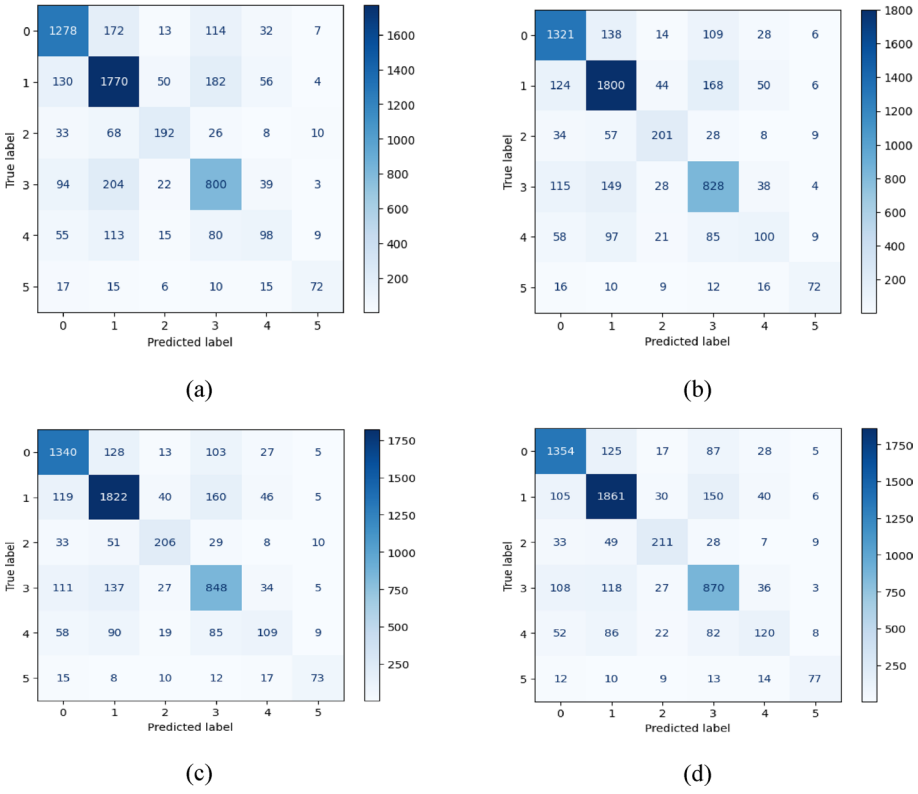


Fig. 7 Confusion Matrix of **a** LSTM, **b** BERT, **c** RoBERTa, and **d** RoBERTa-LSTM; for Mental Health Disorder Analysis

such as precision, recall and F1-score are presented in Table 5. Figure 7 depicts the confusion matrix plots of each model that visually represents the classification performance of each model which clearly indicates how the RoBERTa-LSTM model outperforms the others.

The datasets used in the Fake News Detection, Mental Health Implication and Disorder Analysis models have all been divided into 80:10:10 units to create Training Set, Testing Set and Validation Set respectively to make sure the model works in the real-life situation. The values of the loss functions have been utilized to quantify the differences between the speculated and the actual values of each model with the RoBERTa-LSTM Classifier. Figure 8 has shown the training and validation loss curves of each model and Fig. 9 has shown the training and validation accuracy curves of each model.

5.4 Analysis of the integrated model for the mental health implications of fake news

To survey the psychological effects of fake news on social media, this paper uses an unfiltered dataset of 8,000 news posts pulled through the Reddit API. After cleaning up these posts, they were then tested with the Fake News Detection Model. According

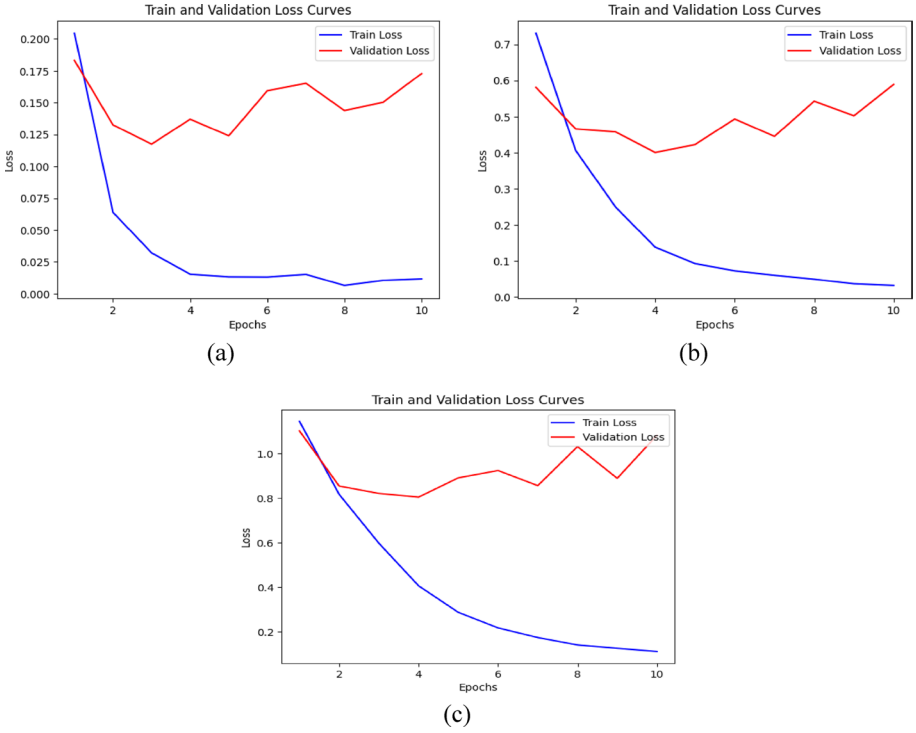


Fig. 8 Training (in blue) and Validation (in red) Loss Curves of Roberta-LSTM for **a** Fake News Detection, **b** Mental Health Implication and **c** Disorder Analysis Models

