University of Hull, UK.

Sentiment Analysis of User Reviews to Evaluate Usability in Mobile Health Applications Using Natural Language Processing

Name: Blessing Bolatito Alakunle

Student ID: 202372187

Supervisor: Tareq AI Jaber

Second Examiner: Zekun Guo

Prepared as part of the requirements for the award of the MSc in Data Science and Artificial Intelligence under the Faculty of Science and Engineering.

January, 2025.

Dedication

This work is dedicated to God, whose infinite grace, unconditional love, and unwavering guidance carried me through every step of this journey. To my parents, Mr. and Mrs. Alakunle, and my siblings, whose steadfast love, encouragement, and prayers have been my foundation and greatest support. To my beloved partner, Daniel Olatunbosun, for his unwavering belief in me and for standing by my side through every challenge, and to my close friend, Adeola Ogunleye, whose encouragement and friendship have been a constant source of strength.

To Olamide Fashina, Chika Aghaulor, and my other esteemed colleagues, whose collaboration and camaraderie have enriched this experience and made it truly memorable. To the Sulaimon's, Onyeukwus and Omasirim Emmanuel for their profound support throughout my academic journey. I am deeply grateful, and the list of those who have supported me is truly endless.

To my supervisor, Tareq Al Jaber, and the dedicated faculty members of the Department of Data Science and Artificial Intelligence at the University of Hull, whose passion for teaching and commitment to excellence have inspired and motivated me throughout this academic journey. Lastly, this work is dedicated to myself—for the resilience, determination, and hard work that brought me to this milestone.

Abstract

The usability of mobile health (mHealth) applications is vital for their effectiveness in improving healthcare delivery, yet poor usability often leads to user frustration and app abandonment. With the growing reliance on mHealth apps for chronic disease management, fitness tracking, and general healthcare monitoring, ensuring their user-friendliness has become critical. Despite advancements in app development, many users face challenges related to navigation, technical glitches, and interface design, which hinder the potential benefits these apps can offer. Addressing these usability concerns is essential for maximizing user engagement and achieving better health outcomes. This study evaluates the usability of mHealth applications through sentiment analysis of user reviews, utilizing traditional machine learning models such as Random Forest and Naive Bayes, and advanced deep learning techniques like Long Short-Term Memory (LSTM).

A pre-processed dataset of 111,989 user reviews was analysed. Results show that LSTM outperformed other models, achieving an accuracy and F1-score of 97%, while Random Forest followed with 93% accuracy. Feature importance analysis highlighted keywords such as “easy”, “great”, and “love”, reflecting positive user sentiments toward ease of use and satisfaction. Negative sentiments were often linked to “navigation issues” and “error handling”, revealing key usability challenges. Usability themes—Ease of Use, Navigation Issues, and Error Handling—were extracted through natural language processing techniques.

Furthermore, sentiment trends over time demonstrated significant improvements in user satisfaction following app updates that prioritized navigation fixes and interface enhancements. Analysis across different app versions showed earlier versions faced higher negative sentiments due to technical issues, whereas later versions reflected improved usability and positive feedback.

This research underscores the importance of leveraging sentiment analysis for usability evaluation, providing developers with actionable insights to enhance app design, address usability challenges, and ultimately improve user satisfaction and healthcare outcomes. Future work can explore advanced transformer models and multilingual analysis to further enrich these findings.

Table of Contents

Chapter 1: Introduction & Background

1.1 Introduction to the topic

The introduction of mobile health (mHealth) applications has brought a remarkable shift in healthcare delivery by providing users with innovative tools for monitoring their health, managing chronic conditions, and accessing medical services conveniently through mobile devices like smartphones. This technological advancement has unlocked opportunities to enhance patient involvement in their care, offer personalized health solutions, and reduce disparities in healthcare access (Ventola, 2014). Studies show that mHealth tools are particularly effective in chronic disease management by enabling continuous monitoring and fostering patient-centred healthcare approaches (Bashshur et al., 2011).

Despite their potential, the success and widespread use of mHealth applications heavily depend on their usability. Poorly designed apps, characterized by overly complex interfaces, confusing layouts, and inadequate user experiences, often discourage users and reduce the effectiveness of these tools in promoting better health outcomes (Nielsen, 1994). For example, a systematic review highlights that many users abandon mHealth applications due to difficulties in navigating the app, lack of user-friendly design, and privacy concerns (Free et al., 2013). Therefore, improving usability through rigorous evaluation is essential to meet user expectations and enhance the effectiveness of mHealth applications in achieving better health outcomes.

1.2 Relevance of the study: Usability Challenges in mHealth Applications

Usability plays a crucial role in determining how effective mobile health (mHealth) applications are in achieving their intended goals. As defined by Nielsen (1994), usability refers to the ease with which users can interact with an application, the simplicity of learning its features, and the overall satisfaction derived from using it. When applications lack usability, users may experience frustration, disengagement, and, in many cases, abandon the app altogether (Ventola, 2014). Research indicates that common issues faced by users include poorly designed navigation paths, overly complicated interfaces, and insufficient accessibility features (Free et al., 2013). To overcome these obstacles, it is vital to conduct comprehensive evaluations of user interactions and actively incorporate their feedback into the app design and development process.

1.3 Critical Review: Sentiment Analysis as a Tool for Usability Evaluation

Sentiment analysis, a technique within natural language processing (NLP), focuses on the computational analysis of opinions, emotions, and sentiments in textual data. This method has gained prominence as a key tool for assessing user feedback, especially in fields like mobile health (mHealth) applications (Cambria et al., 2018). Sentiment analysis examines user-generated content to uncover emotions and opinions within text, offering insights into user perceptions of applications. Positive sentiments often signal satisfaction with features, while negative sentiments can reveal challenges such as poor usability or technical problems (Ruelens, 2021).

Earlier research has explored various frameworks and methodologies for usability evaluation in mHealth applications. For instance, Wang et al. (2022) conducted a scoping review on the usability evaluation of mHealth apps for elderly individuals, highlighting the need for age-friendly interfaces and appropriate evaluation methods. Another study by Petersen and Hempler (2017) developed and tested a mobile application to support diabetes self-management for people with newly diagnosed type 2 diabetes, emphasizing the importance of user involvement in the development. While insightful, their findings were specific to a single health condition, limiting broader applicability. To address such limitations, Harrington et al. (2017) proposed a method for developing universally accessible mHealth apps for older adults, focusing on inclusive design principles to enhance user experience across diverse populations

Building on these contributions, this study conducts a detailed sentiment analysis of user reviews from a variety of mHealth applications. By identifying recurring usability themes and analysing user sentiment trends, this research aims to provide actionable insights that can guide the development of more user-friendly mHealth solutions.

