Engineering Reports

WILEY

Engineering Reports Open Access

RESEARCH ARTICLE OPEN ACCESS

Sentiment Analysis-Based Model for Monitoring User Engagement With Mental Health Chatbots

Ian Igado Mmbayi | Consolata Gakii | Faith Mueni Musyoka

Department of Computing and Information Technology, University of Embu, Embu, Kenya

Correspondence: Ian Igado Mmbayi (igadoian001@gmail.com)

Received: 17 March 2025 | **Revised:** 3 June 2025 | **Accepted:** 6 June 2025

Funding: The authors received no specific funding for this work.

Keywords: aspect-based analysis | BERT | chatbots | machine learning | mental health | sentiment analysis | SMOTE | user reviews

ABSTRACT

Mental health challenges, particularly among youth, are compounded by stigma and limited access to professional care. This has driven demand for scalable digital solutions like chatbots. This study introduces a sentiment analysis-based model to assess user satisfaction with mental health chatbots, analyzing 82 102 reviews from six popular apps on Google Play and Apple's App Stores. A multi-class sentiment classification of positive, negative, and neutral was implemented, enhanced by Synthetic Minority Over-sampling Technique for class balancing, comparing five traditional machine learning models with Bidirectional Encoder Representations from Transformers, a transformer model. Random Forest achieved 98.18% accuracy among traditional models, while BERT outperformed all with 99.17% accuracy, surpassing prior benchmarks. Aspect-based analysis revealed that Emotion and Usability drive positive feedback, while Reliability issues fuel negative sentiments, offering actionable insights for developers to enhance chatbot design. This work advances digital mental health research by integrating multi-class classification, transformer models, and aspect-based analysis, demonstrating a scalable framework for evaluating user feedback.

1 | Introduction

Mental health is a critical global challenge, with disorders like depression and anxiety affecting over 300 million people, particularly youth, and undermining societal well-being [1, 2]. The escalating prevalence of these conditions, coupled with barriers such as stigma, limited access to professionals, and high costs, has intensified the demand for innovative, scalable solutions [3]. AI-driven mental health chatbots have gained traction as cost-effective tools, delivering real-time support through cognitive-behavioral therapy, mindfulness exercises, and psychological guidance [4]. Their rapid adoption, evidenced by millions of downloads on app stores, underscores the urgent need to evaluate their effectiveness and user satisfaction to ensure they address diverse mental health needs [5]. These chatbots offer a promising alternative, yet their real-world impact depends on understanding user experiences through robust analytical methods.

Sentiment analysis of chatbot reviews holds immense potential but is hindered by significant limitations. Most studies rely on binary classification (positive vs. negative), which fails to capture neutral or ambivalent sentiments essential for understanding complex user emotions [5, 6]. For instance, binary models overlook subtle dissatisfaction or mixed feelings, limiting their ability to identify specific drivers like reliability, content relevance, or usability [7]. Moreover, existing research rarely compares traditional machine learning models with advanced transformers like BERT, restricting the depth of sentiment analysis

in mental health contexts [8]. These methodological gaps prevent chatbots from evolving into responsive tools that meet users' emotional and therapeutic needs.

The growing dependence on mental health chatbots necessitates fine-grained analysis to enhance their emotional responsiveness, technical reliability, and therapeutic efficacy. Multi-class sentiment analysis, capturing positive, negative, and neutral sentiments, provides a richer understanding of user interactions, revealing nuanced patterns missed by binary approaches [6, 9]. Integrating traditional machine learning with transformers like BERT offers a comprehensive approach, addressing some methodological limitations and enhancing feedback evaluation [10]. Such advancements are critical for tailoring chatbots to diverse user needs and improving mental health outcomes.

This study analyses 82 102 mental health chatbot reviews to provide actionable insights for improving user satisfaction and effectiveness. A multi-class sentiment classification was implemented, using SMOTE for class balancing, and conducted aspect-based analysis to assess reliability, content, usability, and emotional support. Five traditional machine learning models (Random Forest, SVM, Logistic Regression, SGD, Naïve Bayes) were evaluated alongside BERT, offering a novel comparative framework. Predictions for 1151 unlabeled reviews demonstrate the framework's practical value in real-world applications. This study contributes through:

- Developing a sentiment analysis model that employs multi-class classification to capture positive, negative, and neutral user sentiments, enabling nuanced insights into user engagement with mental health chatbots.
- Implementing aspect-based analysis to identify key factors influencing user experience (reliability, content, usability, and emotion), providing actionable recommendations for chatbot design improvements.

The paper is structured as follows: Section 2 reviews literature, Section 3 details methodology, Section 4 presents results, Section 5 discusses findings, and Section 6 concludes.

2 | Literature Review

Sentiment analysis is pivotal for evaluating user feedback on digital mental health tools like chatbots, addressing a global crisis affecting over 300 million people [1]. This section reviews research on mental health chatbots, lexicon-based sentiment analysis, healthcare applications, and machine learning models, highlighting gaps that justify advanced approaches in sentiment classification.

2.1 | Mental Health Chatbots and User Feedback

Mental health chatbots, such as Woebot and Wysa, provide cognitive-behavioral therapy, mindfulness exercises, and emotional support, overcoming barriers like stigma and cost [3, 4]. App store reviews offer insights into usability, emotional resonance, and technical stability, but analyzing them is challenging due to complex user sentiments [9]. Abd-Alrazaq et al. [5] analyzed ~5000 chatbot reviews, noting positive feedback on accessibility but frequent technical issues, such as app crashes. Caldeira et al. [11] found users valued simple interfaces but raised concerns about privacy and personalization. Shickel et al. [12] observed positive usability feedback but noted difficulties capturing nested sentiments, such as positive reviews with underlying concerns. Gkinko and Elbanna [7] criticized prior studies for using small datasets (< 20 000 reviews) and qualitative methods with limited generalizability. The absence of demographic data e.g., age, sex in reviews further restricts user-specific analyses, a persistent limitation for tailoring chatbot improvements [13].

2.2 | Lexicon-Based Sentiment Analysis

Lexicon-based sentiment analysis uses predefined word lists to classify text as positive, negative, or neutral, valued for its interpretability. TF-IDF identifies key terms (e.g., "helpful," "crash") in app reviews, but struggles with nuanced emotions prevalent in mental health feedback [6]. Saju et al. [14] noted that lexicon-based methods are straightforward but fail to detect sarcasm or humor in sensitive domains. Ribeiro et al. [15] found that missing context leads to misclassification biases for intense emotions. Khan et al. [10] developed multilingual lexicons for sentiment analysis on social media, using TF-IDF to identify key terms. Their approach achieved robust binary classification but struggled with neutral or ambivalent sentiments, critical for mental health chatbot feedback, which often includes complex emotions [9]. This limitation, also evident in app store reviews with technical and emotional drivers, underscores the need for multi-class classification in our study. Lexicon-based approaches rarely address class imbalance or support fine-grained sentiment tasks, necessitating advanced methods [7].

2.3 | Sentiment Analysis in Healthcare

Sentiment analysis enhances healthcare by analyzing patient feedback, with emerging applications in mental health. Abd-Alrazaq et al. [5] used binary classification on ~5000 chatbot reviews, identifying positive and negative sentiments but missing neutral feedback, limiting nuanced insights. Mehmood et al. [8] achieved 88% accuracy in binary analysis of health-related social media but failed to capture specific drivers like Reliability or Usability. Wen [16] detected depression signals on Twitter, suggesting potential for non-clinical monitoring, but did not address structured platforms like chatbots. Menon and George [17] identified public health trends via social media sentiment analysis but overlooked interactive digital tools, such as chatbots. Aspect-based analysis, which could pinpoint specific feedback drivers (e.g., Content, Emotional Support), is underutilized in healthcare, restricting actionable insights for developers [7]. Small datasets, typically 5000–20 000 reviews, further constrain generalizability, particularly for specialized mental health applications [6].