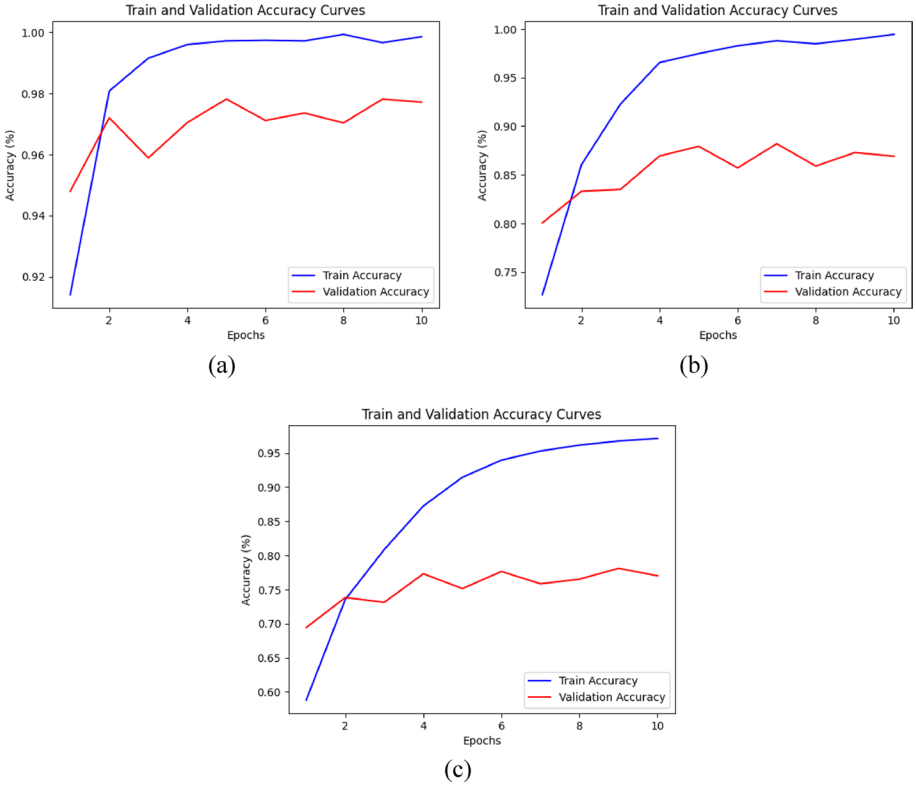


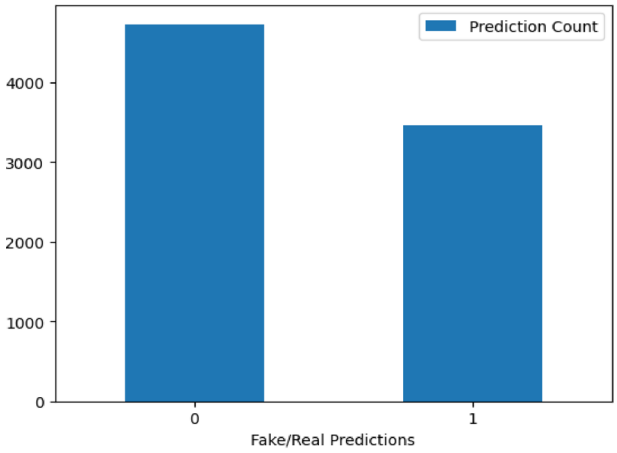
Fig. 9 Training Accuracy (in blue) and Validation Accuracy (in red) Curves of Roberta-LSTM for **a** Fake News Detection, **b** Mental Health Implication and **c** Disorder Analysis Models

to the trustworthiness of the facts of such posts, the result of this model distinguishes between the fake and the real as 0 and 1 respectively as shown in Fig. 9(a). As can be seen, it has been identified that about 4800 posts are fake and 3200 are real. To analyse the consequences of fake news on the users, the resulting responses to the fake news were retrieved via the Reddit API and then entered into the mental health implication model. In this case, we have checked whether the posts had some implications of mental health problems by categorizing the posts as having no indication of mental health problems (labeled as 0), posts implying mental health problems in forms of sarcasm and annoyance (labeled as 1) and posts implying strongly that the mental health problems occurred (labeled as 2). The output of this model in Fig. 9(b), shows the peak number of posts that reflect the mental health with a factor of one half of the total responses indicates a strong association between exposure to fake news and indicators of deteriorating mental health. Lastly, these replies were again analyzed through the Disorder Analysis model to determine the health disorders that may be caused by the fake posts, with the replies then being classified into the health challenges they imply: anxiety (labelled as 0), BPD (labelled as 1), bipolar (labelled as 2), depression (labelled as 3), schizophrenia (labelled as 4) and other mental illnesses (labelled as 5). Figure 9(c) has indicated the percentage of each disorder evaluated; BPD has the highest percentage of 38 with anxiety and depression at 35 and 22 respectively, with bipolar, schizophrenia and other illnesses having negligible values. The difference between the number of posts that show signs of mental illnesses and those that do not indicates the actuality of the mental health consequences of fake news on social media.

