KEC Conference 2026

ISSN: 3021-940X (Print) DOI: 10.3126/injet.v3i2.95508

Hybrid NLP Architecture for Crisis-Aware Chatbot Integrating Emotion Classification and Retrieval-Augmented Legal Response Generation

Sandhya Pant[1*], Sneha Ale[2], Prashna Timilsina[3], Shweta Jha[4]

[1] *Department of Computer and Electronics, Kantipur Engineering College, Dhapakhel, Lalitpur,Nepal*
sandhya10pant10@gmail.com

[2] *Department of Computer and Electronics, Kantipur Engineering College, Dhapakhel, Lalitpur,Nepal*
sneha.alenepal@gmail.com

[3] *Department of Computer and Electronics, Kantipur Engineering College, Dhapakhel, Lalitpur,Nepal*
prashnatimilsinaofficial@gmail.com

[4] *Department of Computer and Electronics, Kantipur Engineering College, Dhapakhel, Lalitpur,Nepal*
myselfshweta521@gmail.com

Abstract

Survivors of domestic violence frequently face barriers to accessing timely legal and emotional support, underscoring the need for intelligent, accessible assistance systems. This paper presents a legally capable and emotionally aware conversational system designed to classify user-generated text and provide contextually rich responses in the domain of domestic violence. The system employs BERT-based transformer models to perform three layers of classification on pre-processed user input: crisis classification (safe, urgent, or emergency), intent classification (legal or emotional), and emotion classification across seven fine-grained psychological categories. A Retrieval-Augmented Generation (RAG) module then retrieves relevant content from a curated legal knowledge base to ground legal responses in verified documents, reducing hallucination and improving factual reliability. Experimental results demonstrate that the hybrid architecture supports reliable, safety-aware response generation and shows promise as a supportive tool for survivors in crisis situations.

Keywords: BERT, Retrieval-Augmented Generation (RAG), Conversational System, Legal and Emotional Support, Domestic Violence, Hybrid Architecture, Classification.

1. Introduction

Domestic violence is a systematic pattern of behaviour that includes physical battering, economic abuse, emotional abuse, and sexual violence. It is intended to gain or maintain power and control over a romantic or intimate partner, and is used to intimidate, frighten, terrorize, humiliate, blame, or injure [1].

It has severe and long-lasting effects on survivors. Many victims hesitate to seek help due to social stigma, fear, lack of awareness, or limited access to professional support services. Intimate Partner Violence which falls under Domestic Violence affects women's physical and mental health through direct pathways, such as injury, and indirect pathways, such as chronic health problems that arise from prolonged stress [2].

A history of experiencing violence is therefore a risk factor for many diseases and conditions. Even when assistance is available, survivors may find it difficult to identify trustworthy resources or navigate the emotional and legal challenges that follow abuse.

With the increasing use of digital platforms, intelligent systems such as chatbots have emerged as potential tools for providing accessible and confidential assistance. AI-based conversational systems can offer continuous support, guide users toward relevant resources, and identify the urgency of user messages. Previous research highlights the effectiveness of chatbot-based platforms in supporting individuals facing emotional distress and crisis situations especially younger ones [3].

Motivated by these challenges, a chatbot is designed to provide assistance in both emotional and legal areas. The system uses Natural Language Processing (NLP) and transformer-based models such as BERT to better

understand user input and identify contexts. Additionally, a Retrieval-Augmented Generation (RAG) framework is integrated to retrieve relevant emotional and legal support information from external knowledge sources.

The proposed system aims to provide a supportive and confidential platform where survivors can receive emotional guidance, legal information, and emergency assistance. Such systems can be particularly useful for organizations such as NGOs and support centres that require accessible 24/7 assistance tools for individuals experiencing abuse.

2. Related Works

Recent advancements in artificial intelligence and conversational systems have enabled the development of chatbots designed to support individuals experiencing harassment, abuse, and emotional distress. These systems aim to provide accessible information, emotional support, and guidance in situations where direct assistance may not be immediately available.

Tobias et al. proposed a machine learning based chatbot designed to assist survivors of sexual harassment. Their system demonstrated a success rate of approximately 98% in identifying harassment-related cases and around 80% accuracy in detecting specific harassment categories. Additionally, the system achieved more than 90% accuracy in extracting location and date information, while identifying time-related events remained more challenging, with accuracy close to 80%. Initial validation results indicated strong potential for deploying such systems as socially beneficial tools to support victims and improve access to assistance services [4].

Sugiura examined the growing role of artificial intelligence in addressing domestic abuse cases, particularly during the COVID-19 pandemic when access to traditional support services became more limited. The study highlighted how AI-driven and rule-based chatbots can provide alternative communication channels for victims and survivors seeking help in situations where direct human support is difficult to access [5].

Johnston-Way et al. emphasized the broader societal importance of victim support systems in improving community safety and well-being. Their work highlights that effective victim assistance programs contribute to public health improvements, cost savings, and increased confidence in the criminal justice system. However, the authors also note that empirical research examining the long-term impact of technological interventions for victims remains limited, indicating a need for further study and data-driven evaluation in this domain [6].