2.4 | Machine Learning and Transformer Models

Traditional machine learning models, such as Random Forest, Support Vector Machines (SVM), and Naïve Bayes, are widely used in sentiment analysis but plateau at ~85%–88% accuracy for complex datasets [8]. Yadollahi et al. [18] applied supervised models, noting strong performance but cautioning their limitations in emotionally nuanced domains like mental health. Class imbalance, prevalent in review datasets, undermines model accuracy, with techniques like SMOTE rarely utilized to address this issue [6]. Transformer-based models like BERT capture contextual relationships, enhancing classification accuracy [19]. Khan et al. [19] applied an attention-based transformer to detect emotions from handwriting and drawing samples, leveraging self-attention to capture nuanced patterns with high accuracy. While not focused on healthcare or text, their success in complex emotion detection supports BERT's applicability to mental health chatbot reviews, where nuanced feedback is prevalent [7]. Azam et al. [20] achieved 94% accuracy detecting depression on social media using deep learning, but their focus on unstructured data limits applicability to chatbots. Tang et al. [21] showed deep learning models outperform traditional ones in capturing subtle sentiments, yet their use in mental health chatbots remains unexplored. Comparative studies of traditional and transformer models are scarce, particularly for emotionally complex feedback [7].

2.5 | Research Gaps

Prior studies predominantly rely on binary classification, small datasets, and traditional machine learning models, failing to capture neutral sentiments, aspect-based insights, or contextual nuances in mental health chatbot feedback. The lack of demographic data limits user-specific analyses, while the scarcity of transformer-based and comparative studies hinders methodological progress. These gaps necessitate advanced sentiment analysis techniques, such as multi-class classification and transformer models, as addressed in this study, to enhance chatbot design and user satisfaction.

3 | Methodology

3.1 | Data Collection

This study focused on user reviews of mental health chatbots collected from two major platforms: Google Play Store and Apple's App Store. This included real-time chatbots powered by AI that provide support for mental health issues such as anxiety, depression, and stress. The study was carried out to identify chatbot apps that have the main functionality of mental support, apply the chatbot function to provide immediate response to the user, and have a great number of reviews for both these platforms so that the sentiment analysis dataset was complete. Below Figure 1 summarizes the flow to the steps undertaken.

Seven chatbot apps meet the above criteria out of which are seven Cups, Amaha (previously known as InnerHour) that concentrates on stress and mood; Sintelly that provides Cognitive Behavior Therapy (CBT) for Post-Traumatic Stress Disorder (PTSD); VOS that delivers Therapy; Wysa that is a Therapy Chatbot; and Youper that is a CBT Therapy Chatbot. Each of these applications had more than 500 000 downloads and 5000 or more reviews pointing to the extensive use and adequate feedback data thus seeking space provision regarding mental health support even more using digital resources.

The Heedzy web scraping tool was evident [22] to collect the comments. The dataset that was gathered comprised 82 102 reviews, comprised of the actual content of the review, the star rating the

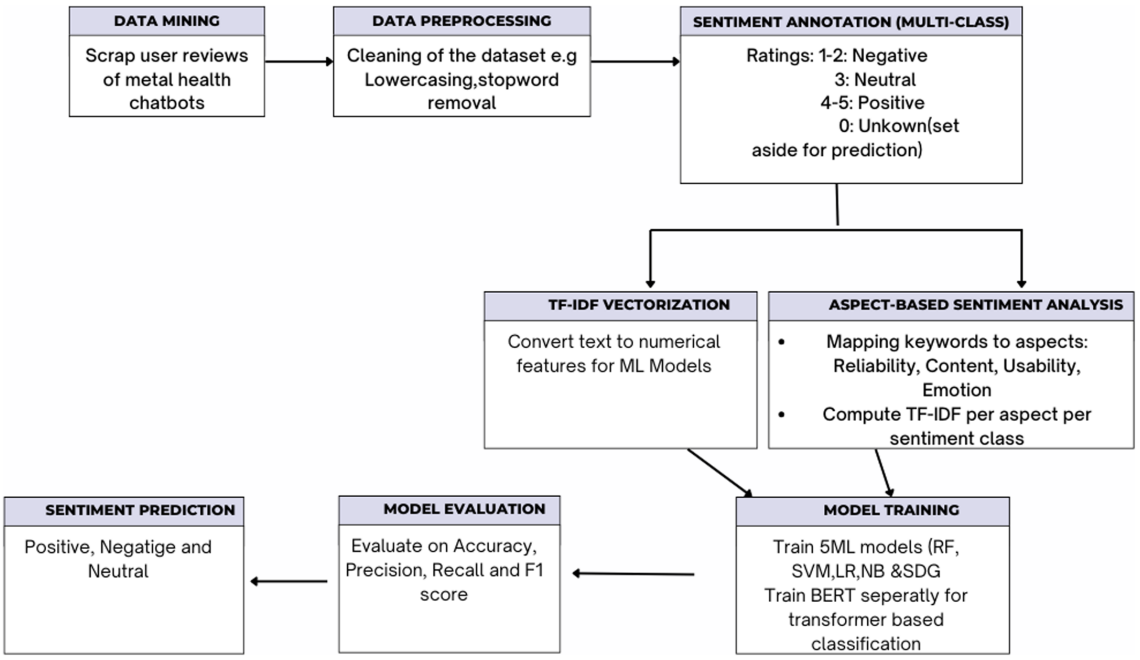

FIGURE 1 | Sentiment analysis flowchart.

review was awarded (from a minimum of 1 to a maximum of 5), the time and date the review was made, and Android or iOS on which the platform was published. This gave a vast dataset that helped to explore the user sentiment toward mental health chatbots.

3.2 | Data Preprocessing

Before conducting the analysis, several cleaning steps were applied to the collected data (reviews) for analysis. The preprocessing step started with transforming all text content into lowercase to establish consistency while eliminating case sensitivity problems. The Natural Language Tool Kit (NLTK) stop word list, as per the study of Patel and Passi [23], helped to eliminate disorder by removing common words "the," "is," and "in" from the reviews. The WordNet Lemmatizer processed all reviews by converting word forms into their base lexical units to maintain data consistency. A sentiment analysis distortion-prevention measure involved removing punctuation with the additional removal of special characters together with numbers thus, inadvertently stripped garbled emoji sequences (e.g., "ðŸ'–" from reviews like "The best ðŸ'–ðŸ'–ðŸ'–"), which likely originated as emojis but appeared as encoding artifacts in the CSV export. Reviews were also normalized by correcting misspellings and expanding abbreviations (e.g., "gr8" to "great"). The normalization process reduced prolonged characters by transforming statements from "soooo happy" to "so happy." Duplicate reviews were removed to maintain dataset balance thus eliminating excessive sentiment over-representation.

3.3 | Data Annotation

Users' star ratings were utilized to assign sentiment labels based on established study methodologies [24, 25]. Reviews with 1-star and 2-star ratings were classified as negative sentiment, while those with 4-star and 5-star ratings were labeled as positive sentiment. Reviews with 3-star ratings were categorized as neutral, reflecting a mixed or balanced sentiment. Additionally, reviews with a 0-star rating were marked as unknown and excluded from analysis due to insufficient sentiment clarity and would be later used for prediction. This labeling process resulted in four primary sentiment categories: positive, negative, neutral, and unknown. The distribution of reviews across these sentiment categories is presented in Table 1 below.

3.4 | Class Balancing

The review dataset contained an extreme distribution of classes since positive sentiments outnumbered negative ones and neutral ones, which created difficulties in precise sentiment analysis. To address this, the training data was taken through Synthetic Minority Over-sampling Technique to create a balanced dataset that would minimize prediction biases and improve classifier performance across all classes. The class distribution was checked to verify higher positive review counts before generating synthetic samples for the minority classes (negative and neutral) until all three sentiment categories reached an equal number of 52 247 reviews each. Unlike random under-sampling, SMOTE avoids discarding valuable information from the majority class and instead interpolates new data points from existing minority class samples. This approach preserved the richness of the original dataset while mitigating bias toward the positive class. The balance achieved enabled more reliable sentiment classification and prevented the model from overfocusing on the dominant class.

TABLE 1 | Dataset distribution by sentiment.

Sentiment	Number of reviews
Positive	52 247
Negative	6183
Neutral	2179
Unknown	1151

3.5 | Data Transformation and Feature Extraction

To perform data transformation and feature extraction, term frequency-inverse document frequency (TF-IDF) vectorization was employed to convert the textual reviews into a manageable form for machine learning. TF-IDF is a commonly used method for translating textual data into numerical features by measuring the importance of the word in the document (the review) and throughout the rest of the dataset [26]. The TF-IDF score is calculated as follows:

$$\mathrm{TF-IDF}(t,d) = \mathrm{TF}(t,d) * \mathrm{IDF}(t)$$

- Term Frequency (TF): It measures the frequency of a term appearing in the review Hu et al. [27].

$$\mathrm{TF}(t,d) = \frac{f_{t,d}}{n_d}$$

  $f_{t,d}$: The number of times the term t appears in document d.

  n_d: The total number of terms in document d.