1.4 Research Objectives

The primary objectives of this research are to:

1. To classify user reviews into positive and non-positive sentiments using machine learning techniques, focusing on usability-related feedback for mHealth applications.
2. To identify key features and usability issues that significantly influence sentiment classification, through feature importance analysis.
3. To extract usability themes from user reviews using natural language processing (NLP) techniques.
4. To analyse sentiment trends over time and across app versions, assessing updates' impact on user satisfaction and usability.

1.5 Research Questions

1. How can user reviews of mobile health (mHealth) applications be effectively classified into positive and negative sentiments using machine learning techniques, and how do the selected models perform in terms of accuracy, precision, recall, and F1-score?
2. Which review features significantly impact sentiment classification, and what do these features reveal about usability issues?
3. What usability-related themes can be extracted from user reviews using natural language processing (NLP) techniques?
4. How do user sentiments towards mHealth applications evolve over time and across different app versions, and what actionable recommendations can be derived to improve the usability and user satisfaction of these applications?

1.6 Research Hypothesis

1. **Null Hypothesis (H_0):**
   Usability factors have no significant impact on the sentiments expressed in user reviews of mobile health applications.

2. **Alternative Hypothesis (H_1):**
   Usability factors significantly influence the sentiments expressed in user reviews of mobile health applications.

Chapter 2.0 Methodology

2.1 Dataset Source and Description

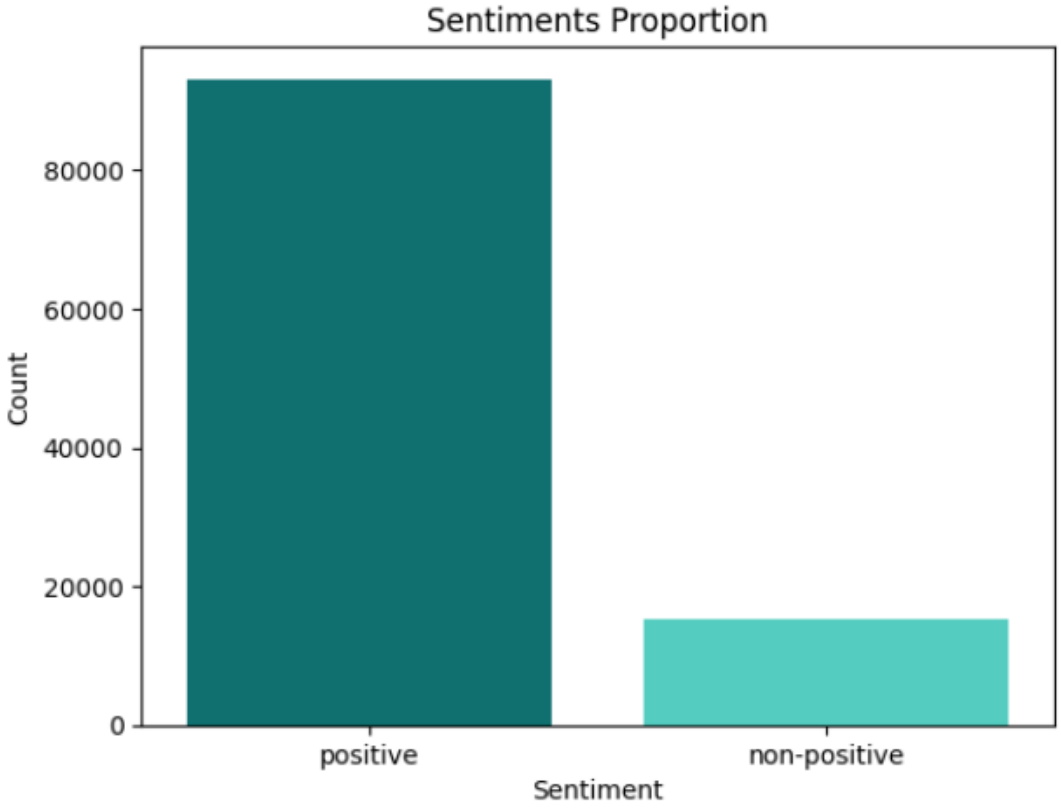
Two datasets were combined, and these were gotten online. The first contains apps which help users with various health tasks, such as tracking their medical data, managing conditions like blood pressure, and communicating with doctors.

The second has **111,989 user reviews** from people who have used these apps. The reviews provide insights into how users feel about the apps, their ease of use, and overall experience. This combined dataset is great for analysing how app features affect user satisfaction. It includes both categorical data (such as app names and user ratings) and numerical data (like the number of installs and average scores), making it perfect for understanding what influences user feedback.

2.2 Dataset Imbalance and Oversampling

This study adopted a binary classification approach, categorizing user reviews into positive and non-positive sentiments (negative and neutral). Binary classification is widely used in sentiment analysis due to its simplicity and ability to provide clear actionable insights. For example, Liu (2017) utilized a binary sentiment classification method to analyse citation texts, effectively distinguishing between positive and negative sentiments to extract meaningful insights.

An initial analysis of the dataset revealed a significant imbalance in sentiment labels, with most reviews classified as positive and a smaller proportion labelled as non-positive. This imbalance, as visualized in the sentiment distribution plot (Figure 1), poses a challenge for training machine learning models. Models trained on imbalanced data tend to overfit to the majority class, resulting in poor predictive performance for the minority class (He & Garcia, 2009). Such bias can distort the insights derived from the analysis and compromise the reliability of the results.



Sentiment Distribution of User Reviews

To address this issue, the **Synthetic Minority Oversampling Technique (SMOTE)** was employed. SMOTE is a widely used resampling method that generates synthetic examples for the minority class by interpolating between existing samples (Chawla et al., 2002). By balancing the sentiment labels, this technique ensured that the model had an equal representation of positive and non-positive reviews during training, improving its ability to generalize across both classes.

2.3 Data Cleaning and Pre-processing

Data cleaning and pre-processing prepare raw data for analysis by removing inaccuracies and restructuring it into a usable format (Nielsen, 1994). Key steps included:

- Dropping Irrelevant Columns: Removed columns irrelevant to analysis or with excessive missing values (Free et al., 2013).
- Handling Missing Values:
    - Categorical Data: Replaced missing values with placeholders for consistency (Ventola, 2014).

- Textual Data: Replaced missing text entries with meaningful placeholders.
- Extensive Missing Data: Filled columns with generic placeholders to ensure dataset completeness (Bashshur et al., 2011).

These steps ensured a clean, consistent dataset ready for analysis.