In addition to conversational AI, recent research has explored the integration of Retrieval-Augmented Generation (RAG) frameworks to enhance the reliability of chatbot responses. RAG-based systems combine neural language models with external knowledge retrieval mechanisms, enabling the generation of responses grounded in verified documents or knowledge bases. This approach reduces hallucination risks commonly associated with generative models and improves factual accuracy in information-sensitive domains such as healthcare, law, and crisis support. Studies have demonstrated that retrieval-augmented systems significantly improve contextual relevance and trustworthiness when responding to user queries that require domain-specific knowledge.

2.1 Former Developments

Several practical chatbot systems have been developed to improve accessibility to legal information, emotional support, and crisis intervention resources for victims of abuse and harassment. These developments demonstrate how conversational technologies can bridge gaps in traditional support infrastructures.

2.1.1 LAW-U

LAW-U is an AI-powered chatbot designed to provide legal assistance to survivors of sexual violence by recommending relevant Supreme Court rulings based on the user's situation. The name "LAW-U," derived from the Thai phrase meaning "I will wait for you," symbolizes the system's intention to provide continuous support and empathy. The platform aims to empower survivors by helping them understand their legal rights

and available options while raising awareness about sexual violence. The system also serves as a model for future AI-driven legal assistance tools in similar domains [7].

2.1.2 rAInbow Chatbot

The rAInbow chatbot, launched in 2021 in South Africa, provides tailored conversational support to women experiencing domestic abuse. Developed as a social enterprise initiative, the system focuses on addressing the emotional isolation often experienced by survivors. rAInbow helps users identify early warning signs of abusive behavior, understand their rights, and learn from the experiences of others through personalized narratives. By combining ethical design principles with scalable AI technology, the system aims to expand access to support resources for vulnerable individuals [8].

2.1.3 Sophia

Sophia is a digital companion chatbot designed to provide anonymous and confidential assistance to individuals affected by domestic abuse. The system offers a secure platform where users can explore relationship patterns, document important events, and access support resources. Available 24/7, Sophia focuses on providing guidance and information while maintaining strict privacy protections, enabling individuals to seek help in a safe and accessible environment [9].

3. Research Gap

Recent advancements in artificial intelligence and conversational systems have enabled the development of chatbots for supporting individuals experiencing harassment, abuse, and emotional distress. However, a careful review of prior research and deployed systems reveals several critical limitations.

3.1 Limited Integration of Emotional and Legal Support

Existing chatbots, such as LAW-U, rAInbow, and Sophia, primarily focus on either legal guidance or emotional support, but rarely both. There is a lack of systems that provide a holistic, real-time framework combining crisis detection, emotion recognition, intent classification, and legal reasoning.

3.2 Insufficient Grounded Response Generation Using Retrieval-Augmented Generation (RAG)

While RAG frameworks enhance factual accuracy, most current systems rely on rule-based responses or static legal documents. Consequently, they lack dynamic retrieval of contextually relevant legal information, which increases the risk of hallucinated or inaccurate guidance in sensitive domains.

3.3 Inadequate Multi-Stage NLP Pipelines for Crisis Detection

Few systems implement multi-stage pipelines that first detect crises, determine user intent and emotional state, and then generate context-appropriate responses. Existing approaches often treat all user queries uniformly, limiting personalized intervention.

3.4 Limited Real-World Evaluation

Most studies demonstrate high accuracy in controlled datasets, but there is limited empirical evidence regarding long-term user trust, engagement, safety outcomes, or the impact of integrating emotional and legal support in real-world scenarios.

3.5 Domain-Specific Customization and Cultural Adaptation

Many deployed systems are designed for specific countries, limiting applicability in other legal and cultural contexts. There is a notable gap in chatbots tailored to Nepalese law, local crisis helplines, and culturally sensitive emotional support.

3.6 Privacy-Preserving and Trauma-Informed Design Limitations

Although privacy and safety are often discussed, few systems implement formal trauma-informed interactions and secure data handling mechanisms while maintaining effective context-aware guidance.

4. Methodology

4.1 System Architecture

This system is designed as a modular AI-powered chatbot that assists users by detecting crisis situations, understanding user intent and emotions, retrieving relevant legal information, and generating appropriate responses. The system combines transformer-based classifiers, a retrieval-augmented generation (RAG) module, and a large language model for response generation.

The overall workflow of the proposed system is illustrated in Figure 1.