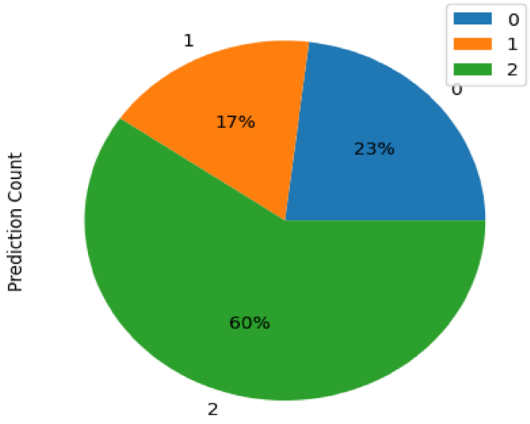
As illustrated in Fig. 10, the raw Reddit data was segmented and analyzed across three key dimensions: fake news detection, mental health implications, and disorder classification. Subfigure (a) presents the distribution of posts used to train and test the fake news detection model, reflecting a balanced mix of authentic and deceptive content. Subfigure (b) highlights the classification of posts based on their psychological tone, ranging from neutral to those suggestive of mental health concerns, underlining the model's sensitivity to nuanced emotional cues. Finally, subfigure (c) showcases how the data was further categorized for disorder analysis, mapping user-generated content to specific mental health conditions. Together, these distributions provide a foundational overview of how the dataset supports the layered structure of the integrated RoBERTa-LSTM framework.

6 Statistical test validation

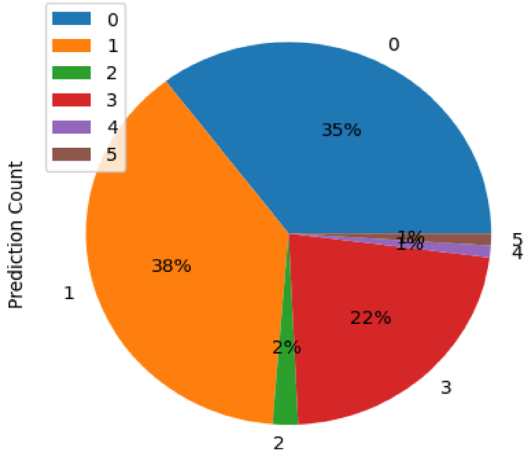
The main purpose of this study was to investigate the role of exposure to fake news as a cause of mental health problems in social media users. In order to understand a smooth integration of these two aspects, we had to identify whether the two categorical variables are interdependent. Without the connection, they would not have converged and the results would have been meaningless. In order to determine the robustness of the relationship between fake news and mental health, the research utilized Pearson Chi-Squared Test of Independence–a statistical tool that is applied to identify the existence of a significant correlation between two variables. This was aimed at determining whether mental health can be directly related to degradation due to exposure to misleading or deceitful information. This test was coded in Python and with the Scipy module used to analyze the Contingency Table computed by the integrated model structure as in Table 6. It is a crosstab 2×3 table between variables. Based on the responses to



(a)



(b)



(c)

Fig. 10 Distribution of Raw Reddit Data for **a** Fake News Detection Model **b** Mental Health Implication Model **c** Disorder Analysis Model

Table 6 Contingency Table

	0	1	2	Total
Fake	2554	1907	6203	10,664
Real	2305	1578	4992	8875
Total	4859	3485	11,195	19,539

Table 7 Contingency Table of Fake and Real News Across Mental Health Categories

	0	1	2	Total
Fake	2651.95	1902.04	6110.01	10,664
Real	2207.05	1582.96	5084.99	8875
Total	4859	3485	11,195	19,539

the fake posts retrieved on Reddit, 2554 responses show no mental health sign (0), 1907 show that mental health was used as sarcasm or displeasure (1), and 6203 show that there was a definite sign of mental illness (2). Equally, 2305, 1578 and 4992 replies have been fetched respectively as the replies of the posts classified as real.

At the core of this analysis lies the Null Hypothesis (H_0), which assumes to examine whether a statistically significant association exists between exposure to fake news and mental health implications. To challenge this assumption, the study used the p-value method, setting a conventional significance threshold (alpha) at 0.05. A p-value lower than this threshold would justify rejecting H_0 and confirm a statistically meaningful link between the two variables.

In this case, the data, summarized in Table 6, yielded a p-value of 0.003871, well below the 0.05 benchmark. This result leads to the rejection of the Null Hypothesis, indicating a significant correlation between fake news exposure and deteriorating mental health thus confirms a statistically significant dependency between the two categorical variables.