2.4 Text Preprocessing for Sentiment Analysis

Text preprocessing standardizes data for sentiment analysis, improving data quality and feature extraction (Cambria et al., 2018). Key steps:

- Removing Punctuation and Numbers: Eliminated noise (Ventola, 2014).
- Filtering Stop Words: Removed common, non-sentiment-carrying words.
- Lowercasing Text: Ensured uniformity (Liu, 2017).
- Tokenization and Lemmatization: Split text into words and reduced them to root forms (Cambria et al., 2018).
- Handling Unusual Tokens and Duplicates: Removed irrelevant tokens and repeated records (Harrington et al., 2017).
- Excluding Empty Records: Ensured data reliability by removing records without meaningful content.

These steps prepared the data for accurate sentiment analysis.

Chapter 3.0: Implementation

3.1 Comparing a range of traditional Models for Sentiment Analysis

Traditional machine learning models like Logistic Regression, Naive Bayes, Random Forest, and Gradient Boosting are commonly used for sentiment analysis. Each model offers specific advantages and limitations, and the choice of model often depends on the dataset and project goals. This section evaluates these models and explains the rationale for selecting two of them for this study.

Logistic Regression

Logistic Regression is a widely used baseline model due to its simplicity and ease of interpretation. It performs well on high-dimensional datasets, such as those using TF-IDF or Bag-of-Words feature representations. However, its primary limitation lies in its inability to capture non-linear relationships, which can reduce its performance on complex datasets (Nielsen, 1994).

Naive Bayes

Naive Bayes, particularly the Multinomial variant, is popular for text classification tasks due to its computational efficiency and simplicity. It works well with large-scale textual datasets, especially when feature distributions are relatively straightforward. However, the assumption of feature independence can be a drawback when analysing textual data, where word interactions are often significant (Rennie et al., 2003).

Random Forest

Random Forest is an ensemble learning method that combines multiple decision trees to improve classification accuracy and robustness. It excels at capturing non-linear relationships and is less prone to overfitting compared to simpler models. Additionally, Random Forest provides valuable insights into feature importance, which can aid in understanding the key factors influencing sentiment. However, its higher computational cost and lower interpretability compared to simpler models can be challenging in resource-constrained environments (Breiman, 2001).

Gradient Boosting

Gradient Boosting models, such as XG Boost, are known for their exceptional performance in classification tasks. They handle noisy and large datasets effectively and often outperform other

methods. However, these models are computationally intensive and require careful hyperparameter tuning, making them less practical for projects with limited resources (Free et al., 2013).

For this study, Random Forest and Naive Bayes were chosen based on their respective strengths:

- Random Forest was selected for its ability to handle non-linear relationships and its robustness against noisy data. It is particularly well-suited for complex datasets where interpretability of key features is important.

3.2 Comparing Deep Learning Models for Sentiment Analysis

Deep learning models, including Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), Bidirectional LSTM (Bi LSTM), and Transformer-based models like Distil BERT and BERT, have been extensively utilized in sentiment analysis. Each model presents unique advantages and challenges, with their suitability often determined by dataset characteristics and application requirements. This section examines these models based on their performance in prior studies and explains the rationale for selecting the most appropriate model for this project.

Long Short-Term Memory (LSTM)

LSTM networks are effective at capturing long-range dependencies in sequential data, making them suitable for sentiment analysis tasks involving lengthy texts. A recent study demonstrated the strength of LSTMs in sentiment classification, achieving high accuracy and F1 scores (Kumar et al., 2024). However, LSTMs can be computationally demanding, particularly for large datasets, which may lead to longer training times.

Gated Recurrent Units (GRU)

GRUs simplify the architecture of LSTMs by combining the forget and input gates into a single update gate, reducing computational complexity. While GRUs often achieve competitive accuracy in sentiment analysis tasks, they may not perform as well as LSTMs when handling datasets with long-term dependencies (Chung et al., 2014).

Bidirectional LSTM (Bi LSTM)

Bi LSTMs enhance LSTMs by processing sequences in both forward and backward directions, allowing for a more comprehensive understanding of context. A study on hybrid CNN-LSTM models highlights the effectiveness of Bi LSTMs in sentiment analysis (Wang et al., 2023).

Despite their effectiveness, Bi LSTMs require more computational resources, which can be a limitation in resource-constrained environments.

Distil BERT

Distil BERT, a distilled version of BERT, balances efficiency with accuracy by reducing the model's size while retaining much of its performance. It has shown strong results in sentiment analysis tasks, particularly for noisy and large datasets (Papia et al., 2024). However, its transformer-based architecture still demands significant computational resources.

BERT

BERT, with its bidirectional transformer architecture, sets a high standard for sentiment analysis by capturing complex word relationships and contextual meaning. However, its high computational cost and the need for extensive fine-tuning make it less practical for certain projects (Cambria et al., 2018).

LSTM was selected for this project due to its proven effectiveness in analysing sequential data and its ability to classify usability-related sentiments accurately. Tan et al. (2022) highlighted its robustness in sentiment analysis, making it an ideal choice for the dataset used in this study.

3.3 **Feature Importance Analysis (Using Random Forest)**

Feature importance analysis using Random Forest revealed that the most influential features for sentiment classification include 'easy,' 'love,' and 'great,' with importance scores of 0.046, 0.043, and 0.038, respectively (Table 1). These features align with user sentiment toward ease of use and overall app satisfaction.

Feature Importance and Extracted Usability Themes for Sentiment Classification

Extracted Usability Themes:

	Feature	Importance	Theme
0	easy	0.0461	Ease of Use
1	love	0.0430	User Satisfaction
2	great	0.0385	Ease of Use
3	log	0.0166	Navigation Issues
4	helpful	0.0111	Ease of Use
5	work	0.0110	Error Handling
6	doesn	0.0106	Other
7	convenient	0.0097	Ease of Use
8	password	0.0093	Navigation Issues
9	excellent	0.0092	User Interface
10	app	0.0092	General Usage
11	won	0.0079	General Usage
12	update	0.0073	Navigation Issues
13	sign	0.0072	Navigation Issues
14	error	0.0070	Error Handling
15	good	0.0069	User Satisfaction
16	hard	0.0068	Error Handling

Keyword Analysis and Word Clouds

The word clouds generated for the user reviews (shown in Figure 2) further illustrate the dominant terms influencing sentiment:

Word Cloud Visualization of Positive and Negative Sentiments in mHealth User Reviews

3.4. Extraction of Usability Themes

Fig 7 categorizes the most influential features into usability themes such as ease of use, navigation issues, and error handling. The prevalence of 'ease of use' highlights the importance of user-friendly interfaces, while navigation issues point to areas needing improvement.

