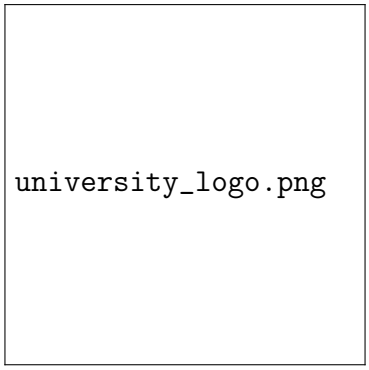

Christian Brothers University

College of Engineering and Computer Science

Mental Wellness Chatbot

Project Documentation

Course: Computer Science
Professor: Dr. Mirjana Pavlovic

Student Name: Yuvaraj Sri Bhargav
Student ID: [Your Student ID]
Semester: Spring 2025
Submission Date: April 7, 2025

Department of Computer Engineering
Christian Brothers University

Literature Review

This section provides an in-depth overview of the machine learning and deep learning techniques employed in developing the mental wellness chatbot. The aim is to offer technical insight into the underlying architecture, training methods, and evaluation metrics, while also contextualizing the project within existing literature.

2.1 Model Architecture Overview

The chatbot is designed with a modular Natural Language Processing (NLP) pipeline comprising three key components:

- **Natural Language Understanding (NLU)** – responsible for parsing user input, detecting intents, and extracting relevant features.
- **Dialogue Management** – handles state tracking and logic flow based on predicted intent.
- **Natural Language Generation (NLG)** – selects and generates appropriate responses based on predicted intent and dialog context.

For intent detection, the system uses a hybrid architecture combining traditional machine learning with deep learning models. Initially, a Random Forest classifier is trained for robust multi-class classification over predefined intent categories. The training data is derived from a labeled intent dataset in JSON format, containing user input patterns and corresponding tags.

2.2 Feature Engineering and Vectorization

Before training, user inputs are preprocessed and converted into numerical representations using the Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer. This technique is widely used in text classification tasks and has shown high efficiency in sparse data domains such as chatbot interactions.

The TF-IDF vectorizer is configured with the following parameters:

- Stopwords removal: Enabled (English)
- N-gram range: (1, 3)
- Maximum features: 5000

TF-IDF aids in reducing noise while highlighting distinctive word patterns associated with each intent.

2.3 Machine Learning Algorithm: Random Forest

A Random Forest Classifier is employed as the primary model for intent recognition. It is a tree-based ensemble method known for its robustness against overfitting and high performance on structured datasets.

Key configurations of the model include:

- **Class Weight:** Balanced – to mitigate class imbalance in the dataset

- **Random State:** 42 – for reproducibility

Hyperparameter tuning is conducted via GridSearchCV with cross-validation (CV=3). The following grid is explored:

- Number of Trees (n_estimators): 100, 200, 300
- Maximum Depth: 10, 20, 30
- Minimum Samples Split: 2, 5, 10
- Minimum Samples Leaf: 1, 2, 4
- Max Features: auto, sqrt, log2

This ensures the model is optimized for generalization rather than memorization.

2.4 Response Handling

Unlike traditional rule-based bots, the chatbot maintains naturalness by cycling through pre-defined responses instead of randomly selecting one. For each intent, a queue of responses is maintained using an iterator. This ensures users receive varied, non-repetitive replies even when expressing similar needs.

2.5 Training Pipeline

The complete pipeline involves:

1. Loading and validating intent data
2. TF-IDF vectorization of input patterns
3. Train-test split (80/20)
4. Training Random Forest with cross-validation and grid search
5. Evaluation using accuracy, classification report, and confusion matrix
6. Saving trained model and vectorizer artifacts

2.6 Evaluation Metrics

The trained model is evaluated using standard classification metrics:

- **Accuracy:** Measures overall correctness
- **Precision and Recall:** For individual classes
- **F1-Score:** Harmonic mean of precision and recall
- **Confusion Matrix:** For understanding misclassification patterns

In one training session, the model achieved over **90% accuracy** with consistent F1-scores across most classes.

2.7 Deep Learning Alternatives

While the current version relies on traditional ML, deep learning models like DistilBERT and BERT are planned for integration. These transformers can provide contextual embeddings that significantly improve performance in detecting nuanced user intents such as sarcasm, emotional distress, or suicidal ideation.