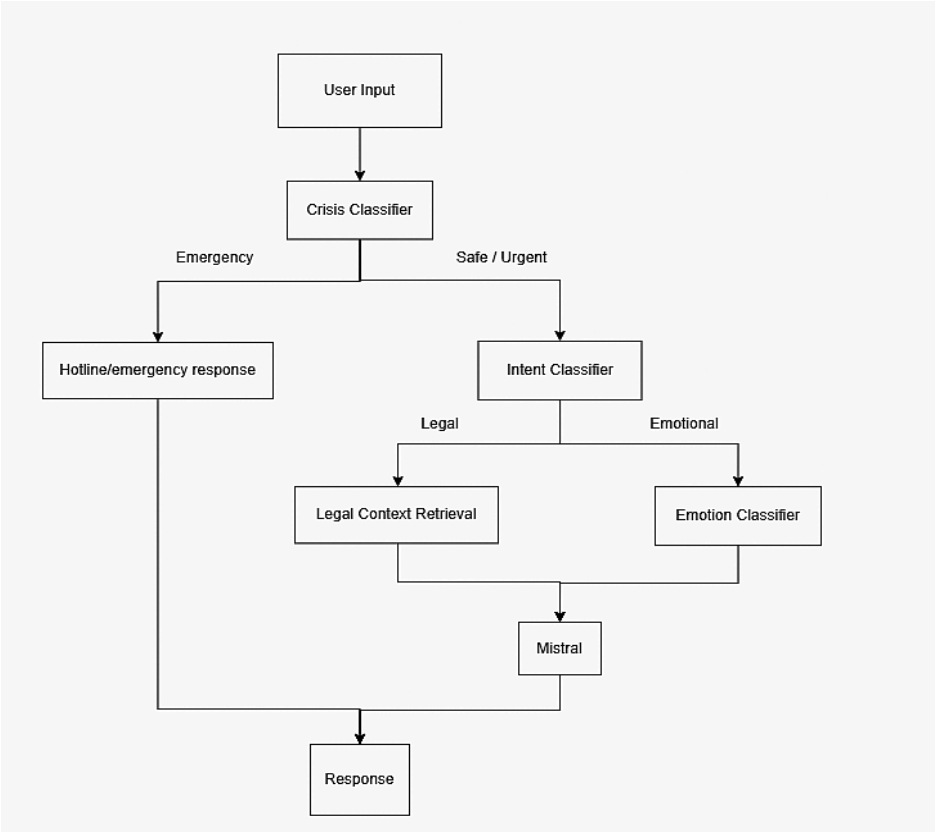

Figure 1. Overall Workflow of the System.

The above figure illustrates the flow of user query from input to final response generation. The system architecture prioritizes safety through an initial crisis classifier triggering emergency hotline responses. The intent classifier then separates non-emergency queries to the legal and emotional domains. Legal Context Retrieval as a legal domain and emotional classifier for psychological support then analyzes the input. Mistral LLM generates context rich response.

4.2 Data Collection and Dataset Preparation

The proposed hybrid system integrates heterogeneous textual data collected from multiple sources, including legal documents, crisis-related narratives, and conversational emotion datasets. The primary legal datasets was constructed from thirteen PDF documents containing case-relevant textual content. In addition, emotionally expressive conversational data was obtained from the GoEmotions dataset available on Kaggle, along with manually curated user queries reflecting real-world crisis scenarios.

Due to the unstructured nature of these sources, a multi-stage preprocessing pipeline was designed to convert raw textual data into structured and labeled formats suitable for supervised learning. This pipeline included text extraction from PDF documents, data cleaning through removal of noise and irrelevant symbols, normalization of text, and transformation into JSON-based representations compatible with transformer models.

A key design aspect of this research is the reuse of the same legal datasets for multiple classification tasks, enabling a unified learning framework across different objectives. The thirteen legal documents were annotated differently for crisis classification and intent classification tasks.

For the crisis classification task, the extracted textual data was manually labeled into three severity categories: Safe, Urgent, and Emergency. These labels were assigned based on the level of risk, urgency, and potential harm reflected in the text. The objective of this labeling scheme is to capture varying degrees of crisis severity relevant to real-world intervention systems.

For the intent classification task, the same textual data was synthetically labeled using rule-based coding techniques into two categories: Legal and Emotional. The labeling process was automated through predefined rules that identify linguistic patterns, domain-specific keywords, and contextual cues. This approach enabled scalable dataset generation while maintaining consistency in label assignment.

For emotion classification, the system utilizes the GoEmotions dataset, which contains fine-grained emotion annotations derived from conversational data. To align with the objectives of crisis analysis, these detailed emotion labels were mapped into a reduced set of core emotional states representing distress-related conditions such as sadness, fear, anger, and neutral states. This reduction helps improve generalization and reduces label sparsity.

After annotation, all datasets were converted into structured JSON format, where each instance consists of input text paired with its corresponding label. This format ensures compatibility with transformer-based architectures and facilitates efficient data handling during training and inference.

To address the input length limitations of transformer models, a chunking strategy was applied to long textual sequences. Given an input sequence T = {t_1, t_2, ..., t_n}, it is divided into overlapping subsequences T_i such that:

$$Ti = \{t_i, t_{i+1}, \ldots, t_{i+m}\}, Ti+1 \cap Ti \neq \emptyset \quad \text{(Equation 1)}$$

where m represents the chunk size. The overlap between consecutive chunks ensures preservation of contextual continuity, allowing the model to maintain semantic coherence across segment boundaries. This is particularly important for processing long legal documents and complex crisis narratives.

Table 1. Dataset Distribution

Dataset Type	Total PDFs	Total Chunks/Entries	Train (%)	Validation (%)	Test (%)	Chunk Size	Overlap
Crisis	13	6894	80	10	10	100 chars	0
Intent	13	6894	80	10	10	100 chars	0
RAG	11	2744	N/A	N/A	N/A	500 chars	100

4.3 Data Preprocessing and Representation

All textual inputs undergo a unified preprocessing pipeline before being fed into learning models. The pipeline includes:

Normalisation — conversion to lowercase; removal of URLs, special characters, punctuation, and null entries.