To reinforce these findings, the chi-squared statistic (χ^2) was also calculated. This method compares each observed value (O) with its expected counterpart (C) as defined in Equation 1 . The expected values, derived from the contingency table shown in Table 7, were then used to compute the chi-squared score following the steps in Equation 2 .

$$\chi^2 = \frac{(rowtotal \times columntotal)}{grandtotal} \tag{1}$$

$$\chi^2 = \sum \{ \frac{(Observed_i - Calculated_i)^2}{Calculated_i} \} \tag{2}$$

The calculated χ2 value is 11.1084. This method compares the computed χ2 value against a standard critical value of the chi-squared statistic to evaluate its significance. The Null Hypothesis (H0) is rejected if the calculated χ2 value surpasses the critical value. As shown in Table 8, the critical chi-squared (χ^2) value for a degree of freedom of 2, based on the size of the contingency matrix, is 5.991. The calculated χ^2 value surpasses this threshold, leading to the rejection of the Null Hypothesis once again. This analysis concludes that the circulation of misinformation over global social media platforms is significantly associated with adverse mental health indicators among users. While the Chi-Squared Test confirms statistical dependency between the variables, causal inferences are beyond the scope of this study and warrant further longitudinal investigation.

Table 8 Chi-Square Test Statistics for Fake News and Mental Health Association

Outcomes	
Alpha value	**0.05**
p—value	0.003871
Degree of freedom	2
Critical value of χ^2	5.991
Calculated value of χ^2	11.1084

7 Implications for industry and academia

This study contributes to academic understanding of the intersection between social media, misinformation and mental health. It also gives details on the potential consequences of false information on the mental health of people. Earlier research has shown that violent and hostile content on social media can influence the way people talk and the behavior of users. Among such cases, this would be the analysis of violent episodes in texts on social media characterized by a low source availability and, in this context, automated systems to recognize damaging information and mitigate the psychological impact of that information on a person will be highlighted to be the appropriate approaches [64]. The results of those studies support the idea that effective content moderation methods, including hybrid transformer-based RoBERTa-LSTM classifier and integrated architecture developed in this paper, need to be developed.

This methodology gives the academicians a guideline to research and develop on the same methodologies in future developments in the field. The proposed hybrid paradigm can be put to practical use by the industry-related professionals that are involved in the development of content moderation algorithms. Platforms can enhance their ability to identify and reduce the adverse impacts of fake news to the mental health of users. Also, studies on hostility management have focused on the need to develop intervention systems to respond to offensive and injurious messages on the internet [65]. These strategies can make a considerable donation to user safety and well-being when implemented in content moderation policies.

Further, the research points to areas that may be explored further, including the integration of a more diverse set of data. This is consistent with studies which examine opinion polarity identification and sentiment analysis specifically in user-produced content like movie reviews which illustrate how various views can be automatically coded to understand content more efficiently [66]. Researchers could explore these opportunities to improve their understanding of the global implications of the fake news outbreak on mental health. This can be through the integration of features or programs that can reduce the transmission of harmful false information and offer mental health support.

The study presents the potential adverse effects of free expression on social media platforms. Players in the industry must reevaluate rules and policies in order to strike an acceptable balance between free speech and the reduction of the negative ramifications of false information. Moreover, by examining the data of social media, researchers can investigate the ethical aspect of AI technology use in mental health analysis. This includes aspects related to privacy, consent, and the ethical use of AI in sensitive industries.

Social media platforms and IT corporations can utilize these insights to improve their content moderation algorithms, focusing on early detection and flagging of fake news that may cause psychological harm. By understanding the mental health effects,

platforms can implement more effective interventions, such as warning labels or providing users with mental health resources when they engage with potentially harmful content. Industry practitioners should reassess regulations and guidelines to achieve a harmonious equilibrium between freedom of speech and minimizing the adverse effects of misinformation. These findings can also be leveraged by governments to shape policies and regulations that address the spread of misinformation with a focus on public health. Moreover, these results can be useful to develop targeted therapies and support systems for individuals negatively impacted by misinformation. Additionally, mental health professionals can advocate for the integration of digital literacy and mental resilience training into educational curricula to better equip people to deal with the psychological effects of fake news.

To discuss the ethical implications of applying the proposed model to real social media data, it is essential to consider several key aspects. Firstly, data privacy and user consent are paramount. This study has used anonymized, publicly available data from Reddit, ensuring that no personally identifiable information (PII) was compromised. However, future applications of the model must strictly adhere to privacy regulations like General Data Protection Regulation (GDPR) or Digital Personal Data Protection Act (DPDPA) and prioritize obtaining explicit user consent. Additionally, the potential for bias and unfairness in the model's outcomes was addressed by using a diverse dataset, though recognizing that complete elimination of bias is challenging. Ongoing fairness audits will be necessary to mitigate any unintended discriminatory effects. Another critical ethical consideration involves the sensitivity of mental health data. The research emphasizes that while the model can provide insights into mental health implications, it should not be used as a diagnostic tool without the involvement of mental health professionals, as misinterpretation could have serious consequences. Moreover, the balance between mitigating misinformation and preserving freedom of expression is crucial. The model's application in content moderation should be carefully calibrated to avoid unjust suppression of legitimate discourse, maintaining transparency about the criteria used. The broader ethical use of AI in sensitive domains like mental health and misinformation detection also raises important questions. Continuous engagement with ethicists, legal experts, and affected communities is vital to ensure that the deployment of such technologies aligns with societal values and ethical standards. This includes being mindful of the potential consequences of errors in the model's predictions and the responsibilities of those deploying these technologies. Finally, as the model and research evolve, it is important to regularly revisit and reassess these ethical considerations to ensure the technology's responsible and fair use in real-world scenarios.

Overall, this study can reflect on academic discussion, future study and industrial precepts through enlightenment on a complex dependence on the misinformation, social media and mental illness. It advocates the integrative perspective in consideration of the social impact of technological advancement.

