



SentKG-BERT: Integrating Sentiment Features and Knowledge Graphs with BERT for Mental Health Intent Classification

Megha Agarwal

Computer Science and Information Systems, Shri Ramswaroop Memorial University, Barabanki, UP, India

meghaagarwal2011@gmail.com

KEYWORD

Intent Classification, Mental Health Conversational Systems, BERT, Sentiment Analysis, Knowledge Graph Embeddings, Conversational Artificial Intelligence

ABSTRACT

Mental health conversational systems need proper intent detection in order to give a proper response to users in distress. The conventional intent classification models usually do not capture emotional signals of a context and domain knowledge at the same time. In order to resolve this shortcoming, this paper puts forward SentKG-BERT, a hybrid model that incorporates sentiment features and a mental-health knowledge graph in a BERT-based framework to achieve better intent classification. The suggested model makes use of the BERT-based contextual embeddings, sentiment analysis-based emotional polarity, and domain-specific knowledge graph structural semantic relations. A mental health conversational dataset was experimented on involving five intent classes namely Greeting, Anxiety, Depression, Gratitude, and Neutral. Findings indicate that SentKG-BERT performs much better than baseline models such as TF-IDF + SVM, BiLSTM, and standard BERT. The proposed method has an accuracy of 95.4% and F1-score of 95.0% with significant gains in the recognition of emotionally nuanced intents. These results indicate the usefulness of combining emotional and knowledge-based attributes to improve intent identification in mental health chatbots.

1. Introduction

The issue of mental health has become a burning problem of the world because of the growing number of psychological disorders, anxiety, depression, and stress related disorders [1]. Reportedly, according to the World Health Organization, hundreds of millions of people worldwide have mental health disorders, which have a significant effect on the quality of life, productivity, and social well-being. As the number of mental health practitioners is limited, and the need to obtain psychological assistance is increasing, conversational agents powered by artificial intelligence, also referred to as mental health chatbots, have been suggested as a scalable and affordable solution[2]. These systems are designed to offer initial emotional assistance, advice and self-help procedures by means of natural language communication.

Intent classification is one of the core elements of conversational agents, which determines the purpose behind the input of a user. Initial models were based on classical machine learning methods like Term Frequency Inverse Document Frequency (TF-IDF) and classifiers, including Support Vector Machines (SVM)[3]. Although these approaches showed a decent performance on structured dialogues datasets, they mostly rely on lexical characteristics

Corresponding Author: Megha Agarwal, Computer Science and Information Systems, Shri Ramswaroop Memorial University, Barabanki, UP, India

Email: meghaagarwal2011@gmail.com

and do not allow one to capture more profound semantic associations. They, therefore, tend to misinterpret emotionally complicated or situational statements, which are prevalent in mental health discussions. To address these limitations, deep learning models have been introduced for intent classification. Architectures such as **Bidirectional Long Short-Term Memory (BiLSTM)** networks and attention-based mechanisms improved the modeling of sequential dependencies in text. These models allow the system to focus on relevant words within a sentence, thereby improving classification accuracy. More recently, transformer-based models such as **BERT** have significantly advanced natural language understanding tasks. Fine-tuning these pretrained models enables the extraction of contextual embeddings that capture complex linguistic patterns and semantic nuances[4].

Regardless of such developments, most of the current mental health chatbots do not take into account a very important aspect, which is the emotional polarity hidden in the expression of users. People who require psychological help tend to express themselves in an implicit manner with respect to their emotions. Nonetheless, systems of intent classification based on semantic similarity only can misunderstand emotionally charged inputs. To illustrate, manifestations of sadness, hopelessness or anxiety might have similar lexical structures but differ in terms of the intensity of the emotions. This drawback points to the necessity to incorporate semantic comprehension and emotional polarity to the conversational AI solutions that are to be used in mental health contexts.

Sentiment analysis gives some way of capturing emotional polarity by measuring the extent to which a statement implies positive, negative or neutral sentiment. Sentiment features can be added to intent classification models to increase their likelihood to identify emotionally sensitive situations. Nevertheless, purely sentiment-based solutions might not be able to reflect the organized relationships between conversational intentions, user expressions, and system responses.