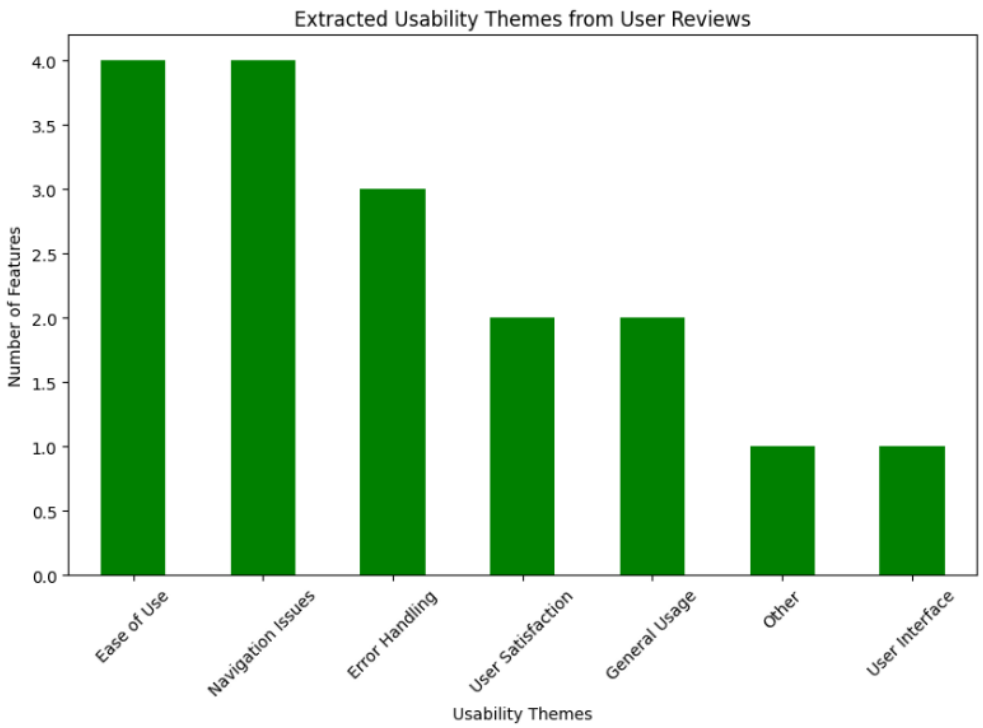


Fig 3: Distribution of Extracted Usability Themes from User Reviews

3.5 Sentiment Evolution Over Time(Figure 4)

The sentiment analysis revealed fluctuations over time. In the early periods, mixed sentiments (76–83% positive) highlighted inconsistencies in user experience, with occasional spikes in negative sentiment due to usability challenges. The middle period saw a decline in positive sentiment and a rise in negative sentiment, likely linked to app updates introducing technical issues. In the later periods, positive sentiment stabilized at 85–87%, while negative sentiment dropped below 15%, reflecting improvements in

features, navigation, and error resolution.

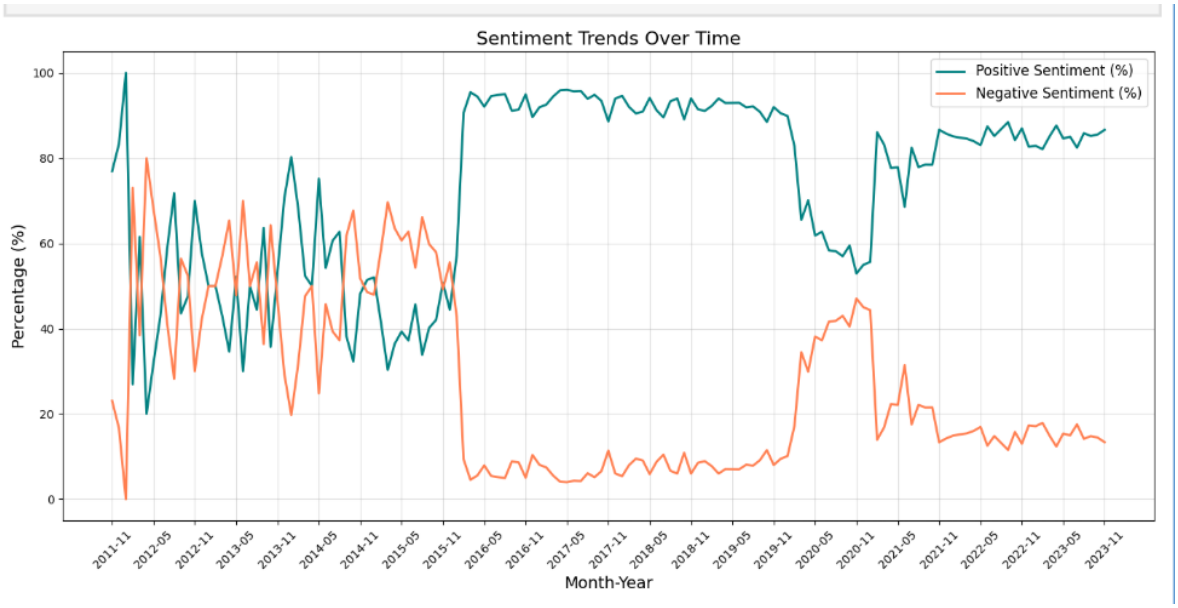


Fig 4: Sentiment Trends Over Time

3.6 Sentiment Distribution Across App Versions

From figure 5, Sentiment analysis showed that early app versions (1 and 2) had high negative sentiment (over 40%), likely due to poor navigation, technical bugs, and unclear design. Later versions (4, 6, 8, and 9) saw over 80% positive sentiment, reflecting improvements in ease of use, navigation, and bug fixes. However, reviews from 'Unknown' versions had mixed sentiments, indicating a need for better version tracking and consistent user experiences.