- Inverse Document Frequency (IDF): Scores how rare the term is across the whole dataset, gives more weight to less standard terms, and may thus contain more sentiment value Zhang et al. [28].

$$\mathrm{IDF}(t) = \log\left(\frac{N}{1+n_t}\right)$$

  N: The total number of documents in the corpus.

  n_t: The number of documents containing the term t.

  Adding 1 in the denominator avoids division by zero for terms not present in any document.

The vectorization technique gave a sparse matrix in which each review was represented as a set of numeric values related to the TF-IDF scores of its words. The words with high TF-IDF scores impacted the sentiment classification the most.

3.6 | Aspect-Based Analysis

Holistic sentiment analysis often obscures specific user concerns in mental health chatbot reviews, limiting actionable developer

insights. To address this, an aspect-based analysis was developed to categorize feedback into four dimensions: Reliability (e.g., app stability), Content (e.g., therapy quality), Usability (e.g., interface design), and Emotion (e.g., emotional support). A keyword dictionary was crafted by analyzing high-weighted TF-IDF terms from the 52 247-review dataset, identifying domain-relevant patterns (e.g., "crash," "bad" for Reliability, "helpful," "calm" for Emotion). Terms were iteratively curated based on their TF-IDF prominence and contextual relevance to chatbot functionality, ensuring robust aspect mappings. Unlike overall sentiment approaches, this method disaggregates feedback into granular categories, enhancing precision. The analysis utilized SMOTE-balanced data to equitably represent positive, negative, and neutral sentiments, aligning with the multi-class framework. This tailored, data-driven approach provides clear insights into user experiences, enabling targeted chatbot improvements without relying on complex embeddings.

3.7 | Data Splitting

To ensure that the model had enough data to learn from but also had some held out for independent evaluation, the dataset was split into 80% training and 20% testing. The SMOTE-balanced training set ensured equal class representation, supporting unbiased multi-class classification.

3.8 | Machine Learning Models

Six models were used for multi-class sentiment classification (positive, negative, neutral), leveraging SMOTE-balanced data and an 80–20 train-test split with 10-fold cross-validation. Five traditional models used TF-IDF vectorized reviews, while BERT processed raw text, enhancing robustness for complex sentiments:

- Random Forest (RF) is an ensemble learning method that utilizes multiple decision trees to enhance classification accuracy. Random Forest works exceptionally well on high-dimensional data, i.e., data with more features like text [29].

$$\hat{y} = \text{Mode}\left(T_1(x), T_2(x), \ldots, T_k(x)\right)$$

Where:

$T_i(x)$: Prediction of the i-th decision tree for input x.

k: Total number of trees in the forest.

- Stochastic Gradient Descent (SGD) is a powerful algorithm that can handle large-scale and sparse datasets such as those from TF-IDF [30].

$$J(\theta) = \frac{-1}{m}\sum_{i=1}^{m}\left[y_i \log\left(h_\theta\left(x_i\right)\right) + \left(1 - y_i\right)\log\left(1 - h_\theta\left(x_i\right)\right)\right]$$

Where:

$$h_\theta\left(x_i\right) = \frac{1}{1 + e^{\theta^T x_i}}\text{(sigmoid function)}$$

m: Number of samples.

$_{Yi}$: Actual label of i-th sample.

- Multinomial Naive Bayes (MNB): One of the best-known probabilistic classifiers, this is highly effective for text classification tasks because it assumes independence of features (words) given the class [31].

$$\hat{y} = \arg max_k P\left(C_k\right) j\prod_{\text{j=1}}^{n} P\left(x_j | C_k\right)$$

Where:

$P\left(C_k\right)$: Prior probability of class C_k.

$P\left(x_j \mid C_k\right)$: Probability of word x_j given class C_k, often computed with Laplace smoothing.

- Logistic Regression (LR): A classifier for binary classification problems that is simple yet effective. Logistic Regression predicts the probability that a review is in a positive or negative class [32].

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

Where:

$\theta^T x$: Weighted sum of input features.

- Support Vector Machine (SVM): SVM has been designed to separate the most suitable hyperplane between the positive and the antagonistic classes in high-dimensional data [33].

$$f(x) = sign\left(w^T x + b\right)$$

Where:

w: Weight vector.

b: Bias term.

- BERT (Bidirectional Encoder Representations from Transformers): leverages pre-trained transformer layers to capture contextual word relationships, ideal for raw text reviews [34]. The model was fine-tuned for multi-class classification using the cross-entropy loss:

$$L = \frac{-1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}, log\left(\hat{y}_{i,c}\right)$$

where: (N) is the number of samples, (C) is the number of classes (3: positive, negative, neutral), $y_{i,c}$ is the true label (1 or 0), and $\hat{y}_{i,c}$ is the predicted probability from the softmax output of BERT's [CLS] token.

3.9 | Model Evaluation Metrics

Several key metrics assessing models' performances were used, including accuracy, precision, recall, and F1 score. These metrics provide a comprehensive view of the models' ability to classify sentiment correctly across the dataset:

- Accuracy: It measures how well the model does across all predictions.

$$\text{Acurracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Where:

TP: True Positives (correct positive predictions).

TN: True Negatives (correct negative predictions).

FP: False Positives (incorrect positive predictions).

FN: False Negatives (incorrect negative predictions).

- Precision: How many of the optimistic predictions were correct.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- Recall: The quality it evaluates relates to the ability of the model to identify all positive reviews.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- F1 Score: The harmonic mean is presented as a balanced assessment of the model's performance and is beneficial in the case of imbalanced classes [35].

$$\text{F1 Score } 2.\frac{\text{Precision. Recall}}{\text{Precision} + \text{Recall}}$$

To provide robust evaluation, 10-fold cross-validation during training was used. In this process, the training set was divided into 10 subsets, and the model was trained by nine, tested by the remaining subset, and executed 10 times. This method reduced the overfitting and gave a reasonable estimate of model generalization.

4 | Results

4.1 | Preprocessing and Class Balancing

After completing all data preprocessing steps, the final count was 61 758. Reviews labeled with a rating of 0 were classified as "unknown" and excluded from the training and validation phases but retained for post-model prediction analysis. The initial labeled dataset included 52 247 positive, 6183 negative, and 2179 neutral reviews, highlighting a significant class imbalance that could bias sentiment classification. To address this, the Synthetic Minority Over-sampling Technique was employed to balance the dataset, generating synthetic samples for the minority classes (negative and neutral) until each sentiment class reached 52 247 reviews. This technique yielded a balanced dataset with equal representation of all three sentiment classes, as shown in Table 2.

Unlike random under-sampling, SMOTE preserved the variety of positive reviews, maintaining the detailed feedback from the original extensive dataset. This balanced dataset enhanced the results' broad applicability, supporting robust multi-class sentiment analysis.

TABLE 2 | Sentiment distribution before and after under-sampling.

Sentiment	Before class balancing	After class balancing
Positive	52 247	52 247
Negative	6183	52 247
Neutral	2179	52 247

4.2 | Key Sentiment Drivers Identified Using TF-IDF

TF-IDF analysis was applied to extract the most influential terms within each sentiment category by quantifying word importance in individual reviews relative to the entire corpus. This allowed the identification of dominant patterns and recurring expressions that shaped the sentiment classification across the dataset of 61 758 reviews. The top 20 terms per sentiment class provided insight into user feedback dynamics, as illustrated in Figure 2.

In the combined dataset, the highest-ranking terms were *helpful, love, and talk*. These words appeared frequently in reviews and carried high discriminative value for sentiment classification. Their prominence suggests that emotional benefit, user affection for the app, and conversational ability were consistently mentioned regardless of sentiment. Additional terms such as *thank, recommend, and anxiety* appeared in contexts that reflected both therapeutic value and user needs. Meanwhile, *problem, bad, and work* stood out as negative signals, indicating recurring issues with functionality or perceived performance.