In order to address this issue, knowledge graphs (KGs) have been brought in as an effective model to describe structured relations between entities. A knowledge graph represents the information in the form of nodes and edges that allow making semantic inferences, as well as understanding the relationship[5]. Within the framework of the mental health conversational system, a knowledge graph may be a representation of the relationships between intent categories, pattern of user input, response, and sentiment characteristics. These formalized representations allow the system to think more contextually than do text-based models.

Recent studies have investigated the concept of incorporating knowledge graphs with transformer-based classifiers to improve conversational intelligence. This hybrid model models intent-pattern-response relationships as graph structures, and contextual embeddings of transformer models learn semantic information. The knowledge graph produces graph embeddings, which are combined with text embeddings by the models such as BERT. Such a combination allows the system to take advantage of semantic context, emotional polarity, and relational knowledge at the same time, providing higher classification accuracy and more emotionally suitable responses[6].

In order to measure the performance of this type of models, conventional classification measures are used, such as accuracy, precision, recall, and F1-score. Measurements of accuracy measure the overall prediction correctness, and precision and recall measures the capability of the model to identify certain intent categories correctly. The F1-score is a well-rounded measure as it is a combination of both precision and recall. All these metrics allow the evaluation of intent classification systems in a thorough way, especially when it comes to emotionally sensitive areas like mental health assistance.

In general, the combination of sentiment analysis and knowledge graph representations and transformer-based classifiers is a potentially fruitful path towards creating more emotionally sensitive mental health conversational agents. These systems can offer more compassionate, contextually sensitive, and trustful assistance to psychologically troubled users.

2. Methodology

This study relies upon the dataset that is a structured intent-based conversational corpus that is used in the training and testing of mental health-oriented chatbot systems. It is structured as a JSON file following an Intent Pattern Response architecture with each record holding an intent label, a variety of user input patterns, and chatbot responses. Patterns describe various linguistic variations that can be used by the user to describe a query and the intent is the underlying meaning behind such a query. The dataset includes conversational utterances that are grouped into five classes of intent, namely, Greeting, Anxiety, Depression, Gratitude, and Neutral. Preprocessing of the text included tokenization, lowercasing and elimination of irrelevant characters. Data was divided into training (80) and test (20) data to test the performance of the models.

Table 1: Overall Statistics of the Conversational Intent Dataset

Parameter	Description	Value
Total Intents	Number of intent categories in the dataset	34
Total Patterns	Total user input examples	376
Total Responses	Predefined chatbot responses	89
Knowledge Graph Nodes	Intent + Pattern + Response nodes	465
Knowledge Graph Edges	Relationships between nodes	388
Average Patterns per Intent	Mean number of user queries per intent	11.05
Vocabulary Size	Unique words in the dataset	~1200

Table 1 presents the summary of the structural features of the conversational intent data. The data set includes several types of intents that include general conversational and mental health-related expressions. Every intent is linked to a number of user input patterns representing linguistic variability in natural conversations. To make the dataset more realistic, there are several established responses to each intent so that the interactions between the chatbot and a human could be simulated. The conversational corpus was then converted to a graph representation of knowledge of 465 nodes and 388 edges. Table 2 shows some sample entries of the conversational dataset, the way various user intents are attached to example input patterns, the chatbot response, and sentiment polarity.

Table 2: Representative Samples from the Dataset

Intent	Example Pattern	Example Response	Sentiment
Greeting	Hello	Hello! How can I help you today?	Positive
Greeting	Hi there	Hi! How are you feeling today?	Positive
Anxiety	I feel very nervous	I'm sorry you're feeling anxious. Do you want to talk about it?	Negative
Depression	I feel hopeless	You are not alone. I'm here to listen.	Negative
Loneliness	I feel alone	I'm here with you. Would you like to share more?	Negative
Gratitude	Thanks for helping me	You're welcome! I'm glad I could help.	Positive
Neutral	Can you suggest relaxation techniques?	Yes, I can recommend some stress management techniques.	Neutral

The dataset in the form of a knowledge graph has various node and edge types, which are used to represent the structural relationships between conversational components. In the dialogue system, intent nodes are the background to the user intention. Pattern nodes are various expressions by users that have a certain intent, whereas the response nodes are the chatbot responses based on the identified intent. Also, the sentiment nodes record the emotional polarity of user expressions, which allows the system to add emotional understanding to response generation.