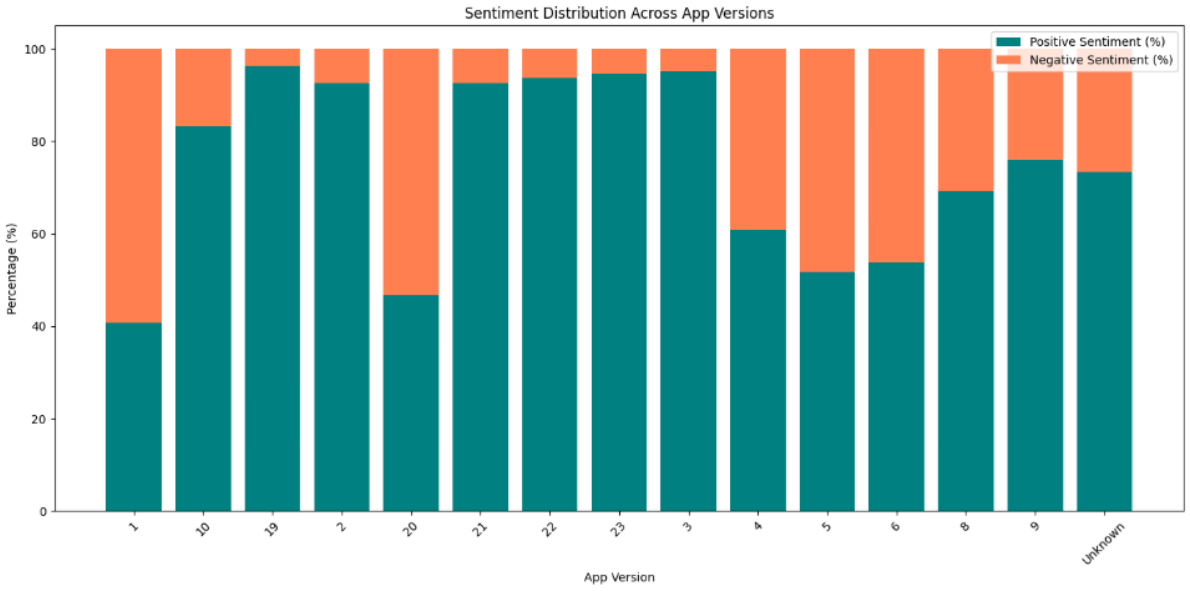


Figure 5: Graph of Sentiment Distribution Across App Versions.

Chapter 4: Results

This section presents the results of sentiment classification models applied to user reviews of mHealth applications. Comparisons are made using key metrics, including accuracy, precision, recall, F1-score, and AUC-ROC, with insights into the usability-related themes derived from user reviews.

4.1 Classification Results of Machine Learning Models

The performance of traditional machine learning models, Multinomial Naive Bayes and Random Forest is summarized in Table 2 and Figure 6.

Table 2: Model Performance Comparison.

Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score
Multinomial Naive Bayes	0.92	0.90	0.91	0.90
Random Forest	0.93	0.92	0.91	0.92

Random Forest achieved the best results with an accuracy of 93%, precision of 92%, and F1-score of 92%, surpassing Naive Bayes across all evaluation metrics. These findings are consistent with earlier research, such as Ventola (2014), which emphasized Random Forest's effectiveness in managing complex textual datasets.

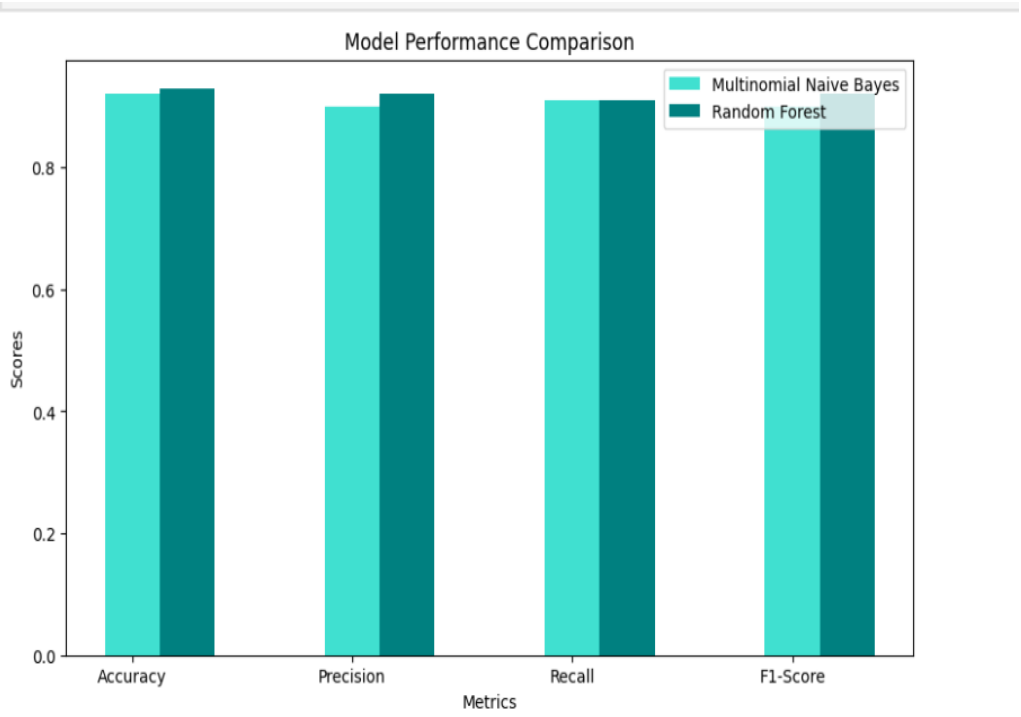


Figure 6: Graph illustrating the Model Comparison.

Results of Deep Learning Model

LSTM Model Evaluation

Table 3: Classification Report for LSTM

```
Classification Report:
              precision    recall  f1-score   support

Non-Positive       0.97      0.98      0.97     18730
    Positive       0.98      0.97      0.97     18531

    accuracy                           0.97     37261
   macro avg       0.97      0.97      0.97     37261
weighted avg       0.97      0.97      0.97     37261

1165/1165 ━━━━━━━━━━━━━━━━━━━━ 32s 28ms/step
ROC-AUC Score: 0.99
```

The LSTM model achieved a test accuracy of 97.46% with a training-validation accuracy difference of less than 0.5% (Table 3).