8 Shortcomings

The RoBERTa-LSTM hybrid model, despite its strengths, has notable limitations. A primary concern is its handling of long sequences, as RoBERTa's fixed input size, typically limited to 512 tokens, can result in the loss of important information while processing longer texts, negatively impacting performance. Furthermore, the combination of

RoBERTa and LSTM increases the model's complexity, leading to higher computational demands and extended training times. This complexity also heightens the risk of overfitting, particularly on smaller datasets.

The three datasets have notable shortcomings that could impact their effectiveness in model training and research. Source diversity is a concern across all datasets, as the datasets are limited to Twitter and Reddit, reducing their representativeness and generalizability. For the Fake News Dataset (6420 posts) and Mental Health Implication Dataset (2460 posts), their small sizes could lead to overfitting. The Disorder Analysis Dataset (58,116 data points) suffers from significant class imbalance, with certain disorders like BPD being overrepresented compared to others like Schizophrenia, potentially biasing the model's predictions. Additionally, labeling subjectivity in datasets, particularly those involving mental health, might reflect annotator biases or inconsistencies, impacting the reliability of the data and the accuracy of the models trained on it. Overall, these biases and limitations can significantly affect the validity and generalizability of the study's findings.

9 Conclusion and future scope

Technology has changed the way we live, work and socialize. Although social media has made the world a smaller global village, its application in the uncontrolled proliferation of misinformation in the service of foreign interests has played an active role in psychological distress and damage. In this paper, we provide an in-depth understanding of the psychological health impact of misinformation and why early detection and response are of the essence. At the heart of this research is a novel RoBERTa-LSTM hybrid model, which achieved impressive accuracy rates: 98.4% for fake news detection, 87.8% for identifying mental health implications, and 77.3% for disorder classification.

By integrating these three models and applying them to an unbiased, real-world dataset sourced from Reddit, the study offers a data-driven evaluation of how digital falsehoods affect mental well-being across social platforms. Statistical validation using Pearson's Chi-Squared Test (p-value = 0.003871) revealing a statistically significant association between misinformation exposure and adverse mental health outcomes. These findings may act as a contributing factor for critical foundation not only for identifying harmful influences but also for informing the design of proactive strategies that promote emotional resilience online. Figure 11 outlines a research-to-impact journey, starting with misinformation and mental health detection, and culminating in hope speech detection and its role in building brand trust and digital well-being.

Looking ahead, this research serves as an essential precursor to the development of hope speech detection frameworks. Building on the ability to identify risk factors, future work can focus on amplifying positive, supportive communication across diverse linguistic and cultural contexts, including low-resource languages. Such advancements have the potential to support businesses, educational institutions, and online communities in cultivating healthier, more inclusive digital ecosystems. By bridging the gap between harmful content detection and hope speech promotion, this line of research can play a transformative role in enhancing digital well-being, strengthening brand trust, and promoting community resilience. Future research may explore the integration of more advanced transformer architectures and improved hybrid models to further enhance classification performance. While this research primarily targets English-language data,

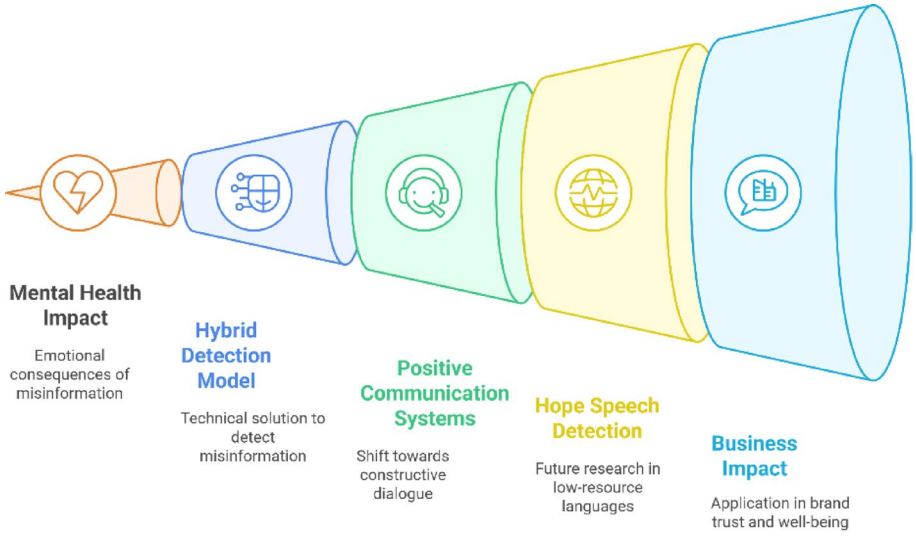


Fig. 11 Combating Misinformation for Positive Impact

its model is adaptable to other low-resource languages. Therefore, this research can be further extended by including a more diverse dataset, including data from languages with complex scripts like Arabic or Thai which could help refine the model's ability to process non-Latin scripts. By testing the model with social media data from these languages, we can better understand how fake news spreads in different linguistic and cultural settings. Additionally, expanding the model to consider variables such as gender, ethnicity, and socioeconomic status could provide deeper insights into how different demographic groups are affected by and respond to fake news. This will further enhance the model's ability to capture the global repercussions of the fake news epidemic and provide targeted solutions for mitigating its effects across diverse populations. The use of larger, and better-validated datasets, including those with improved annotation quality or expert verification, could help reduce noise and bias while improving reliability. Additionally, cross-dataset evaluation, model explainability, and bias-aware training strategies may be investigated to support robust and responsible deployment in sensitive application domains.