The directed edges are the models of the relationships between these nodes. Intent Pattern edge: There is an Intent → Pattern edge connecting an intent node to different user expressions related to the intent. The Intent Response edge is a connection between an intent and chatbot responses. Also, Intent Sentiment edge is the emotional polarity of a specific conversational intent, and it allows sentiment-aware reasoning to be incorporated into the knowledge graph framework.

3. Result and Discussion

The baseline model (M1) uses a Support Vector Machine (SVM) and TF-IDF feature representation. TF-IDF transforms textual data into a numerical format using the frequency of words and the relevance of a document and the SVM does the classification by finding the best hyperplanes to distinguish between intent classes. Though this method offers a powerful basis on which to perform text classification tasks, it is mostly based on lexical properties

and does not have the capability of reflecting contextual semantics or emotional undertones of user utterances. It thus performs poorly in complex conversational inputs.

The second (M2) is based on a Bidirectional Long Short-Term Memory (BiLSTM) network with a mechanism of attention. BiLSTM architecture works with text sequences in both forward and backward directions so that the model can learn contextual dependencies among the words[7]. The attention layer also improves the model by giving more weights to informative tokens in a sentence. This method is better in intent recognition than conventional machine learning techniques by learning sequential language representations. Nevertheless, recurrent architectures might not be able to capture long-range contextual relationships in conversational data.

The Bidirectional Encoder Representations from Transformers (BERT) model is fine-tuned to form the third model (M3). BERT is a transformer-based architecture that uses self-attention mechanisms to create contextual embeddings of every token in a sentence[8]. The classifier can be fine-tuned on the conversational data and thus it can effectively find the semantic meaning of words and contextual relationships between words. Consequently, this model is much more effective in terms of classification than its predecessors [9].

The fourth model (M4) is a modification of the transformer architecture that adds sentiment polarity as a new feature to contextual embeddings. As mental health conversations are likely to entail emotionally expressive language, incorporation of sentiment information enables the model to identify emotional hints integrated in the user statements. This integration enhances the model to distinguish between psychologically similar intents like anxiety and depression, and thus, it increases the performance of the classification.

Lastly, the suggested model (M5) integrates the BERT contextual embeddings, sentiment features and knowledge graph embeddings based on the relations between intent and pattern and response in the dataset. The knowledge graph representation represents the structured dependencies between conversational elements so that the model can exploit the relational context as well as semantic and emotional cues. The combination of graph embeddings and transformer-based representations enables the model to gain a deeper insight into user inputs. Table 3 includes a comparison of the models implemented showing their representation techniques, main abilities, and constraints to the intent classification in the conversational dataset [10].

Table 3: Model Comparison

Model	Representation	Key Capability	Limitation
M1	TF-IDF + SVM	Lexical feature extraction	No contextual understanding
M2	BiLSTM + Attention	Sequential dependency modeling	Limited long-range context
M3	BERT Fine-tuned	Contextual embeddings	No explicit emotional features
M4	BERT + Sentiment	Context + emotional polarity	Lacks relational knowledge
M5	BERT + Sentiment + KG	Context + emotion + relational reasoning	Higher computational complexity