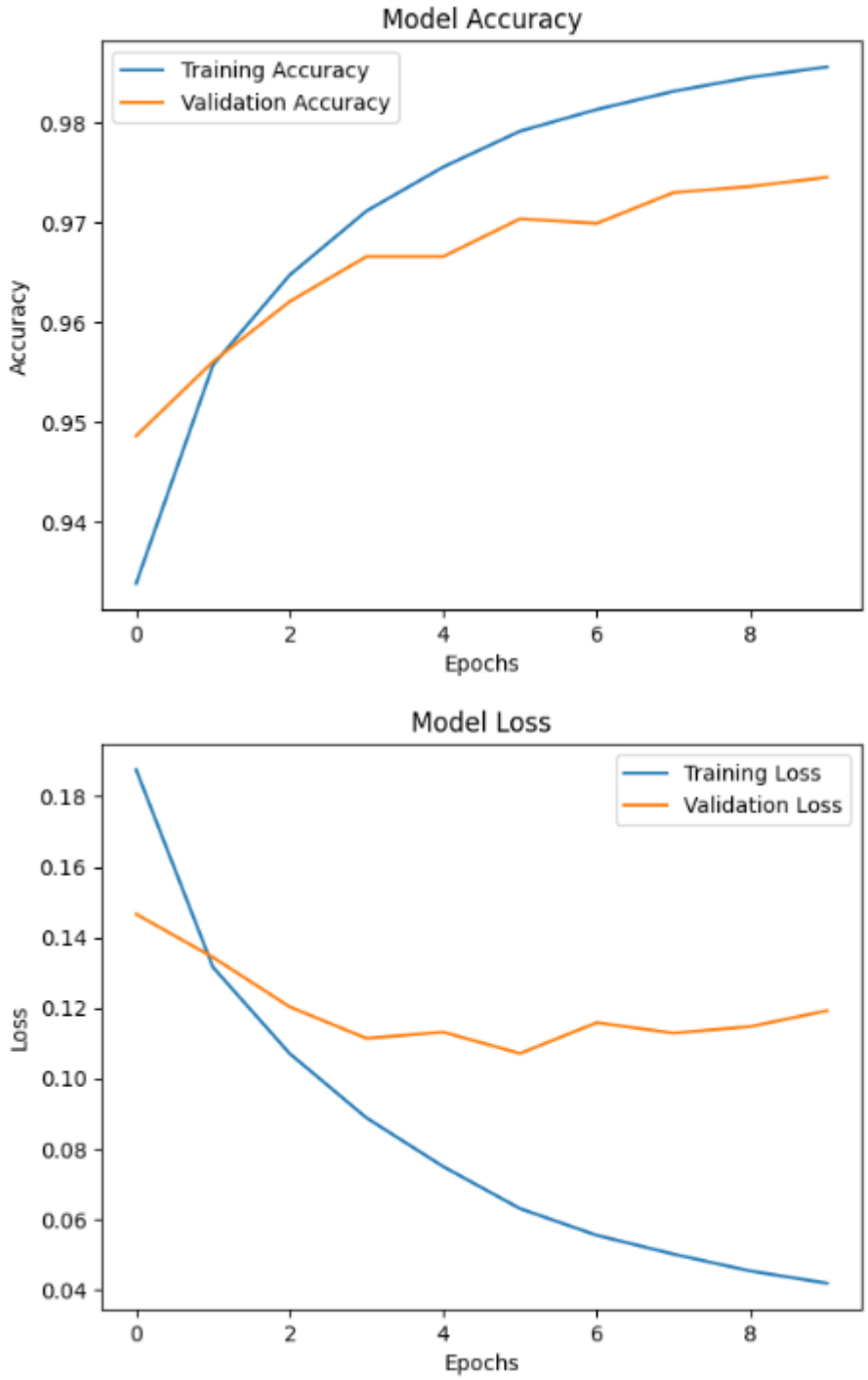


Fig 7: LSTM Model Accuracy and Loss

The confusion matrix (Figure 8) reveals a balanced classification across both sentiment classes, with a high true positive rate for positive reviews. Furthermore, the ROC-AUC (Figure 9) score of 0.99 confirms the model's exceptional discriminatory capability.

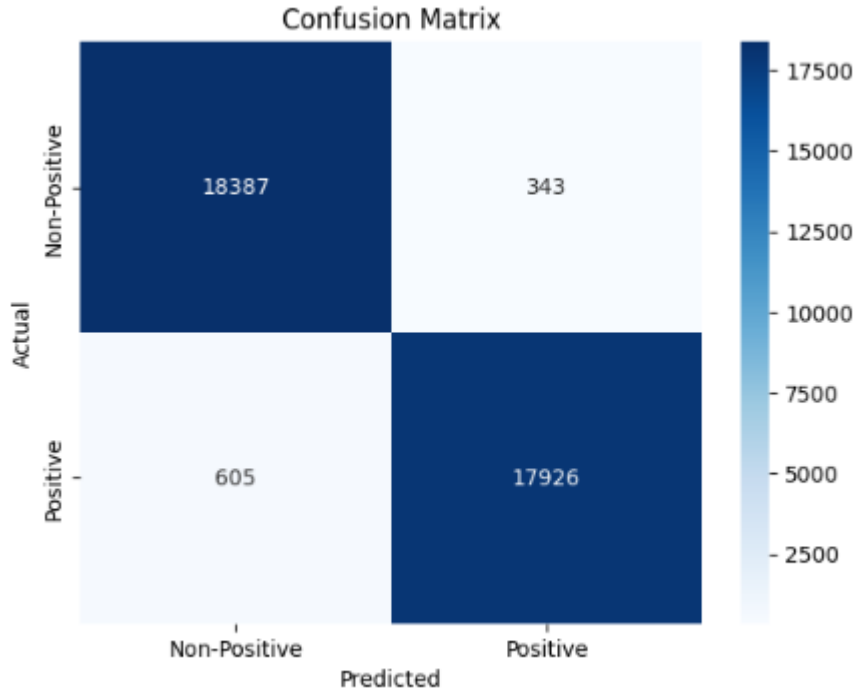


Figure 8: LSTM Confusion Matrix

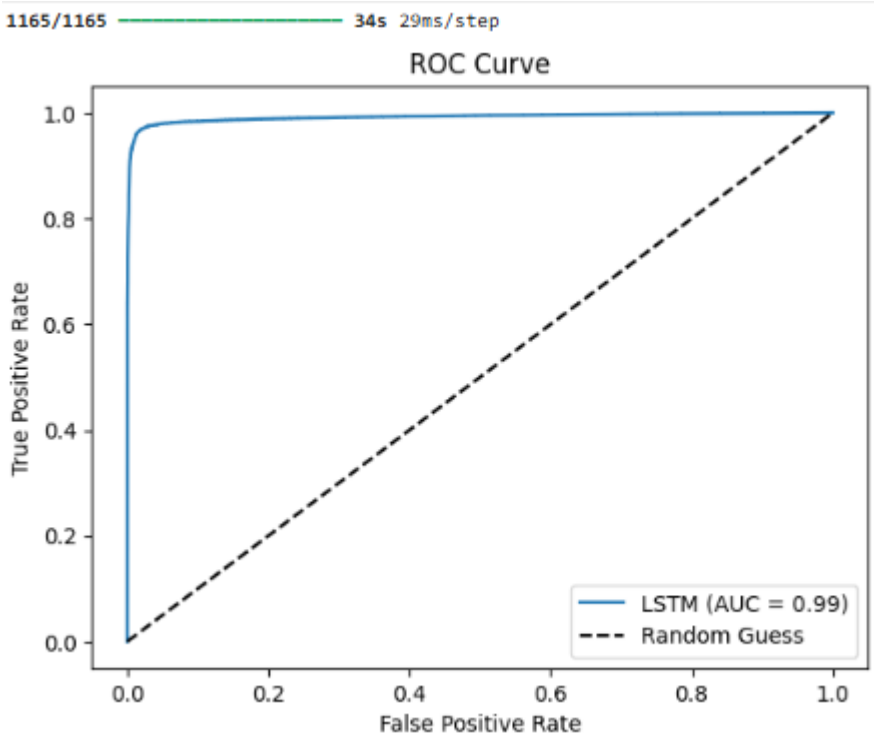


Fig 9: ROC Curve for LSTM

In all, LSTM outperformed other models, achieving the highest accuracy (97%) and F1-Score (97%), due to its ability to capture sequential relationships in textual data. Random Forest followed closely, providing strong performance with an accuracy of 93%.