4.2.1 | Key Terms in Positive Sentiments

The most common and meaningful words in positive reviews included *helpful, love, and talk*. These words highlighted how users felt the chatbot as supportive and emotionally affirming. This indicated that the interactions with the chatbots helped users deal with emotions, anxiety, or loneliness. The word *love* most appeared when describing appreciation for the features of the app, the quality of the content, or the general availability of the app when they were going through an emotional struggle. *Talk* appeared as a central theme in positive feedback as a measure of the smooth flow of conversation or the responsiveness of the bot. The users commonly mentioned chatbots as good listeners or a safe place to converse through their issues as an emphasis on the roles of the apps as a reachable companion.

Some of the most significant high-frequency terms that appeared were *thank, recommend, and calm*, which expressed gratitude and a willingness to recommend the app. *Depression, anxiety, and mental* appeared in comments that mentioned the capacity of the chatbot for emotional regulation techniques or reassurance. The occurrence of the words *friend, therapist, happy, and awesome* further supported the emotional connection and therapeutic significance felt by users as shown in Figure 3. The evidence from these findings indicates that positive sentiment is not merely due to the efficacy of mental health assistance but also due to the capacity of the chatbot for the creation of a human-like comforting presence.

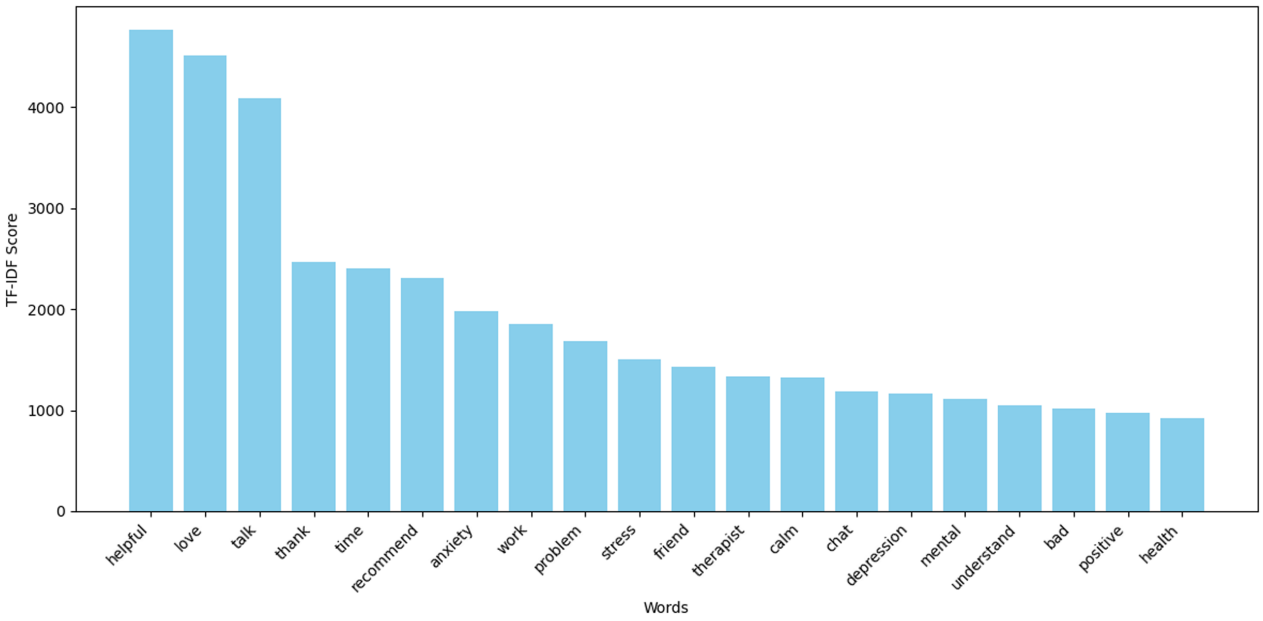

FIGURE 2 | Top 20 TF-IDF terms driving sentiment across all reviews.

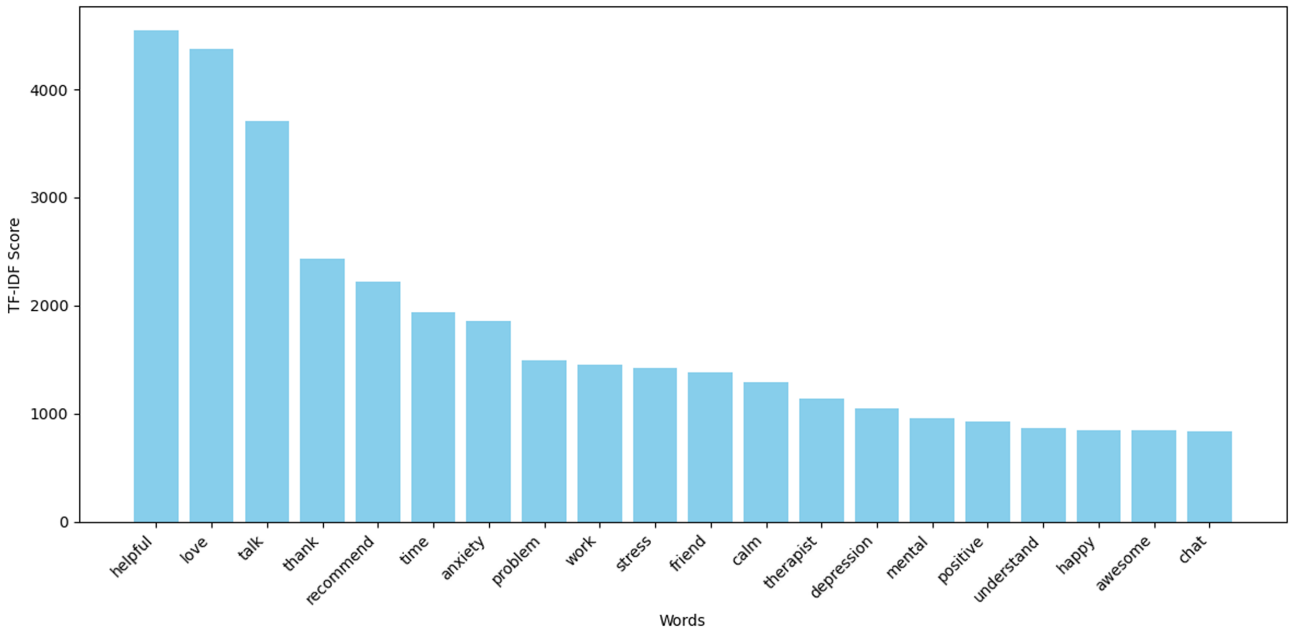

FIGURE 3 | Top 20 TF-IDF terms driving sentiment in positive reviews.

4.2.2 | Key Terms in Negative Sentiments

In contrast, negative reviews were shaped by performance complaints and usability frustrations. The top word, *time, frequently* appeared in reviews expressing delays in responses, long wait times for loading, or sessions ending prematurely. *Bad and work* reflected general dissatisfaction, with users stating the app didn't function as intended or failed to meet expectations. *Chat and talk* were also frequent in negative reviews but with very different context compared to positive one's users often criticized the chatbot's repetition, limited responses, or inability to understand complex input. Account and login issues were identified through terms like account and fix, indicating problems with access or registration.

These functional shortcomings were directly tied to expressions of dissatisfaction. Additionally, *response, mental, and understanding* appeared when users felt the bot failed to comprehend their emotional needs or reply meaningfully. This pattern of high TF-IDF scores around technical vocabulary suggests that much of the negative sentiment stemmed from broken interactions, rather than objection to the app's concept itself *in* Figure 4.