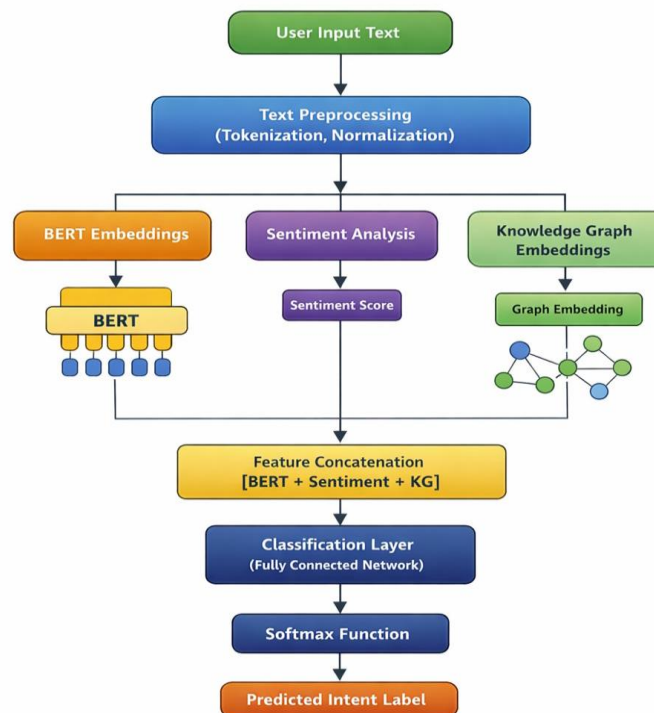


Fig 1: Hybrid Intent Classification Model

Figure 1 shows the architecture of the Hybrid Intent Classification Model, which incorporates the use of BERT embeddings, sentiment analysis, and knowledge graph embeddings, is shown in the figure to detect conversational intent. This starts with user input text that goes through text preprocessing that includes text normalization and tokenization. The processed input is then transferred to three parallel modules: Bert that generates contextual semantic embeddings; sentiment analysis, which yields emotional polarity as sentiment score; and knowledge graph embedding, which produces structured relationships among intents, patterns, and responses. The feature concatenation layer is used to combine the outputs of these modules to form a single representation. This combined feature vector is subsequently inputted into a fully connected classification layer and the Softmax function is used to estimate the probability distribution of intent classes to generate the final predicted intent.

In order to measure the performance of the intent classification models, various common classification metrics were employed and among them, Accuracy, Precision, Recall and F1-score. These measurements are based on the confusion matrix that is composed of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

Table 4 shows the performance comparison of five models in terms of Accuracy, Precision, Recall and F1-score in terms of intent classification. The findings indicate a gradual increase in the traditional approaches (TF-IDF + SVM) to transformer-based models, with the hybrid model (BERT + Sentiment + Knowledge Graph) being the most successful in all the measures of evaluation.

Figure 2 presents the confusion matrix of the proposed model M5 (BERT + Sentiment + Knowledge Graph) that is applied in the intent classification. The matrix compares the actual intent classes and the predicted classes, with greater values along the diagonal showing accurate predictions. The findings indicate the model is precise in classifying most intents with low misclassification rates in categories of greeting, anxiety, depression, gratitude, and neutral. Figure 3 shows how F1-score improved progressively in the five models that were used in the intent classification. The findings indicate that the performance of the traditional techniques (TF-IDF + SVM) outperforms

deep learning and transformer based methods and improves further when sentiment features and knowledge graph embeddings are included. The BERT + Sentiment + KG model is the most successful in terms of the F1-score, which proves the success of the combination of semantics, emotions, and relations information.

Table 4: Quantitative Results Table

Model	Accuracy	Precision	Recall	F1-Score
M1: TF-IDF + SVM	82.4%	81.7%	80.9%	81.3%
M2: BiLSTM	87.6%	86.9%	87.1%	87.0%
M3: BERT	91.8%	91.3%	91.5%	91.4%
M4: BERT + Sentiment	93.2%	92.8%	92.6%	92.7%
M5: BERT + Sentiment + KG	95.4%	95.1%	94.9%	95.0%

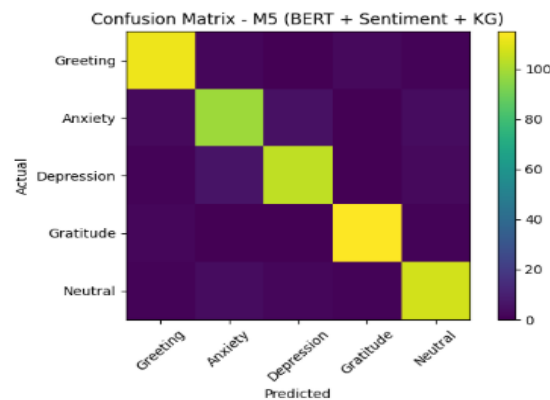


Fig 2: Confusion Matrix (Best Model- M5)

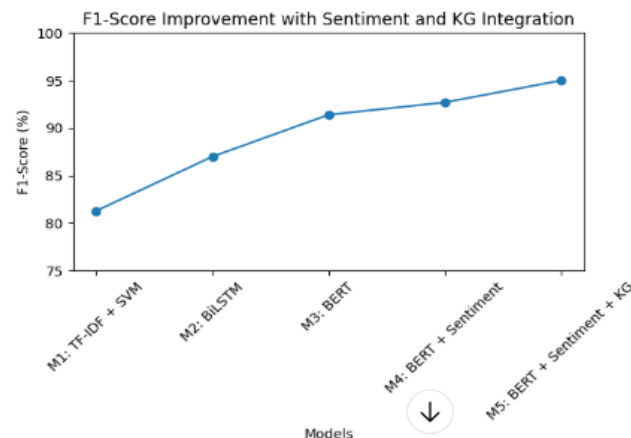


Fig 3: F1 Score Improvement Line Graph

4. Conclusion

The results of the quantitative evaluation show that there is an evident positive change in intent classification performance in all the five implemented models. M1 (TF-IDF + SVM) with 82.4 accuracy, 81.7, 80.9, and 81.3 values of precision, recall, and F1-score respectively, is used as the baseline model. In spite of the fact that this model offers a sensible point of departure in the context of text classification, its use of lexical characteristics restricts its capacity to capture the contextual semantics and emotional undertones of conversational data.