Table 4: Model Comparison across all metrics.

	Model	Accuracy	Precision	Recall	F1-Score
0	Multinomial Naive Bayes	0.92	0.90	0.91	0.90
1	Random Forest	0.93	0.92	0.91	0.92
2	LSTM	0.97	0.98	0.97	0.97

These results highlight the effectiveness of deep learning models like LSTM for sentiment classification while validating the robustness of traditional ensemble methods like Random Forest. This is shown in table 4.

4.2 Extracted Usability Themes

Table 5: Table showing the extracted usability Themes

Extracted Usability Themes:

	Feature	Importance	Theme
0	easy	0.0461	Ease of Use
1	love	0.0430	User Satisfaction
2	great	0.0385	Ease of Use
3	log	0.0166	Navigation Issues
4	helpful	0.0111	Ease of Use
5	work	0.0110	Error Handling
6	doesn	0.0106	Other
7	convenient	0.0097	Ease of Use
8	password	0.0093	Navigation Issues
9	excellent	0.0092	User Interface
10	app	0.0092	General Usage
11	won	0.0079	General Usage
12	update	0.0073	Navigation Issues
13	sign	0.0072	Navigation Issues
14	error	0.0070	Error Handling
15	good	0.0069	User Satisfaction
16	hard	0.0068	Error Handling

The analysis identified key usability themes influencing user satisfaction in mHealth applications. These include Ease of Use (e.g., "easy," "great"), emphasizing intuitive interfaces for better engagement, and Navigation Issues (e.g., "log," "password"), highlighting design flaws that frustrate users. Error Handling (e.g., "error," "hard") pointed to technical issues impacting reliability, while User Satisfaction (e.g., "love," "good") reflected positive experiences. Other themes like General Usage and User Interface underscored the importance of functionality and design. The feature "easy" had the highest impact on sentiment, reaffirming the significance of simplicity and usability in app design, aligning with Nielsen's heuristics (1994).

4.3 Sentiment Trends Over Time

As discussed in the Implementation section, the sentiment trends over time revealed significant fluctuations in user satisfaction, particularly following app updates. These findings are critical to understanding the evolution of usability in mHealth applications

Chapter 5: Discussion

5.1 Implications Limitations and Future Work

How Can User Reviews of Mobile Health (mHealth) Applications Be Effectively Classified into Positive and Negative Sentiments Using Machine Learning Techniques?

The findings confirm that machine learning techniques are effective in classifying user reviews of mHealth applications into positive and negative sentiments (Oyebode et al., 2020). Among the tested models, Long Short-Term Memory (LSTM) emerged as the top performer, demonstrating its suitability for text classification tasks that require context-aware analysis. Additionally, Random Forest and Naive Bayes exhibited strong performance, providing viable alternatives for applications where faster and more interpretable implementations are needed.

Which Review Features Significantly Impact Sentiment Classification, and What Do These Features Reveal About Usability Issues?

Feature importance analysis highlighted that certain keywords significantly influence sentiment classification:

- Positive Sentiments: Keywords like “easy,” “love,” and “great” were strong indicators of positive sentiments, emphasizing the importance of simplicity and user satisfaction in mHealth applications.
- Usability-Specific Terms: Words such as “update” and “convenient” played critical roles, revealing the importance of app updates and ease of use. While many updates improved usability, some introduced challenges, as noted by Liew et al. (2019), who emphasized the importance of reliability in mobile health apps.

These findings confirm that ease of use and task-specific functionalities are pivotal to user satisfaction. They also provide sufficient evidence to reject the null hypothesis (H_0) and accept the alternative hypothesis (H_1), demonstrating that usability factors significantly impact user sentiments.

What Usability-Related Themes Can Be Extracted from User Reviews Using Natural Language Processing (NLP) Techniques?

Using NLP techniques, several usability-related themes were identified:

1. Ease of Use: Keywords like “easy” and “great” highlight simplicity as a crucial factor for user retention, aligning with Nielsen’s usability heuristics (Nielsen, 1994).
2. Navigation Issues: Terms such as “log” and “password” indicated challenges with navigating app features, consistent with findings by Papia et al. (2024), which emphasized navigation as a key usability factor.
3. Error Handling: Words like “error” and “hard” pointed to technical issues like app crashes, underscoring the need for reliable error management to prevent user frustration.
4. User Satisfaction: Positive terms such as “love” and “good” reflected user satisfaction when usability expectations were met, reinforcing the importance of seamless app experiences.

How Do User Sentiments Towards mHealth Applications Evolve Over Time and Across Different App Versions?

User sentiments towards mHealth applications improved over time, with early app versions receiving negative feedback due to navigation issues, technical bugs, and poor interface design. Later versions saw increased positive sentiments, driven by simplified workflows, resolved technical issues, and improved navigation and design. These findings highlight the importance of iterative updates, user feedback, and continuous usability refinement to enhance user satisfaction.

Limitations of the Study

One limitation of this study is the imbalance in the dataset, which, despite the application of SMOTE, may still affect the model's performance on unseen data (Chawla et al., 2002; He & Garcia, 2009). Future research could explore additional data augmentation techniques or larger datasets to mitigate this issue. Another limitation is the focus on English-language reviews, which restricts the generalizability of the findings to non-English-speaking users (Dashtipour et al., 2016) . Future studies should include multilingual data to capture a broader range of user experiences.

Future Work

Future research could focus on improving sentiment classification accuracy and gaining deeper insights by utilizing advanced deep learning models such as BERT and RoBERTa. Additionally, conducting multilingual sentiment analysis would allow for the identification of common usability issues across diverse user groups. Analyzing user cohorts could provide insights into how user satisfaction evolves over time and with app updates. Integrating sentiment analysis with automated usability testing and exploring non-textual data, such as user behavior patterns and performance metrics, would offer a more comprehensive understanding of usability challenges and app performance.