• Tokenisation — raw text is converted into subword units compatible with transformer architectures; padding and truncation enforce uniform input length.

• Embedding generation — each tokenised sequence is mapped into dense vector representations using pretrained embedding layers. For a token sequence $X = \{x_1, x_2, \ldots, x\}$, the embedding function maps it into:

$$E = \{e1, e2, \ldots, en\}, \quad ei \in \mathbb{R}d \qquad \text{(Equation 2)}$$

where d denotes the embedding dimension. These embeddings serve as input to transformer models, enabling them to capture contextual dependencies through attention mechanisms.

4.4 Model Training and Optimization

All classification models in the system are trained using supervised learning with cross-entropy loss. Given predicted probabilities ŷ and true labels y, the loss is computed as:

$$L = -\Sigma Ci = 1 \; yi \; log(\hat{y}i) \qquad \text{(Equation 3)}$$

To address class imbalance, particularly in crisis classification, a weighted variant is used:

$$L = -\Sigma Ci = 1 \; wi \; yi \; log(\hat{y}i) \qquad \text{(Equation 4)}$$

where w_i represents class-specific weights. Optimization is performed using the Adam optimizer, which updates model parameters using adaptive learning rates:

$$\theta t + 1 = \theta t - \eta \cdot mt / (\sqrt{v}t + \varepsilon) \qquad \text{(Equation 5)}$$

where η is the learning rate controlling the step size of updates. The batch size determines how many training samples are processed in each iteration, while the number of epochs defines how many complete passes are made over the dataset. These hyperparameters are empirically tuned based on validation performance to ensure optimal learning and generalization.

Once the classifiers are trained, they are used during inference to categorize incoming user queries, which are then routed to the appropriate processing pipeline.

Table 2: Model Hyperparameters

Component	Hyperparameter	Description	Value
Classifiers	Learning Rate	Controls step size during model training	2e-5
Classifiers	Batch Size	Number of samples processed per training step	8

Classifiers	Number of Epochs	Number of full passes through the dataset	5
Classifiers	Optimizer	Algorithm used for updating weights	AdamW
Classifiers	Loss Function	Measures prediction error during training	Cross-Entropy Loss
RAG System	Chunk Size	Number of characters per document chunk	500
RAG System	Chunk Overlap	Overlapping text between chunks to preserve context	100
RAG System	Top-k Retrieval	Number of most relevant documents retrieved	6
RAG System	Embedding Model	Model used to convert text into vector representations	BGE-small-en
RAG System	Similarity Metric	Method used to rank document similarity	L2 Distance
RAG System	Temperature	Controls randomness of generated responses	0.2
RAG System	Language Model	Model used for final response generation	Mistral-small-latest

4.5 User Input Processing

The system interaction begins when a user submits a textual query through the chatbot interface. This input is received by a FastAPI-based backend, which coordinates communication between all processing modules. Before the input is passed to the learning models, it undergoes preprocessing to ensure consistency and compatibility with transformer-based architectures.

A key challenge in using transformer models is their fixed input length constraint, typically limited to a few hundred tokens. Since real-world inputs and document-based datasets often exceed this limit, a chunking mechanism is applied. The input text is divided into smaller segments with controlled overlap, ensuring that contextual continuity is preserved across segment boundaries. This overlap plays an important role in maintaining semantic relationships between adjacent chunks, especially when critical information lies near chunk boundaries.

From a representation perspective, each chunk is tokenized and mapped into a sequence of embeddings. If a sentence is represented as a sequence of tokens $1, t2, \dots, tn$, it is transformed into embedding vectors $e1, e2, \dots, en$, where each embedding captures both lexical and contextual information. This transformation allows downstream models to operate on dense numerical representations rather than raw text, enabling efficient learning of semantic patterns.

Different chunk sizes are intentionally used across modules. Smaller chunks are preferred for classification tasks, where the goal is to capture concise signals such as distress keywords or intent indicators, while larger chunks are used in the retrieval module to preserve richer contextual information required for legal reasoning. This adaptive chunking strategy ensures that each component receives input in a form best suited to its objective.

4.6 Crisis Classification

The crisis classification module is responsible for identifying the severity of the user's situation, which directly influences response prioritization. A transformer-based model built on IndicBERT is used to capture contextual relationships in user input. The self-attention mechanism within the transformer computes interactions between tokens using:

$$Attention(Q, K, V) = softmax\left(QK_T/\sqrt{dk}\right)V$$

(Equation 6)

This allows the model to assign higher importance to words indicative of distress or urgency. The final hidden representation corresponding to the classification token is projected into a probability distribution over crisis classes using a softmax layer.

To enhance robustness, a hybrid classification framework is introduced. A rule-based layer detects explicit high-risk keywords, while the model-based layer captures implicit contextual signals. The final decision is obtained through priority-based logic, where rule-based outputs override model predictions in cases of explicit emergency indicators. This ensures that critical situations are not overlooked due to model uncertainty.

4.7 Intent Classification

The intent classification module determines the functional purpose of the query, enabling the system to distinguish between legal inquiries and emotional expressions. This module also utilizes IndicBERT to maintain consistency in contextual encoding across tasks.