Author contributions
S.A., S., V.A. and D.K.: conceptualisation, methodology, and validation; D.K., R.P., V.A and A.A.K: formal analysis; D.K., R.P. and A.A.K.: resources; S.A., S. and V.A.: data curation; S., D.K., and R.P: review and editing; S.A, D.K, V.A. and A.A.K.: visualization; R.P., D.K. and A.A.K.: project administration; All authors reviewed the manuscript.

Funding
The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data Availability
All data generated or analyzed during this study are included in this published article.

Declarations

Ethics approval and consent to participate
This research did not involve. Any human or animal participation. All authors have checked and agreed on the submission. Furthermore, the study recognizes the ethical considerations associated with predicting mental health or classifying online content, including privacy, consent, and the potential for misinterpretation, and urges caution when extrapolating conclusions to broader populations.

Competing Interests
The authors declare no competing interests.

Received: 30 October 2025 / Accepted: 3 February 2026

Published online: 18 February 2026

References

1. Brewer MB. The importance of being we: human nature and intergroup relations. Am Psychol. 2007;62(8):728.
2. Rosen LD, Whaling K, Rab S, Carrier LM, Cheever NA. Is Facebook creating "iDisorders"? The link between clinical symptoms of psychiatric disorders and technology use, attitudes and anxiety. Comput Human Behav. 2013;29:1243–54.
3. Parveen F, Jaafar NI, Ainin S. Social media usage and organizational performance: reflections of Malaysian social media managers. Telemat Informatics. 2015;32(1):67–78.
4. M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, A. Flammini, and F. Menczer. Political polarization on twitter. In Proc. 5th International AAAI Conference on Weblogs and Social Media (ICWSM), 2011.
5. Conover MD, Gonçalves B, Flammini A, Menczer F. Partisan asymmetries in online political activity. EPJ Data Sci. 2012;1:6.
6. Michael RB, Breaux BO. The relationship between political affiliation and beliefs about sources of "fake news." Cogn Res. 2021;6:6. https://doi.org/10.1186/s41235-021-00278-1.
7. Patel, P., Kannoorpatti, K., Shanmugam, B., Azam, S., & Yeo, K. C. (2017, January). A theoretical review of social media usage by cyber-criminals. In 2017 International Conference on Computer Communication and Informatics (ICCCI) (pp. 1–6). IEEE.
8. The American Psychiatric Association. Press Release: Americans Say They are More Anxious than a Year Ago; Baby Boomers Report Greatest Increase in Anxiety. New York, May 7, 2018. Available at: https://www.psychiatry.org/newsroom/news-releases/americans-say-they-are-more-anxious-than-a-year-ago-baby-boomers-report-greatest-increase-in-anxiety Accessed April 1, 2019.
9. Rocha YM, de Moura GA, Desidério GA, et al. The impact of fake news on social media and its influence on health during the COVID-19 pandemic: a systematic review. J Public Health (Berl). 2021. https://doi.org/10.1007/s10389-021-01658-z.
10. WHO. International Statistical Classifi cation of Diseases and Related Health Problems, 10th revision. Geneva, Switzerland: World Health Organization, 1992–94.
11. https://www.weforum.org/agenda/2019/01/this-is-the-worlds-biggest-mental-health-problem/
12. Xiong J, Lipsitz O, Nasri F, Lui LMW, Gill H, Phan L, et al. Impact of COVID-19 pandemic on mental health in the general population: a systematic review. J Affect Disord. 2020;277:55–64. https://doi.org/10.1016/j.jad.2020.08.001.
13. Shu K, Mahudeswaran D, Liu H. FakeNewsTracker: a tool for fake news collection, detection, and visualization. Comput Math Organ Theory. 2019;25:60–71. https://doi.org/10.1007/s10588-018-09280-3.
14. Abdullah Alsaeedi and Mohammad Zubair Khan, "A Study on Sentiment Analysis Techniques of Twitter Data" International Journal of Advanced Computer Science and Applications(IJACSA), 10(2), 2019.
15. Gkotsis, G., Oellrich, A., Velupillai, S., Liakata, M., Hubbard, T. J. P., Dobson, R. J. B., & Dutta, R. (2017). Characterisation of mental health conditions in social media using Informed Deep Learning. Scientific Reports, 7(1). https://doi.org/10.1038/srep45141
16. A. Dey, R. Z. Rafi, S. Hasan Parash, S. K. Arko and A. Chakrabarty, "Fake News Pattern Recognition using Linguistic Analysis," 2018 Joint 7th International Conference on Informatics, Electronics & Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR), 2018, pp. 305–309, https://doi.org/10.1109/ICIEV.2018.8641018.
17. Zhang Y, Jin R, Zhou ZH. Understanding bag-of-words model: a statistical framework. Int J Mach Learn & Cyber. 2010;1:43–52. https://doi.org/10.1007/s13042-010-0001-0.
18. Kramer, O. (2013). K-Nearest Neighbors. In: Dimensionality Reduction with Unsupervised Nearest Neighbors. Intelligent Systems Reference Library, vol 51. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-38652-7_2
19. Graves, A. (2012). Long Short-Term Memory. In: Supervised Sequence Labelling with Recurrent Neural Networks. Studies in Computational Intelligence, vol 385. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-24797-2_4
20. Medsker LR, Jain LC. Recurrent neural networks. Design and Applications. 2001;5:64–7.
21. Dey, R., & Salem, F. M. (2017, August). Gate-variants of gated recurrent unit (GRU) neural networks. In 2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS) (pp. 1597–1600). IEEE.
22. O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458.
23. S. Girgis, E. Amer and M. Gadallah, "Deep Learning Algorithms for Detecting Fake News in Online Text," 2018 13th International Conference on Computer Engineering and Systems (ICCES), 2018, pp. 93–97, https://doi.org/10.1109/ICCES.2018.8639198
24. Youness Madani, Mohammed Erritali, Belaid Bouikhalene, Using artificial intelligence techniques for detecting Covid-19 epidemic fake news in Moroccan tweets, Results in Physics, Volume 25, 2021, 104266, ISSN 2211–3797, https://doi.org/10.1016/j.rinp.2021.104266.
25. Biau G, Scornet E. A random forest guided tour. TEST. 2016;25:197–227.
26. Noble WS. What is a support vector machine? Nat Biotechnol. 2006;24(12):1565–7.
27. Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41–46).
28. Sinha A, Raval M, S S. Machine learning based detection of deceptive tweets on Covid-19. Int J Eng Adv Technol. 2021;10:375–80. https://doi.org/10.35940/ijeat.E2831.0610521.
29. Pierri F, Piccardi C, Ceri S. A multi-layer approach to disinformation detection in US and Italian news spreading on Twitter. EPJ Data Sci. 2020;9:35. https://doi.org/10.1140/epjds/s13688-020-00253-8.
30. LaValley MP. Logistic regression. Circulation. 2008;117(18):2395–9.
31. Bansal R., Paka W.S., Nidhi, Sengupta S., Chakraborty T. (2021) Combining Exogenous and Endogenous Signals with a Semi-supervised Co-attention Network for Early Detection of COVID-19 Fake Tweets. In: Karlapalem K. et al. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2021. Lecture Notes in Computer Science, vol 12712. Springer, Cham. https://doi.org/10.1007/978-3-030-75762-5_16
32. Li F, Jin Y, Liu W, Rawat BPS, Cai P, Yu H. Fine-tuning bidirectional encoder representations from transformers (BERT)–based models on large-scale electronic health record notes: an empirical study. JMIR Med Inform. 2019;7(3):e14830.
33. Seddari N, Derhab A, Belaoued M, Halboob W, Al-Muhtadi J, Bouras A. A hybrid linguistic and knowledge-based analysis approach for fake news detection on social media. IEEE Access. 2022;10:62097–109. https://doi.org/10.1109/ACCESS.2022.3181184.