Chapter 6: Conclusion

This study examined user sentiments and usability themes in mHealth applications using machine learning and natural language processing (NLP) techniques. The results revealed valuable insights into user satisfaction and the challenges faced by mHealth app users.

The analysis showed that the LSTM model performed the best in classifying user reviews, achieving an accuracy of 97%. Important usability themes were identified through feature importance analysis, including Ease of Use, Navigation Issues, Error Handling, and User Satisfaction. The study also found that user sentiments improved significantly over time as updates focused on simplifying navigation, fixing bugs, and improving the user experience. Sentiment trends across app versions further emphasized the need for thorough testing and user-centred design to reduce negative feedback.

The findings highlight the value of using sentiment analysis as a tool for understanding user feedback and improving app usability. By addressing critical issues such as poor navigation, technical errors, and unclear interfaces, developers can enhance user satisfaction and engagement.

Moving forward, developers should focus on continuously monitoring user feedback and integrating usability improvements into their updates. A user-centred approach that prioritizes simplicity, technical reliability, and clear communication will ensure that mHealth applications meet the needs of their users and deliver an excellent user experience.

References

1. Bashshur, R.L., Shannon, G.W., Krupinski, E.A. and Grigsby, J. (2011) 'The empirical foundations of telemedicine interventions for chronic disease management', *Telemedicine and e-Health*, 17(10), pp. 769–800. https://go.exlibris.link/wBpLNrHM.

2. **Breiman, L.** (2001) 'Random Forests', *Machine Learning*, 45(1), pp. 5–32. Available at: https://go.exlibris.link/BYVZrslX.

3. Cambria, E., Schuller, B., Xia, Y. and Havasi, C. (2018) 'New avenues in opinion mining and sentiment analysis', *IEEE Intelligent Systems*, 28(2), pp. 15–21. Available at: https://go.exlibris.link/0yWTtjn3.

4. Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P. (2002) 'SMOTE: Synthetic Minority Over-sampling Technique', *Journal of Artificial Intelligence Research*, 16, pp. 321–357. Available at: https://go.exlibris.link/rYTzCyfk.

5. Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014) 'Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling', *arXiv preprint*, arXiv:1412.3555. Available at: https://go.exlibris.link/jBjr9sYL.

6. Dashtipour, K., Poria, S., Hussain, A., Cambria, E., and Hawalah, A.Y. (2016) 'Multilingual Sentiment Analysis: State of the Art and Independent Comparison of Techniques', *Cognitive Computation*, 8(4), pp. 757–771. Available at: https://go.exlibris.link/76HDjf5n.

7. Free, C., Phillips, G., Watson, L., Galli, L., Felix, L., Edwards, P., et al. (2013) 'The effectiveness of mobile-health technology-based health behaviour change or disease management interventions for health care consumers: a systematic review', *PLoS Medicine*, 10(1), e1001362. Available at: https://go.exlibris.link/xnQtYVXH.

8. Harrington, C. N., Ruzic, L. & Sanford, J. A. (2017) 'Universally accessible mHealth apps for older adults: Towards increasing adoption and sustained engagement', in *Universal Access in Human–Computer Interaction. Human and Technological Environments*. Cham: Springer, pp. 3-14. Available at: https://link.springer.com/chapter/10.1007/978-3-319-58700-4_1

9. Kumar, A., Singh, R., and Kaur, S. (2024) 'Three-Class Sentiment Analysis Using LSTM for Sequential Data', *International Journal of Artificial Intelligence and Applications*, 15(1), pp. 34–45. Available at: https://go.exlibris.link/PdMWt9tr.

10. Liew, M.S., Zhang, J., See, J., and Ong, Y.L. (2019) 'Usability Challenges for Health and Wellness Mobile Apps: Mixed-Methods Study Among mHealth Experts and Consumers', *JMIR mHealth and uHealth*, 7(1), e12160. Available at: https://go.exlibris.link/5XyGNRTM.

11. Liu, H. (2017) 'Sentiment Analysis of Citations Using Word2vec', *arXiv preprint arXiv:1704.00177*. Available at: https://go.exlibris.link/yp7lcbHY.

12. Nielsen, J. (1994) *Usability engineering*. Boston: Academic Press. https://go.exlibris.link/bPDyLZqF.

13. **Oyebode, O., Ndulue, C., Alhasani, M., and Orji, R.** (2020) 'Using Machine Learning and Thematic Analysis Methods to Evaluate Mental Health Apps Based on User Reviews', *IEEE Access*, 8, pp. 111141–111158. Available at: https://go.exlibris.link/QgcJSTjM.

14. Papia, S.K., Khan, M.A., Habib, T., Rahman, M., and Islam, M.N. (2024) 'DistilRoBiLSTMFuse: an efficient hybrid deep learning approach for sentiment analysis', *PeerJ Computer Science*, 10, pp. e2349–e2349. Available at: https://go.exlibris.link/hkwzWtBH.

15. Petersen, M. & Hempler, N. F. (2017) 'Development and testing of a mobile application to support diabetes self-management for people with newly diagnosed type 2 diabetes: a design thinking case study', *BMC Medical Informatics and Decision Making*, 17(91). Available at: https://go.exlibris.link/nz8F9N1G.

16. **Rennie, J.D.M., Shih, L., Teevan, J., and Karger, D.R.** (2003) 'Tackling the Poor Assumptions of Naive Bayes Text Classifiers', *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*, Washington DC, pp. 616–623. Available at: https://go.exlibris.link/ttKHjx6W.

17. **Ruelens, A.** (2021) 'Analyzing user-generated content using natural language processing: a case study of public satisfaction with healthcare systems', *Journal of*

Computational Social Science, 5, pp. 731–749. Available at: https://go.exlibris.link/BCcS24Bx.

18. Tan, K.L., Lee, C.P., Lim, K.M., and Anbananthen, K.S.M. (2022) 'Sentiment Analysis With Ensemble Hybrid Deep Learning Model', *IEEE Access*, 10, pp. 103694–103704. Available at: https://go.exlibris.link/vT0LYTzP.

19. Ventola, C.L. (2014) 'Mobile devices and apps for health care professionals: uses and benefits', *Pharmacy and Therapeutics*, 39(5), pp. 356–364. Available at: https://go.exlibris.link/HDKTV21w.

20. Wang, Q., Liu, J., Zhou, L., Tian, J., Chen, X., Zhang, W., Wang, H., Zhou, W. & Gao, Y. (2022) 'Usability evaluation of mHealth apps for elderly individuals: a scoping review', *BMC Medical Informatics and Decision Making*, 22(317). Available at: https://go.exlibris.link/qbt0YsKK.