4.2.3 | Key Terms in Neutral Sentiments

Neutral reviews displayed a blend of both functional commentary and tentative evaluations. The most common words used were *talk and work*, frequently included in evaluations that neither

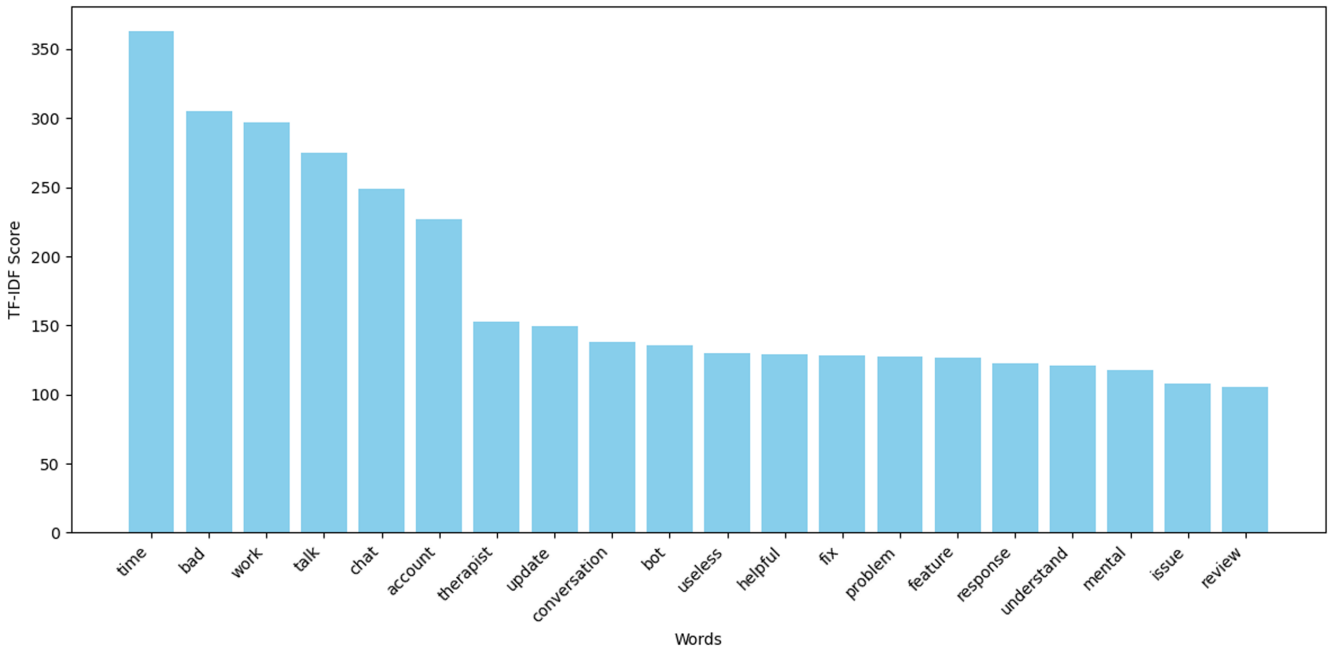

FIGURE 4 | Top 20 TF-IDF terms driving sentiment in negative reviews.

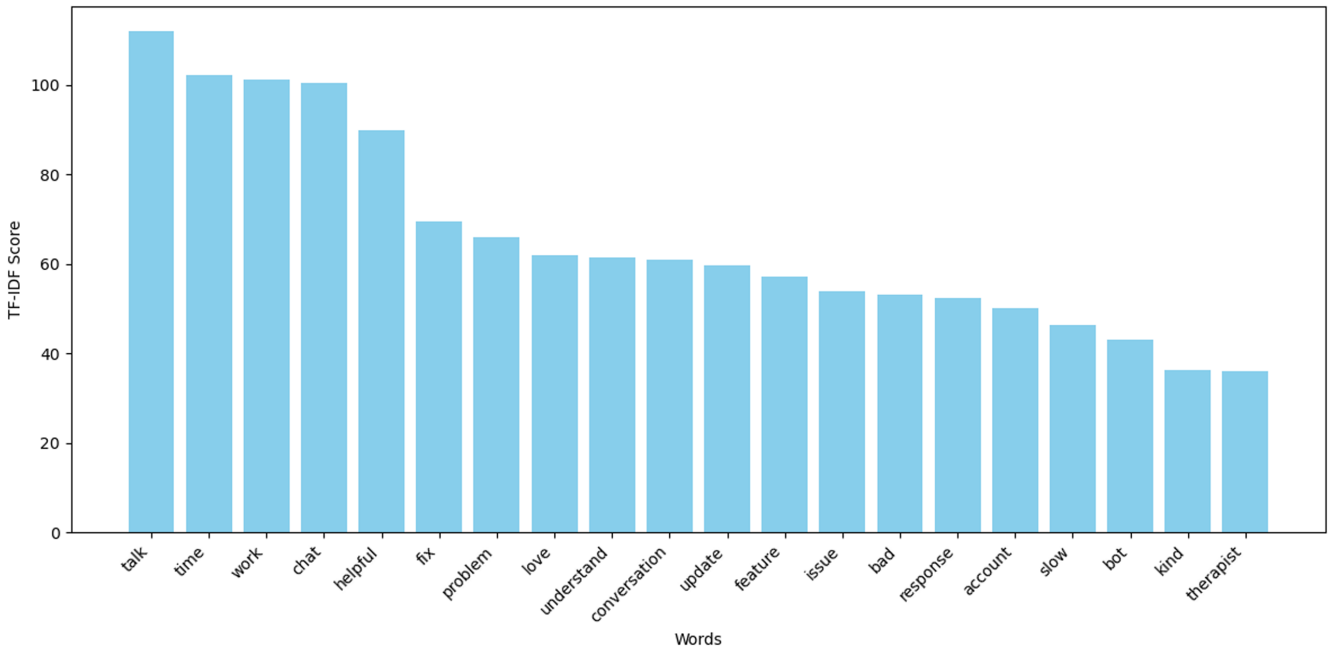

FIGURE 5 | Top 20 TF-IDF terms driving sentiment in neutral reviews.

commended nor criticized the chatbot but explained its overall functioning. For example, users could comment that the chatbot "talks fine" or "works okay," implying doubt or cautious endorsement. *Time and chat* followed as the next most common words used in the evaluations when users explained their experience but provided no clear judgment. Such evaluations accepted the functioning of the chatbot but presented slight issues or even suggestions.

Technical words such as *problem, issue, feature, and update* indicated that users knew there were things that they felt should be improved but not strongly enough to post a negative review. Notably, *helpful and love* appeared, but in weaker forms that suggested that some users liked certain features but remained neutral. The appearance of understand, conversation, and kind pointed toward users testing the extent that the bot was capable of a natural discussion, but not with a high level of emotional investment. This set of reviews represents users that remained neutral, seeing the potential for the chatbot but not feeling it as consistently affecting them Figure 5.

The TF-IDF rankings as per Table 3 above by sentiment presented a fine-grained analysis of the patterns of language that defined user feedback. Positive comments focused on emotional

TABLE 3 | Top terms by TF-IDF score in positive, negative, and neutral reviews.

Positive terms	TF-IDF score	Negative terms	TF-IDF score	Neutral terms	TF-IDF score
Helpful	4535.29	Time	363.83	Talk	112.01
Love	4371.76	Bad	306.66	Work	101.32
Talk	3695.78	Work	296.33	Time	101.23
Recommend	2224.28	Chat	248.91	Chat	100.35
Calm	1292.15	Bug	135.95	Helpful	91.04

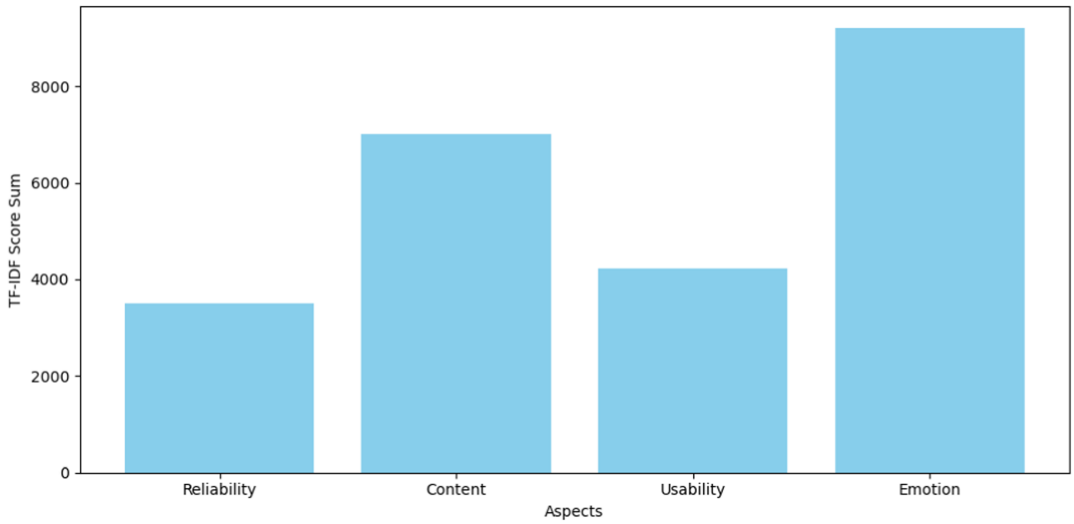

FIGURE 6 | Aspect drivers of sentiment in positive reviews.

support and high levels of satisfaction, negative comments centered around issues with the app and issues with access, and neutral comments expressed wariness, moderate satisfaction, or mild irritation. These findings provide a basis for comprehending how sentiment is embedded in natural user speech and the location of salient patterns within conversations with a chatbot.