34. Sheng, Q., Cao, J., Zhang, X., Li, R., Wang, D., & Zhu, Y. (2022). Zoom out and observe: News environment perception for fake news detection. arXiv preprint arXiv:2203.10885.
35. Raza S, Ding C. Fake news detection based on news content and social contexts: a transformer-based approach. Int J Data Sci Anal. 2022;13:335–62. https://doi.org/10.1007/s41060-021-00302-z.
36. Alyoubi S, Kalkatawi M, Abukhodair F. The detection of fake news in Arabic tweets using deep learning. Appl Sci. 2023;13:8209. https://doi.org/10.3390/app13148209.
37. Oberiri Destiny Apuke, Bahiyah Omar, Fake news and COVID-19: modelling the predictors of fake news sharing among social media users, Telematics and Informatics, Volume 56, 2021, 101475, ISSN 0736–5853, https://doi.org/10.1016/j.tele.2020.101475.
38. Baek K, Holton A, Harp D, Yaschur C. The links that bind: uncovering novel motivations for linking on Facebook. Comput Human Behav. 2011;27(6):2243–8. https://doi.org/10.1016/j.chb.2011.07.003.
39. Keyes CLM. 'Promoting and protecting mental health as flourishing: A complementary strategy for improving national mental health.' Amer Psychologist. 2007;62(2):95–108.
40. Kim J, Lee J, Park E, Han J. A deep learning model for detecting mental illness from user content on social media. Sci Rep. 2020. https://doi.org/10.1038/s41598-020-68764-y.
41. Zou, L., Zheng, J. & McKeown, M. J. Deep learning based automatic diagnoses of attention deficit hyperactive disorder. In Proc. 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP) 962–966 (Montreal, Canada, 2017).
42. Lin, Huijie & Jia, Jia & Qiu, Jiezhong & Zhang, Yongfeng & Shen, Guangyao & Xie, Lexing & Tang, Jie & Feng, Ling & Chua, Tat-Seng. (2017). Detecting Stress Based on Social Interactions in Social Networks. IEEE Transactions on Knowledge and Data Engineering. PP. 1–1. https://doi.org/10.1109/TKDE.2017.2686382.
43. Zhang, Z., Lin, W., Liu, M., & Mahmoud, M. (2020). Multimodal Deep Learning Framework for Mental Disorder Recognition. 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). https://doi.org/10.1109/fg47880.2020.00033
44. Joshi, D. J., Makhija, M., Nabar, Y., Nehete, N., & Patwardhan, M. S. (2018). Mental health analysis using deep learning for feature extraction. Proceedings of the ACM India Joint International Conference on Data Science and Management of Data - CoDS-COMAD '18. https://doi.org/10.1145/3152494.3167990
45. O'Dea B, et al. Detecting suicidality on Twitter. Internet Interv. 2015;2(2):183–8.
46. Joshi D, Patwardhan DM. An analysis of mental health of social media users using unsupervised approach. Comput Human Behav Rep. 2020;2:100036. https://doi.org/10.1016/j.chbr.2020.100036.
47. Sumathi, M. R., & Poorna, B. (2016). Prediction of mental health problems among children using machine learning techniques. International Journal of Advanced Computer Science and Applications, 7(1).
48. Masri, R.Y.; Jani, H.M., "Employing artificial intelligence techniques in Mental Health Diagnostic Expert System," in Computer & Information Science (ICCIS), 2012 International Conference on , vol.1, no., pp.495–499,12–14 June 2012 https://doi.org/10.1109/ICCISci.2012.6297296
49. Zarate D, Ball M, Prokofieva M, Kostakos V, Stavropoulos V. Identifying self-disclosed anxiety on Twitter: a natural language processing approach. Psychiatr Res. 2023;330:115579.
50. Rawat, N., Chauhan, S., & Awasthi, L. K. (2024, March). Deep Learning Approaches for Predicting Mental States through Tweet Analysis. In 2024 5th International Conference on Innovative Trends in Information Technology (ICITIIT) (pp. 1–6). IEEE.