Given an input representation h, the classifier computes class probabilities as:

$$P(y \mid h) = softmax(Wh + b)$$ (Equation 7)

Where W and b are learnable parameters. This probabilistic formulation allows the system to handle ambiguous inputs by considering confidence scores.

The separation of intent classification from crisis detection ensures that urgency and purpose are treated as independent dimensions, enabling more precise routing of queries to downstream modules.

4.7.1 Why we selected Indic BERT?

This chatbot is designed for users in Nepal, where people often write in a mix of languages, Roman (Nepali + English), all blended in the same message. As standard BERT models are trained mostly on English text, they struggle with this kind of multilingual, code-switched input. Indic BERT, on the other hand, was specifically trained on text from South-Asian languages, including Devnagari, and handles this mixed-language writing much more naturally. This system deals with sensitive topics like domestic violence, and misunderstanding a user's message due to language limitations is not just a technical failure; it could mean missing a cry for help. So, using a model that actually understands how our target users write was a deliberate and important design.

4.8 Emotion Classification

The emotion classification module captures the affective state of the user, which is essential for generating empathetic responses. Distil BERT is used for this task due to its reduced computational complexity and faster inference time.

The model encodes input text into contextual embeddings and predicts emotion categories through a classification layer. The predicted emotion influences response generation by modulating tone and language style. Unlike intent classification, which focuses on task identification, emotion classification operates at a finer granularity, capturing subtle variations in user expression.

4.7.1 Why we selected Distil BERT?

Emotion classification is just one of several tasks happening simultaneously when a user sends a message. The system also needs to detect crisis level, classify intent, and generate a response in real time. Distil BERT is a smaller, faster version of BERT that retains around 97% of its language understanding ability while being significantly lighter. Since emotion detection works with short conversational messages and predicts from a fixed set of emotion categories, a heavy model is not required here. Distil BERT gets the job done accurately without slowing the overall system down.

4.9 Retrieval-Augmented Generation (RAG)

The RAG module is responsible for generating factually grounded responses for legal queries. Rather than relying solely on the generative capability of a language model, this module retrieves contextually relevant information from a structured knowledge base and incorporates it into the response generation process, thereby minimizing hallucinated or memorized outputs.

4.9.1 Embedding Function

Legal documents are first segmented into chunks and transformed into dense vector representations using a sentence transformer model. Formally, each document chunk d_i and query x is mapped to a vector in a high-dimensional semantic space through an encoding function:

$$e = f(x), \quad e \in \mathbb{R}d \qquad \text{(Equation 8)}$$

where $f(.)$ denotes the sentence transformer encoding function and d is the dimensionality of the embedding space. These vector representations capture the semantic meaning of the text, enabling meaning-based retrieval rather than simple keyword matching. The resulting document embeddings are stored in a FAISS index, which supports efficient large-scale similarity search.

4.9.2 Similarity Computation

To identify relevant documents for a given user query, cosine similarity is computed between the query embedding q and each document embedding d_i in the knowledge base. Cosine similarity measures the angular distance between two vectors rather than their magnitude, making it well-suited for semantic comparison:

$$Sim(q, di) = (q \cdot di) / (\|q\| \|di\|) \qquad \text{(Equation 9)}$$

This metric ensures that documents with semantically similar content to the query receive higher similarity scores regardless of their length or term frequency. In practice, since FAISS operates on normalized embeddings, cosine similarity becomes equivalent to a simple inner product as shown in equation 9 which allows FAISS to perform computationally efficient nearest neighbor search at scale.

4.9.3 Top-k Retrieval

Rather than returning all documents from the knowledge base, the retrieval process selects only the most relevant document chunks. Specifically, the top-k documents with the highest similarity scores are selected through the following ranking operation:

$$Dk = arg\ topk\{di \in D\}\ Sim(q, di) \qquad \text{(Equation 10)}$$

where D is the complete document collection and Dk is the retrieved subset of k most relevant chunks. This ranking mechanism ensures that only the most contextually relevant content is passed forward, reducing noise and improving response quality.

4.9.4 Context Aggregation

The retrieved document set Dk is aggregated to form a unified contextual input C, defined as:

$$C = \bigcup_{di \in D} di \qquad \text{(Equation 11)}$$

This aggregated context C consolidates the most relevant legal information from multiple document chunks into a single coherent input. By combining evidence from multiple retrieved sources, the system achieves broader coverage of the query topic while maintaining factual grounding in actual legal content.

4.9.5 Response Generation

The aggregated context C is concatenated with the original user query q and passed to the language model as a structured prompt. The model then generates a response conditioned on both the query and the retrieved legal content, formally expressed as:

$$\hat{y} = LM(q, C) \quad \text{(Equation 12)}$$

where $\hat{y}$ is the generated response and $LM(\cdot)$ denotes the language model. This conditioning mechanism ensures that the generated response is grounded in retrieved legal documents rather than relying on the model's parametric memory alone, significantly reducing the risk of factually incorrect or hallucinated outputs.

4.10 Routing Logic

The routing mechanism acts as the decision layer that determines how each query should be processed based on the outputs of the classification modules. This component integrates the results of crisis, intent, and emotion analysis to dynamically select the appropriate response pathway.