4.3 | Aspect-Based Sentiment Analysis

Aspect-based sentiment analysis was used to better understand the nature of the user feedback. A keyword dictionary that is based on data was obtained using TF-IDF term prominence mapping of words under four core aspects: Reliability, Content, Usability, and Emotion. TF-IDF weighted scores were calculated to measure the relative importance of each of the aspects in the three sentiment categories (positive, negative, and neutral).

4.3.1 | Aspect Emphasis in Positive Sentiment

In positive reviews as per Figure 6 below, the largest TF-IDF value was obtained for Emotion with Content following it, then Usability and Reliability. This shows that the users expressed most strongly the emotional value they got from the chatbots using words that represented feeling heard, calm, or support. The high ranking of Content shows that the users valued the therapy exercises, cognitive methods, or supported conversations provided by the chatbots. Meanwhile the terms under Usability mentioned the convenience of interactions, easiness of navigation and accessibility and the terms that appeared under Reliability mentioned consistent behavior or prompt responses. Collectively these aspect scores show that the users with positive opinions were mainly motivated by emotional involvement and backed up with substantial content and a stable user process.

4.3.2 | Aspect Emphasis in Negative Sentiment

Negative comments as per Figure 7 below, were led by Reliability, with Content, Usability, and Emotion following behind. This trend reflects that dissatisfied users' concerns focused most on technical problems such as bugs, crashes, login problems, or frozen systems. Although Content and Usability appeared as well, they were typically mentioned in critical situations like unhelpful dialog, confusing layouts, or malfunctioning features. Words with emotional terms appeared less commonly, implying that emotive language was less prevalent among dissatisfied users, or that emotional needs remained unmet and hence not written out. These scores reflect that functionality and system reliability were the most urgent issues in dissatisfied comments.

4.3.3 | Aspect Emphasis in Neutral Sentiment

Neutral reviews as per Figure 8 below had lower but generally balanced TF-IDF scores for all categories, with Reliability taking

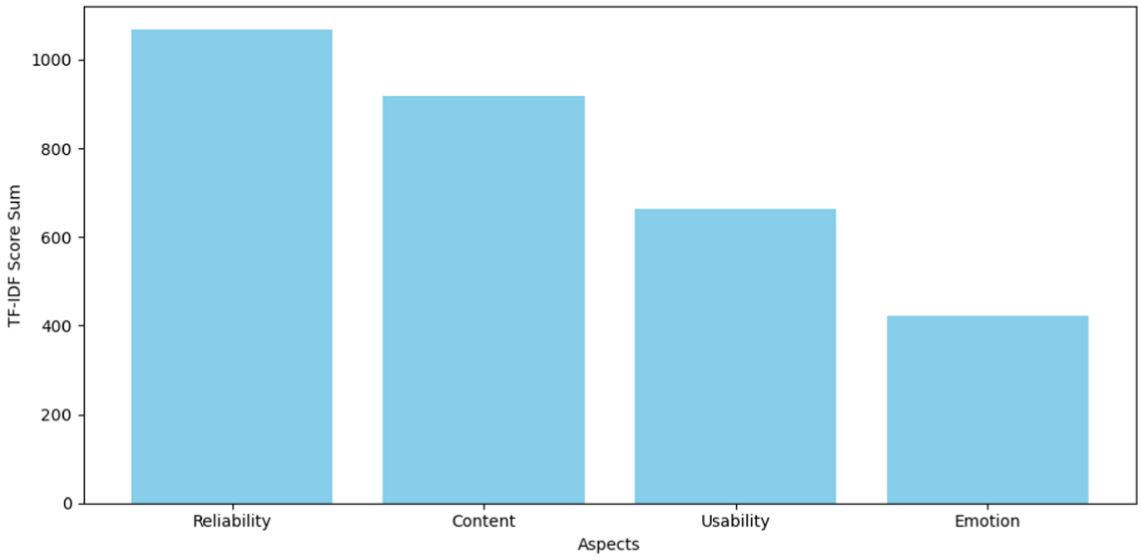

FIGURE 7 | Aspect drivers of sentiment in negative reviews.

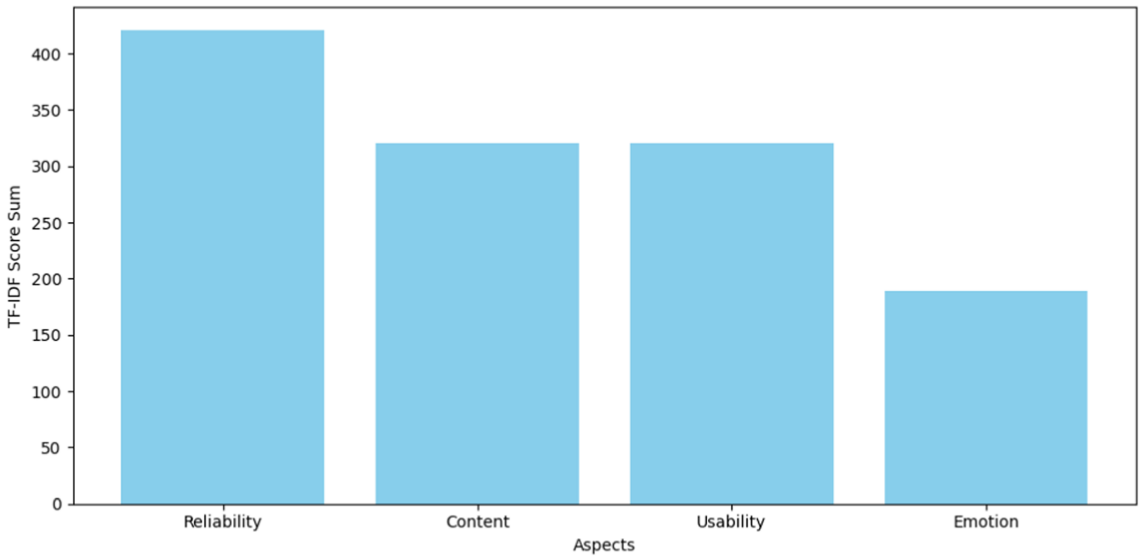

FIGURE 8 | Aspect drivers of sentiment in neutral reviews.

the forefront, then Usability, Content, and Emotion. The pattern indicates that neutral reviewers in general were assessing how the platform performed and looked, with no extreme positive or negative emotions. Most users within this class described the chatbot like "worked okay" or "has good features," mentioning specific features or aspects for improvement. The low emotional rating is an indicator of the neutral review tone since the users used a more cautious approach of describing features or citing minor flaws with little emotional attachment.

The aspect-based TF-IDF results as per Table 4 above helped isolate which chatbot features and user experiences dominated sentiment expressions. While positive reviews were shaped by emotional and content-related language, negative reviews focused on performance and system reliability. Neutral feedback reflected a more measured evaluation across all aspects, with relatively lower intensities.

4.4 | Model Performance Metrics

The performance of six machine learning models Logistic—Regression, Support Vector Machine (SVM), Stochastic Gradient Descent (SGD), Naïve Bayes, Random Forest, and BERT was—evaluated using accuracy, precision, recall, and F1-score across three sentiment classes: positive, negative, and neutral. This comprehensive evaluation was conducted on a test for both the traditional machine learning models and the BERT transformer model, both processed through TF-IDF vectorization (traditional models) or embedding layers (BERT). The results are summarized in Table 5.