51. Al Banna MH, Ghosh T, Al Nahian MJ, Kaiser MS, Mahmud M, Taher KA, et al. A hybrid deep learning model to predict the impact of COVID-19 on mental health from social media big data. IEEE Access. 2023;11:77009–22.
52. Kokane, V., Abhyankar, A., Shrirao, N., & Khadkikar, P. (2024, May). Predicting Mental illness (Depression) with the help of NLP Transformers. In 2024 Second International Conference on Data Science and Information System (ICDSIS) (pp. 1–5). IEEE.
53. Gupta, K., Jinad, R., & Liu, Q. (2023, June). Comparative analysis of nlp models for detecting depression on twitter. In 2023 International Conference on Communications, Computing and Artificial Intelligence (CCCAI) (pp. 23–28). IEEE.
54. Deshpande, V., Lee, M., Yao, Z., Zhang, Z., Gibbons, J. B., & Yu, H. (2024). LocalTweets to LocalHealth: A Mental Health Surveillance Framework Based on Twitter Data. arXiv preprint arXiv:2402.13452.
55. Luna McBride: "Constraints_Train.csv," in "COVID-19 Tweet Truth Analysis," Kaggle. [Online]. Available: https://www.kaggle.com/code/lunamcbride24/covid19-tweet-truth-analysis/input. Accessed: April. 13, 2022.
56. Surana, Praatibh; Yusuf, Mirza; Singh, Sanjay (2022). ADAM-SDMH: A DAtaset from Manipal for Severity Detection in Tweets related to Mental Health. figshare. Dataset. https://doi.org/10.6084/m9.figshare.19029656.v2
57. Kamarul Adha. (2022). Mental Disorders Identification (Reddit) . Kaggle. https://doi.org/10.34740/KAGGLE/DSV/4579285
58. Naseer M, Windiatmaja JH, Asvial M, Sari RF. RoBERTaEns: deep bidirectional encoder ensemble model for fact verification. Big Data Cogn Comput. 2022;6(2):33. https://doi.org/10.3390/bdcc6020033.
59. Van Houdt G, Mosquera C, Nápoles G. A review on the long short-term memory model. Artif Intell Rev. 2020. https://doi.org/10.1007/s10462-020-09838-1.
60. Semary NA, Ahmed W, Amin K, Pławiak P, Hammad M. Improving sentiment classification using a RoBERTa-based hybrid model. Front Hum Neurosci. 2023;17:1292010. https://doi.org/10.3389/fnhum.2023.1292010.
61. Tan KL, Lee CP, Anbananthen KSM, Lim KM. RoBERTa-LSTM: a hybrid model for sentiment analysis with transformer and recurrent neural network. IEEE Access. 2022;10:21517–25. https://doi.org/10.1109/ACCESS.2022.3152828.
62. Yunita A, Pratama MI, Almuzakki MZ, Ramadhan H, Akhir EAP, Firdausiah Mansur AB, et al. Performance analysis of neural network architectures for time series forecasting: a comparative study of RNN, LSTM, GRU, and hybrid models. MethodsX. 2025;15:103462. https://doi.org/10.1016/j.mex.2025.103462.
63. Bansal, V., Tyagi, M., Sharma, R., Gupta, V., & Xin, Q. (2022). A Transformer Based Approach for Abuse Detection in Code Mixed Indic Languages. ACM transactions on Asian and low-resource language information processing.
64. Sharma, D., Gupta, V., Singh, V. K., & Pinto, D. (2024). Should we stay silent on violence? An ensemble approach to detect violent incidents in Spanish social media texts. Natural Language Processing, 1–20.
65. Raina, S. T., Mathur, A., Goyal, C., & Gupta, V. (2021). A novel method for hostility management. In Proceedings of International Conference on Computational Intelligence, Data Science and Cloud Computing: IEM-ICDC 2020 (pp. 93–101). Springer Singapore.

66. Gupta V, Jain N, Shubham S, Madan A, Chaudhary A, Xin Q. Toward integrated CNN-based sentiment analysis of tweets for scarce-resource language,Hindi. Transactions on Asian and Low-Resource Language Information Processing. 2021;20(5):1–23.

Publisher's Note