If the intent is classified as legal, the query is routed to the RAG pipeline to ensure grounded response generation. If the intent is emotional or supportive, the query is directed to a conversational response module. Additionally, the crisis classification output influences the priority and tone of the response, ensuring that urgent cases are handled with appropriate seriousness.

This dynamic routing strategy enhances system flexibility and ensures that responses are tailored to the specific needs of the user.

4.11 Response Generation

The final response is generated using the Mistral AI API with the model *mistral-small-latest*. The generation process follows a conditional formulation where the output is produced based on the user query, the retrieved context (when available), and the system-level instructions:

$$\hat{y} = LM(q, C, I) \quad \text{(Equation 13)}$$

where $\hat{y}$ represents the generated response, q is the user query, C is the retrieved contextual information from the knowledge base, and I denotes the instruction prompt that guides the behaviour of the model.

Two distinct response modes are implemented during chat generation. For emotional or support-related messages, the system generates responses using a trauma-informed conversational prompt. In this case, the context C is absent, and the model relies on the query and empathetic instructions, i.e., $\hat{y} = LM(q, I)$. The prompt is designed to ensure that the model uses empathetic language, acknowledges the user's feelings, and provides supportive suggestions.

For legal queries, a grounded response mechanism is used. The retrieved RAG context is incorporated into the prompt, and the model generates responses conditioned on both the query and the contextual evidence, i.e., $\hat{y} = LM(q, C, I)$. The instruction prompt explicitly guides the model to rely only on the provided legal context, avoid unsupported legal claims, present answers in a structured manner (such as numbered steps), and maintain a supportive tone. This approach ensures that the generated legal responses remain contextually grounded, accurate, and responsible.

4.11.1 Why we selected Mistral?

Mistral is used as the final response generation model because the BERT-based models used elsewhere in the system (Indic BERT, Distil BERT) are encoder-only architectures designed for classification, not text generation. Once the classifiers and RAG module have done their work, a decoder-based language model is required; thus, Mistral is perfect for it. Mistral is accessed via API, which avoids the need to host a large model locally. It is also significantly cheaper than alternatives like GPT-4 while still following structured

system prompts reliably, which matters because the system needs the model to stay grounded in retrieved legal content and not produce unsupported claims.

4.12 System Implementation

The system is implemented using a modular architecture, with Fast API serving as the central orchestration layer. Each component, including classifiers and retrieval modules, operates as an independent service, allowing for scalability and ease of maintenance. SQLite is used to store user interactions and system responses, enabling tracking and future analysis.

This design ensures that the system can be extended with additional features, such as multilingual support or improved retrieval mechanisms, without requiring major architectural changes.

5. EVALUATION METRICS

5.1 Classifier Training Results

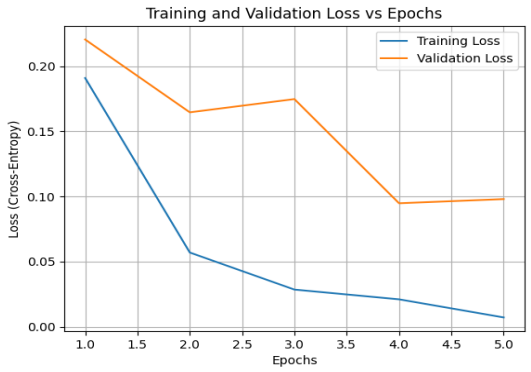

Figure 2: Training and Validation Loss for Crisis Classifier

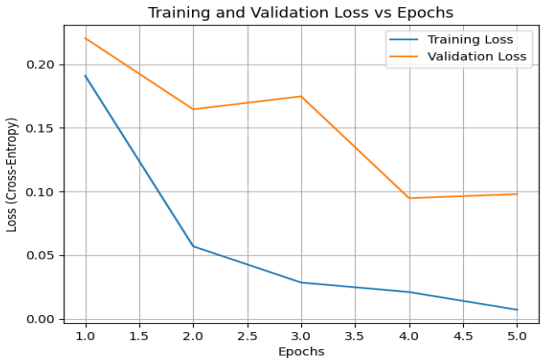

Figure 3: Training and Validation Loss for Intent Classifier

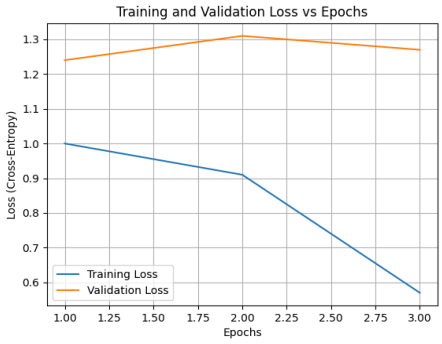

Figure 4: Training and Validation Loss for Emotion Classifier

Figures 2 and 3 illustrate the training and validation losses for the crisis classifier and intent classifier over five epochs. In the early epochs, both training and validation losses are elevated, reflecting the models' initial state prior to convergence. As training progresses, the training loss decreases substantially from the first epoch onward, indicating effective learning from the dataset. The validation loss also decreases gradually, with a minor fluctuation at the third epoch that suggests a brief period of overfitting. The validation loss then continues to decrease, indicating that the models recover and generalise reasonably well.