BERT emerged as the top-performing model, achieving 99.18% accuracy with perfect or near-perfect precision and recall for all sentiment classes, particularly excelling in detecting minority

TABLE 4 | Aspect TF-IDF scores by sentiment class.

Aspect	Positive score	Negative score	Neutral score
Emotion	9201.99	421.97	187.70
Content	7021.51	917.53	319.12
Usability	4230.20	664.55	320.15
Reliability	3505.87	1066.00	419.40

TABLE 5 | Performance metrics for sentiment classification models (multi-class).

Model	Accuracy	Precision (Neg/Neu/Pos)	Recall (Neg/Neu/Pos)	F1-score (Neg/Neu/Pos)
Logistic regression	88.16%	90%/84%/91%	90%/88%/87%	90%/86%/89%
SVM	90.20%	91%/88%/92%	93%/91%/87%	92%/89%/90%
Stochastic gradient	82.73%	83%/79%/86%	85%/77%/86%	84%/78%/86%
Naïve Bayes	83.79%	87%/77%/87%	83%/83%/85%	85%/80%/86%
Random forest	98.18%	97%/99%/99%	99%/99%/96%	98%/99%/97%
BERT (transformer)	99.18%	95%/93%/100%	100%/100%/99%	98%/96%/100%

class sentiments such as neutral and negative. Random Forest followed closely with 98.18% accuracy and F1-scores exceeding 97% across all classes, reflecting its ability to generalize well on balanced data. Among the traditional classifiers, SVM and Logistic Regression maintained strong performance with 88%–90% accuracy. SVM showed balanced precision and recall above 90% for all three classes, while Logistic Regression exhibited slightly lower recall for the positive class. Naïve Bayes and SGD demonstrated moderate performance, with Naïve Bayes slightly outperforming SGD in neutral sentiment classification. These results confirm the superior contextual sensitivity of transformer models like BERT, especially for nuanced or less-polarized sentiments. They also underscore the value of balancing through SMOTE and multi-class architecture in enhancing model performance across all sentiment types.

4.5 | Unseen Data Predictions

To further assess the generalizability and real-world applicability of the trained classifiers, sentiment predictions were performed on previously unseen user reviews that were excluded from the model training and evaluation phases. These included 1151 reviews that had either a 0-star rating or no explicit sentiment label during preprocessing. These reviews were retained specifically for blind inference, serving as a proxy for deployment conditions where user feedback arrives without labels. The two best-performing models from earlier experiments, Random Forest and BERT, were selected for this task due to their superior performance in classifying labeled data.

The results as per Table 6 above show that both models classified the most unseen reviews as positive, with BERT predicting 988 positive reviews compared to 965 from Random Forest. BERT also predicted slightly more neutral sentiments (23) compared to Random Forest (18), while Random Forest assigned a higher number of negative labels (166 vs. 138). These patterns suggest BERT was more conservative with negative labels and more confident in detecting neutral tones—consistent with its labeled data

TABLE 6 | Sentiment predictions on unlabeled data ($n = 1151$).

Model	Positive	Negative	Neutral
Random forest	965	166	18
BERT	988	138	23

performance. The relatively small number of neutral predictions across both models may reflect the fact that ambiguous or balanced expressions tend to be overshadowed by dominant sentiment signals during vectorization or embedding.

5 | Discussion

This research explored how machine learning and transformer models perform in classifying user sentiment from mental health chatbot reviews to achieve improved accuracy on subtle expressions of emotion. The multi-class methodology coupled with TF-IDF features extraction and aspectual analysis probed deeper into how users interact and gauge mental health chatbots. Comparative model assessment showed that BERT had the best accuracy (99.18%) with a close second-place showing by Random Forest (98.18%) significantly outpacing other conventional classifiers. The good performance by BERT overall and with regards to detecting minority classes of sentiments (neutral and negative) is consistent with research by Liu et al. [36] and Khan et al. [19], who highlighted the capacity of transformer models to capture emotional nuances and context dependencies of text. The high recall and precision in detecting neutral sentiments demonstrate the strength by BERT to recognize ambiguous or confounded emotional states, a long-standing weakness in previous sentiment research conducted using binary schemes of labeling [6, 16].

Random Forest, a classic model, was competitive in all classes and showed the effectiveness of ensemble approaches on high-dimensional, SMOTE-balanced text data. Having a low misclassification rate on neutral and negative classes confirms the

robustness of the model in real-world classification scenarios and when balanced class distributions and structured TF-IDF representations are used.

Conversely, models such as SGD and Naïve Bayes did poorly with neutral reviews, often classifying them as positive sentiments. This is consistent with existing research [21, 37], which has established that linear classifiers as well as general probabilistic models perform poorly when it comes to multi-class emotion detection, and particularly when the emotion is expressed indirectly or subtly.

Aspect-based sentiment analysis also clarified in more detail which features had the strongest impact on user experience. Emotion and content were prominent in posts with positive feedback, whereas in negative feedback, the leading aspect was reliability. This is reflected in previous research by Caldeira et al. [11] and Gkinko & Elbanna [7], who established that system responsiveness and perceived emotional bonding are strongly linked to user satisfaction when using health apps. The low emotion scores on negative and neutral classes imply that chatbot interactions lacking emotional richness or conversational clarity cause users to disengage or become dissatisfied. Prediction on unlabeled data also showed both BERT and Random Forest to classify most reviews as positive and relatively little as neutral. This is line with Albahri et al. [38], who finds that models optimized on binary or even three-class systems tend to overestimate sentiment polarity when faced with ambiguous or low-intensity expressions. The finding that more neutral sentiments were predicted by BERT than by Random Forest supports the findings that contextual modeling facilitates better ambiguity management.

In addition, the TF-IDF outcomes indicate that affect words like "love", "helpful", and "calm" were the most prominent positive feedback, while "bug", "issue", and "login" were the drivers of negative feedback. These patterns of language highlight the twin requirement of users, the chatbot has to work reliably as well as express empathetic, responsive emotion. Wherever technical performance or affectual tone goes awry, the user emotion becomes negative irrespective of the intended purpose of usage. Combined, these results support the value of multi-class sentiment models and sophisticated context models in assessing mental health chatbot systems. The addition of neutral feedback and aspect drivers improves interpretability and prompts developers to optimize not only chatbot content but also system stability and conversational authenticity.

6 | Conclusion and Recommendations

The study focused on how user sentiment can be tracked by applying machine learning and transformer models to mental health chatbot reviews and analyzing the emotional and functional aspects of user engagement. In adopting a multi-class framework and using aspect-level assessment, the study better reflected how users perceive digital mental health tools both based on satisfaction and based on reliability, usability, and responsiveness to their emotions. The insights highlight the significance of building chatbot systems that are both technically reliable and emotionally supportive and can respond to user requirements with empathy, clarity, and consistency.

Future studies could extend this model to track sentiment in specific mental health contexts, especially among users facing stressors like unemployment, financial hardship, or gambling. Researchers would study how real-time tracking of emotions is incorporated into chatbots to facilitate real-time identification of disengagement, dissatisfaction, or distress. There is also scope to investigate further how neutral and uncertain sentiment identification is optimized, which is a weakness in many models. Interdisciplinary studies may also investigate how aspect feedback like consistency or usability is related to long-term user engagement and mental health as per the specific demographic. Following on from here, further research may aid in the creation of more adaptive and context aware, as well as ethically informed, digital mental health programs.

Author Contributions

Ian Igado Mmbayi: conceptualization, writing – original draft, software, data curation, formal analysis, visualization, methodology. **Consolata Gakii:** supervision, writing – review and editing, methodology. **Faith Mueni Musyoka:** writing – review and editing, supervision, validation.

Ethics Statement

No ethics approval was required as this study analyzed anonymized public app reviews.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. World Health Organization, *World Mental Health Report: Transforming Mental Health for All* (WHO, 2022).

2. K. P. Kruzan, E. E. Fitzsimmons-Craft, M. Dobias, J. L. Schleider, and A. Pratap, "Developing, Deploying, and Evaluating Digital Mental Health Interventions in Spaces of Online Help- and Information-Seeking," *Procedia Computer Science* 206 (2022): 6–22, https://doi.org/10.1016/j.procs.2022.09.081.