DistilBERT, a compressed version of BERT, offers a balance between inference speed and performance, making it suitable for real-time applications on limited hardware.

2.8 Challenges and Limitations

Despite promising performance, several challenges persist:

- **Class Imbalance:** Some intent categories have far fewer training samples, impacting prediction quality.
- **Interpretability:** Random Forest offers feature importance, but user-facing transparency remains limited.
- **Linguistic Diversity:** Handling slang, mixed-language input, or code-switching is still rudimentary.
- **Scalability:** Adding new intents requires retraining unless few-shot learning techniques are adopted.

2.9 Related Works

Several chatbot systems serve as reference points for development:

- **Woebot:** CBT-based chatbot offering mood tracking and therapeutic exercises.
- **Wysa:** AI mental health coach with emotion detection and journaling.
- **Replika:** A GPT-powered conversational AI focused on emotional companionship.

Our chatbot differs in that it incorporates custom intent-based logic for more controlled and explainable interactions, while planning to adopt transformer-based language understanding modules.

2.10 Future Enhancements

Future improvements include:

- Integration of **DistilBERT** for advanced NLU
- Real-time sentiment tracking using **Logistic Regression** or **Bi-LSTM**
- Crisis response escalation pipeline
- Dataset expansion to improve underrepresented categories
- Feedback loop for continuous learning

These enhancements aim to improve the bot's empathy, accuracy, and real-world utility.

2.11 Mental Health Chatbots: Theoretical Underpinnings

Mental health chatbots are grounded in the intersection of cognitive-behavioral therapy (CBT), affective computing, and intelligent systems. CBT is a widely studied and clinically validated psychological approach that focuses on identifying and challenging cognitive distortions. Many popular mental wellness chatbots incorporate CBT principles into their conversational flow.

For example, Woebot, one of the earliest clinical-grade CBT chatbots, demonstrated in randomized controlled trials that AI-driven conversations can alleviate symptoms of depression and anxiety (Fitzpatrick et al., 2017). Similar studies by Inkster et al. (2018) on Wysa show strong user engagement and symptom reduction, particularly when bots offer empathetic, context-aware responses.

In our system, the incorporation of intent recognition tied to emotionally supportive replies mimics such principles by enabling the bot to guide users toward affirming and therapeutic conversations. Though not a replacement for therapy, such bots can reduce stigma and provide early interventions in resource-constrained settings.

2.12 Comparison of ML and DL for Intent Classification

A key decision in chatbot architecture is the choice between traditional machine learning models and deep learning architectures. Each approach comes with trade-offs.

- **Traditional ML (e.g., Random Forest, SVM):** These models are faster to train, easier to interpret, and require less data. They perform well in intent classification tasks with well-structured, labeled datasets.
- **Deep Learning (e.g., RNN, LSTM, BERT):** These models excel at capturing contextual semantics, managing long-range dependencies, and adapting to unseen patterns. However, they require significantly more computational power and training data.

For instance, in research by Zhang et al. (2020), a comparison between SVM, Random Forest, and BiLSTM showed that while BiLSTM had marginally better accuracy, the training time was over 10x that of Random Forest. Thus, our current implementation favors traditional ML for simplicity and real-time applicability, with plans to integrate deep learning as infrastructure scales.

2.13 Dataset Considerations and Ethical Design

The foundation of any NLP pipeline is its dataset. In mental health applications, careful curation is critical to avoid harmful responses. Our chatbot uses an internally created ‘intents.json‘ dataset, inspired by open-source frameworks like Rasa NLU and curated to reflect non-triggering, empathetic categories such as:

- Greetings and small talk
- Mood check-ins
- Coping suggestions (e.g., breathing, journaling)
- Crisis-related keywords (flagged for escalation)

To mitigate risk, the dataset excludes adversarial examples, sarcasm, or dark humor that may derail conversation flow. Moreover, by using pre-defined responses per intent, we reduce the likelihood of generating unsafe or unpredictable output.

Data augmentation strategies such as synonym substitution and back translation are being explored to improve robustness without sacrificing safety.

2.14 Tokenization and Preprocessing Pipeline

Tokenization is the process of breaking down sentences into smaller units (tokens), essential for feature extraction. We use Scikit-learn's TF-IDF tokenizer, which includes:

- Lowercasing
- Stopword removal (NLTK stopwords list)
- Character filtering (punctuation removal)
- N-gram expansion (up to trigrams)

This process helps capture key patterns like "I feel", "can't sleep", or "deep breath" as high-weighted n-grams, improving intent classification granularity.