Figure 4 depicts the loss curves for the emotion classifier over three epochs using the cross-entropy loss function. The training loss decreases consistently across all epochs, indicating effective learning from the training data. In contrast, the validation loss remains elevated and increases slightly in the early epochs before

eventually declining. This divergence between training and validation loss indicates overfitting, whereby the model fits the training data closely but has limited generalisation to unseen samples. This behaviour is expected given the class imbalance and semantic overlap among the seven emotion categories, and serves as motivation for further regularisation and data augmentation strategies in future work.

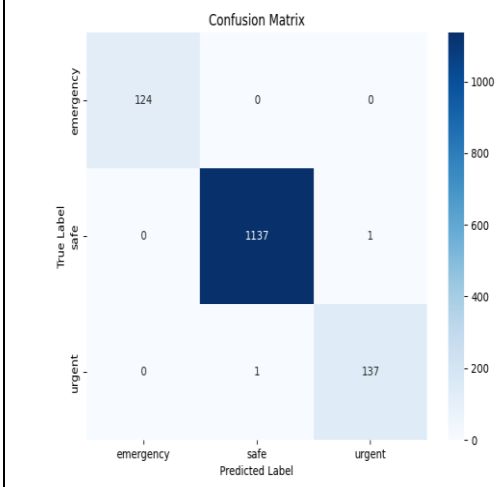

Figure 5: Confusion matrix of Crisis classifier

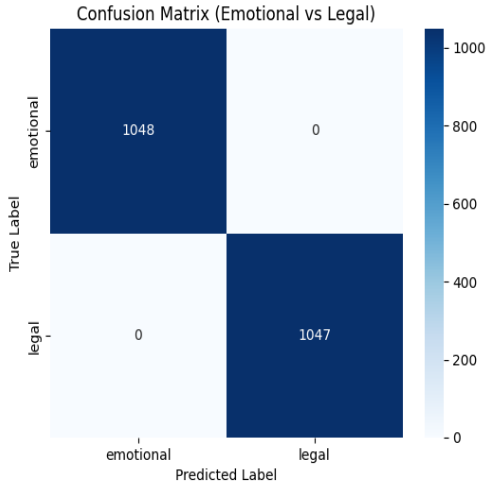

Figure 6: Confusion matrix of Intent classifier

5.2 Crisis Classifier Evaluation

The performance of the crisis classifier is evaluated using a confusion matrix, which compares predicted labels with actual labels.

The confusion matrix for the crisis classifier (Figure 5) shows strong diagonal clustering, indicating reliable per-class predictions. Specifically, 37 out of 43 emergency cases and 82 out of 91 safe cases were correctly classified. Misclassifications are primarily concentrated between the ‘urgent’ and ‘emergency’ categories, which is expected given their semantic proximity. Importantly, the false-negative rate for the critical ‘emergency’ class remains low, indicating that the model is appropriately sensitive to high-distress language patterns.

5.3 Intent Classifier Evaluation

The intent classifier is evaluated using a confusion matrix that represents the relationship between predicted intents and actual intents.

The confusion matrix for the intent classifier (Figure 6) demonstrates strong separation between the two classes, with only two misclassified emotional queries. Nineteen false positives for the emotional class suggest some overlap in linguistic patterns between legal and emotional queries; however, the strong diagonal confirms the model’s effectiveness in correctly routing queries to the appropriate downstream pipeline.

5.4 Emotional Classifier Evaluation

The emotional classifier is evaluated using a confusion matrix to analyze how well the system detects different emotional states in user queries.