3. J. A. Naslund, K. A. Aschbrenner, R. Araya, et al., "Digital Technology for Treating and Preventing Mental Disorders in Low-Income and Middle-Income Countries: A Narrative Review of the Literature," *Lancet Psychiatry* 4, no. 6 (2017): 486–500.

4. M. R. Haque and S. Rubya, "An Overview of Chatbot-Based Mobile Mental Health Apps: Insights From App Description and User Reviews," *JMIR Publications* 11 (2023): e44838–e44838, https://doi.org/10.2196/44838.

5. A. Abd-Alrazaq, A. Rababeh, M. Alajlani, B. M. Bewick, and M. Househ, "Effectiveness and Safety of Using Chatbots to Improve Mental Health: Systematic Review and Meta-Analysis," *Journal of Medical Internet Research* 22, no. 7 (2020): e16021, https://doi.org/10.2196/16021.

6. A. Ahmed, S. Aziz, M. Khalifa, et al., "Thematic Analysis on User Reviews for Depression and Anxiety Chatbot Apps: Machine Learning Approach," *JMIR Formative Research* 6, no. 3 (2022): e27654, https://doi.org/10.2196/27654.

7. L. Gkinko and A. Elbanna, "Hope, Tolerance and Empathy: Employees' Emotions When Using an AI-Enabled Chatbot in a Digitalised Workplace," *Information Technology & People* 35, no. 6 (2022): 1714–1743.

8. K. Mehmood, D. Essam, K. Shafi, and M. K. Malik, "Sentiment Analysis for a Resource Poor Language—Roman Urdu," *ACM Transactions on Asian and Low-Resource Language Information Processing* 19, no. 1 (2019): 1–15.

9. A. Følstad and P. B. Brandtzæg, "Users' Experiences With Chatbots: Findings From a Questionnaire Study," *Quality and User Experience* 5, no. 1 (2020): 3, https://doi.org/10.1007/s41233-020-00033-2.

10. Z. A. Khan, Y. Xia, A. Khan, et al., "Developing Lexicons for Enhanced Sentiment Analysis in Software Engineering: An Innovative Multilingual Approach for Social Media Reviews," *Computers, Materials & Continua* 79, no. 5 (2024): 2771–2793.

11. C. Caldeira, Y. Chen, L. Chan, V. H. B. Pham, Y. Chen, and K. Zheng, "Mobile Apps for Mood Tracking: An Analysis of Features and User Reviews," *National Institutes of Health* 2017 (2017): 495–504.

12. B. Shickel, M. Heesacker, S. A. Benton, A. Ebadi, P. Nickerson, and P. Rashidi, *Self-Reflective Sentiment Analysis* (Association for Computational Linguistics (ACL), 2016), https://doi.org/10.18653/v1/w16-0303.

13. A. Yates, A. Cohan, and N. Goharian, *Depression and Self-Harm Risk Assessment in Online Forums* (Cornell University, 2017), https://doi.org/10.18653/v1/d17-1322.

14. B. Saju, S. Jose, and A. Antony, *Comprehensive Study on Sentiment Analysis: Types, Approaches, Recent Applications, Tools and APIs* (IEEE, 2020), https://doi.org/10.1109/accthpa49271.2020.9213209.

15. F. N. Ribeiro, M. Araújo, P. Gonçalves, M. A. Gonçalves, and F. Benevenuto, "SentiBench—a Benchmark Comparison of State-Of-The-Practice Sentiment Analysis Methods," *EPJ Data Science* 5, no. 1 (2016): 1–29, https://doi.org/10.1140/epjds/s13688-016-0085-1.

16. S. Wen, "Detecting Depression From Tweets With Neural Language Processing," *IOP Publishing* 1792, no. 1 (2021): 012058, https://doi.org/10.1088/1742-6596/1792/1/012058.

17. M. Menon and B. George, "Social Media Use for Patient Empowerment in the Gulf Cooperation Council Region," *Elsevier BV* 1, no. 1 (2018): 21–27, https://doi.org/10.1016/j.ceh.2018.10.002.

18. A. Yadollahi, A. G. Shahraki, and O. R. Zaïane, "Current State of Text Sentiment Analysis From Opinion to Emotion Mining," *Association for Computing Machinery* 50, no. 2 (2017): 1–33, https://doi.org/10.1145/3057270.

19. Z. A. Khan, Y. Xia, K. Aurangzeb, et al., "Emotion Detection From Handwriting and Drawing Samples Using an Attention-Based Transformer Model," *PeerJ Computer Science* 10 (2024): e1887.

20. F. Azam, M. T. Agro, M. Sami, M. H. Abro, and A. Dewani, *Identifying Depression Among Twitter Users Using Sentiment Analysis* (IEEE, 2021), https://doi.org/10.1109/icai52203.2021.9445271.

21. D. Tang, B. Qin, and T. Liu, "Deep Learning for Sentiment Analysis: Successful Approaches and Future Challenges," *WIREs Data Mining and Knowledge Discovery* 5, no. 6 (2015): 292–303, https://doi.org/10.1002/widm.1171.

22. Heedzy, Download App Reviews From iTunes App Store & Google Play. 2016, https://heedzy.com/.

23. R. Patel and K. Passi, "Sentiment Analysis on Twitter Data of World Cup Soccer Tournament Using Machine Learning," *IoT* 1, no. 2 (2020): 14.

24. J. Gebauer, Y. Tang, and C. Baimai, "User Requirements of Mobile Technology: Results From a Content Analysis of User Reviews," *Information Systems and e-Business Management* 6 (2008): 361–384.

25. S. McIlroy, N. Ali, H. Khalid, and A. E. Hassan, "Analyzing and Automatically Labelling the Types of User Issues That Are Raised in Mobile App Reviews," *Empirical Software Engineering* 21 (2016): 1067–1106.

26. M. Bounabi, K. El Moutaouakil, and K. Satori, "Text Classification Using Fuzzy TF-IDF and Machine Learning Models," in *Proceedings of the 4th International Conference on Big Data and Internet of Things* (Association for Computing Machinery, 2019), 1–6.

27. K. Hu, H. Wu, K. Qi, et al., "A Domain Keyword Analysis Approach Extending Term Frequency-Keyword Active Index With Google Word2Vec Model," *Scientometrics* 114 (2018): 1031–1068.

28. W. Zhang, T. Yoshida, and X. Tang, "A Comparative Study of TF* IDF, LSI and Multi-Words for Text Classification," *Expert Systems With Applications* 38, no. 3 (2011): 2758–2765.

29. L. Breiman, "Random Forests," *Machine Learning* 45 (2001): 5–32.

30. S. Santhosh Baboo and M. Amirthapriya, "Comparison of Machine Learning Techniques on Twitter Emotions Classification," *SN Computer Science* 3, no. 1 (2022): 35.

31. T. Eriksson, *Automatic Web Page Categorizationusing Text Classication Methods* (Linnaeus University, 2013).

32. J. Phillips, E. Cripps, J. W. Lau, and M. R. Hodkiewicz, "Classifying Machinery Condition Using Oil Samples and Binary Logistic Regression," *Mechanical Systems and Signal Processing* 60 (2015): 316–325.

33. B. Liu and L. Zhang, "A Survey of Opinion Mining and Sentiment Analysis," in *Mining Text Data* (Springer, 2012), 415–463.

34. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding," in *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)* (Association for Computational Linguistics (ACL), 2019), 4171–4186.

35. T. B. Alakus and I. Turkoglu, "Comparison of Deep Learning Approaches to Predict COVID-19 Infection," *Chaos, Solitons & Fractals* 140 (2020): 110120.

36. A. R. Liu, P. Pataranutaporn, S. Turkle, and P. Maes, "Chatbot Companionship: A Mixed-Methods Study of Companion Chatbot Usage Patterns and Their Relationship to Loneliness in Active Users," *arXiv* (2024), preprint arXiv:2410.21596.

37. N. Y. Elamin, "Sentiment Analysis With Supervised Learning Techniques: A Survey," *Indian Society for Education and Environment* 13, no. 3 (2020): 249–268, https://doi.org/10.17485/ijst/2020/v13i03/148900.

38. A. S. Albahri, A. M. Duhaim, M. A. Fadhel, et al., "A Systematic Review of Trustworthy and Explainable Artificial Intelligence in Healthcare: Assessment of Quality, Bias Risk, and Data Fusion," *Information Fusion* 96 (2023): 156–191.