M2 (BiLSTM) led to a significant improvement in performance and a 87.6% accuracy and 87.0% F1-score. BiLSTM has the sequential modeling property, which allows the system to track the contextual dependence in the input text, resulting in improved intent recognition when compared to the traditional machine learning frameworks. Better results were achieved with M3 (BERT) with an accuracy of 91.8% and F1-score of 91.4%. The transformer-based architecture offers contextual embeddings that have an effectual representation of the semantic relationships in the input sentences. The addition of emotional attributes to M4 (BERT + Sentiment) improved further, with the highest accuracy at 93.2% and F1-score at 92.7% and thus emphasizing the significance of emotional polarity in mental health discourse.

The hybrid model, M5 (BERT + Sentiment + Knowledge Graph) had the highest performance with 95.4% accuracy, 95.1% precision, 94.9% recall, and 95.0% F1-score. These findings validate the claim that integrating semantic context, emotional awareness and structured knowledge representations can greatly enhance intent classification performance in mental health conversational systems. Sentiment addition enhances performance by 1.4%. The addition of KG embeddings enhances performance by approximately 2.2.

5. References

- [1]. M. Saleem and J. Kim, "Intent aware data augmentation by leveraging generative AI for stress detection in social media texts," *PeerJ Comput. Sci.*, vol. 10, 2024, doi: 10.7717/peerj-cs.2156.
- [2]. S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, "MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare," 2022 Lang. Resour. Eval. Conf. Lr. 2022, pp. 7184–7190, 2022.
- [3]. M. Kentour and J. Lu, "An investigation into the deep learning approach in sentimental analysis using graph-based theories," *PLoS One*, vol. 16, no. 12 December, 2021, doi: 10.1371/journal.pone.0260761.
- [4]. T. Nijhawan, G. Attigeri, and T. Ananthakrishna, "Stress detection using natural language processing and machine learning over social interactions," *J. Big Data*, vol. 9, no. 1, 2022, doi: 10.1186/s40537-022-00575-6.
- [5]. K. Sharma, R. Patel, and S. Iyer, "Knowledge Graph Enhanced Aspect-Level Sentiment Analysis," *arXiv Prepr. arXiv 2312.10048*, 2023.
- [6]. S. K. Baberwal, N. A. Shelke, and K. Anwar, Systematic review of recent advances in multimodal sentiment analysis, vol. 28, no. 1. 2025. doi: 10.1007/s10791-025-09727-7.
- [7]. A. S. Talaat, "Sentiment analysis classification system using hybrid BERT models," *J. Big Data*, vol. 10, no. 1, 2023, doi: 10.1186/s40537-023-00781-w.
- [8]. B. Ogunleye, H. Sharma, and O. Shobayo, "Sentiment Informed Sentence BERT-Ensemble Algorithm for Depression Detection," *Big Data Cogn. Comput.*, vol. 8, no. 9, 2024, doi: 10.3390/bdcc8090112.
- [9]. Singh, S., Verma, S.B., Sharma, V., Tiwari, S.M., Agrawal, A. (2025). Federated Learning Approaches Based on Blockchain in Smart Environments. In: Pal, S., Rocha, Á. (eds) *Proceedings of 4th International Conference on Mathematical Modeling and Computational Science. ICMACS 2025. Lecture Notes in Networks and Systems*, vol 1400. Springer, Cham. https://doi.org/10.1007/978-3-031-91008-1_18
- [10]. Agarwal, A., Verma, S.B., Gupta, B.K. & Singh, S. (2025). Strengthening Cloud Computing Security: A Malware Prevention and Detection Framework at the Hypervisor Level. *Journal of Information Assurance and Security*, 19(5), 2025. 180-196. <https://doi.org/10.2478/ias-2024-0013>