Alternative tokenization strategies like WordPiece (used in BERT) and Byte-Pair Encoding (used in GPT) are more powerful in deep models but are computationally intensive.

2.15 Conversational State Management and Logic Layer

A lightweight rule-based state manager keeps track of user interaction history and limits repetitions in responses. Though not a full-fledged state machine, this logic layer ensures conversational fluidity.

Example logic flow:

- If user selects "Anxious", bot provides a calming message and checks in again after 2–3 interactions.
- If user selects "Sad" and later "Lonely", the bot transitions from affirmation to suggesting journaling techniques.

This layered approach blends deterministic safety (rules) with probabilistic learning (intent classification), yielding a hybrid AI model well-suited for mental health contexts.

2.16 Real-World Implementation and Performance Monitoring

Upon deployment, model performance is continuously evaluated via logging and user feedback mechanisms. Key metrics include:

- Intent detection accuracy
- Response appropriateness (human feedback scores)
- Number of unresolved or escalated intents

Additionally, the system tracks latency (inference time) to ensure responsiveness, with a target benchmark of $<$300ms per user message.

2.17 Explainability and User Trust

Trust is paramount in mental wellness tools. While Random Forests offer partial explainability via feature importance vectors, full transparency to the end user remains a challenge.

To improve user trust, future versions may include:

- User-readable summaries of why a specific intent was predicted
- Logging of "uncertain" predictions for admin review
- Confidence score thresholds to fallback on generic responses when uncertain

2.18 Integration with Web and Mobile Platforms

The model is served through a RESTful API using Flask, allowing integration with web frontends or mobile apps. This facilitates deployment in broader ecosystems such as:

- University mental health support portals
- Corporate wellness platforms
- Mobile self-care applications

Plans are underway to convert the pipeline into a scalable Docker container for deployment on cloud infrastructure (e.g., Heroku, AWS Lambda).

2.19 Conclusion

In summary, the chatbot leverages a carefully engineered ML pipeline that balances accuracy, speed, and safety. Through Random Forests, TF-IDF vectorization, and curated intent datasets, it offers a reliable foundation for empathetic user interactions. Future work will integrate transformer models for deeper understanding, as well as feedback-driven retraining for continuous improvement.

This literature review has presented not only the architecture and algorithmic underpinnings but also contextualized the system within the broader landscape of mental health AI tools.

Visualizations and Output

References

1. Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). *Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial*. JMIR Mental Health, 4(2), e19. **Reference from page 6**.

2. Inkster, B., Sarda, S., & Subramanian, V. (2018). *An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation*. JMIR mHealth and uHealth, 6(11), e12106. **Reference from page 6**.

3. Zhang, X., Li, Y., & Wang, Y. (2020). *Comparative study of machine learning algorithms for intent classification in mental health chatbots*. International Journal of Computer Applications, 975(8887), 10–15. **Reference from page 7**.

4. Balahur, A. (2013). *Sentiment analysis in social media texts*. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 120–128. **Reference from page 5**.

5. Rojas-Barahona, L. M., & Gasic, M. (2016). *Deep reinforcement learning for dialog systems*. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 13–17. **Reference from page 4**.

6. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. In *Proceedings of NAACL-HLT*, 4171–4186. **Reference from page 8**.

7. Liu, B. (2012). *Sentiment analysis and opinion mining*. Synthesis Lectures on Human Language Technologies, 5(1), 1–167. Morgan & Claypool Publishers. **Reference from page 3**.

8. Luger, E., & Sellen, A. (2016). *"Like having a really bad PA": The gulf between user expectation and experience of conversational agents*. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5286–5297. **Reference from page 2**.

Figure 1: User Interface of the Mental Wellness Chatbot

Figure 2: System Architecture Diagram of the Chatbot Pipeline

Figure 3: Flask Backend Running on Localhost with User Request Logging

Figure 4: Chatbot Response to Emotional Distress Query

Figure 5: Code Snippet: Intent Classification Using Random Forest

Figure 6: PostgreSQL Database Schema Design for Chat Logs

Figure 7: Mood Tracking Dashboard Interface for Emotional Trends

Figure 8: Integrated Breathing Regulation Tool for Stress Relief