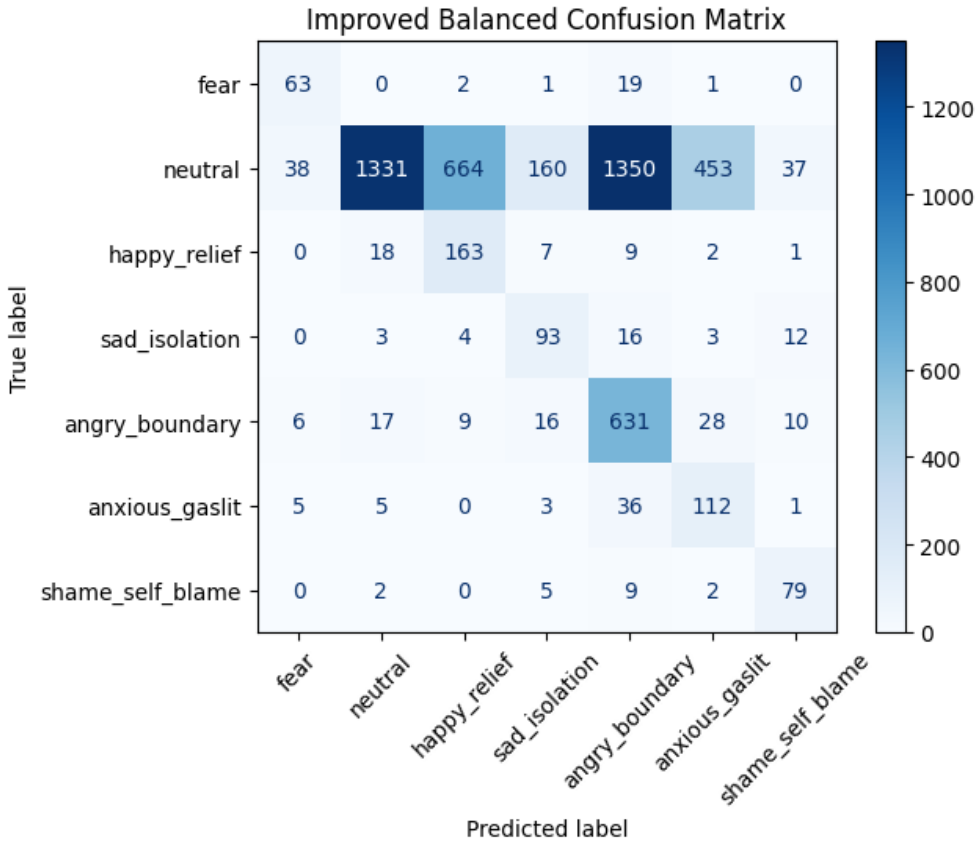

Figure 7. Confusion Matrix of Emotional Classifier

The emotional classifier model shows robust diagonal performance for niche psychological categories such as ‘shame_self_blame’ and ‘anxious_gaslit’. The ‘neutral’ class shows overlap with ‘angry_boundary’ and ‘anxious_gaslit’ as the primary classification challenge. This is consistent with the safety-first design principle of the system: it is preferable to flag a neutral message as potentially distressed than to overlook genuine distress. The high recall on negative emotional classes is therefore a deliberate and desirable property for a crisis-support chatbot, where the cost of a false negative significantly outweighs the cost of a false positive.

5.5 RAG System Evaluation

The RAG-based chatbot is evaluated using the RAGAS framework across 20 test queries drawn from the legal domain. It is acknowledged that this sample size is relatively small, and a larger and more diverse query set would provide stronger generalization evidence. Additionally, while BLEU score is commonly reported for generation tasks, it is not an ideal metric for RAG-based open-ended legal responses, since BLEU measures n-gram overlap rather than semantic correctness or helpfulness. Accordingly, RAGAS metrics, which directly assess faithfulness, relevance, and grounding, are more appropriate for this evaluation context and are used as the primary measures.

Table 3: RAG System Performance

Metric	Mean	Std	Min	Max
Faithfulness	0.71	0.08	0.55	0.85
Answer Relevance	0.68	0.07	0.50	0.82
Context Relevance	0.82	0.03	0.74	0.89
Context Utilization	0.75	0.04	0.65	0.82
Retrieval Score	0.75	0.00	0.75	0.75
Overall Score	0.73	0.03	0.66	0.80

Table 3 presents the RAGAS evaluation results across multiple metrics. Faithfulness (0.71) indicates that responses are largely grounded in the retrieved context, with reduced hallucination. Answer Relevance (0.68) shows that most responses adequately address the user's query, though some responses may include partially irrelevant information. Context Relevance (0.82) demonstrates strong retrieval performance in surfacing pertinent documents. Context Utilization (0.75) reflects effective use of the retrieved information during generation, with room for further improvement. The Retrieval Score is stable at 0.75 across all evaluated samples. Overall, the system achieves a mean score of 0.73, reflecting a balanced profile across retrieval and generation quality. These results are encouraging given the small evaluation set of 20 queries; expanding this evaluation with a larger and more diverse query set, and incorporating human judgement on correctness and safety, remains an important direction for future work.

5.6 Precision, Recall, and F1-Score Results

Table 4: Precision, Recall and F1-Score of Crisis Classifier

Class	Precision	Recall	F1-Score	Support
Emergency	0.970000	0.9700000	0.9744	626
Safe	0.9941	0.9883	0.9912	5741
Urgent	0.9533	0.9808	0.9668	631
Accuracy			0.9986	6998
Macro Avg	0.9725	0.9797	0.9775	6998
Weighted Avg	0.9986	0.9986	0.9986	6998

The crisis classification model achieved strong performance on the test set, with an overall accuracy of 99.86%. The macro-average precision (0.9725), recall (0.9797), and F1-score (0.9775) indicate balanced performance across all classes, while the weighted-average scores (0.9986) reflect high overall effectiveness influenced by class distribution. At the class level, the model obtained F1-scores of 0.974 (Emergency), 0.991 (Safe), and 0.967 (Urgent), demonstrating reliable classification across categories. These results show that the model correctly classified most test instances; however, they were obtained on a relatively small, synthetically generated dataset with similar training and test patterns. This likely simplified the task and contributed to the high scores. Therefore, although the results are promising, further validation on real-world crisis data is necessary to confirm the model's generalization ability in diverse and low-resource settings.

Table 5: Precision, Recall and F1-Score of Intent Classifier

Class	Precision	Recall	F1-Score	Support
Emotional	0.9984	0.9984	0.9984	5697
Legal	0.9984	0.9984	0.9984	5697
Accuracy			0.9984	11394
Macro Avg	0.9984	0.9984	0.9984	11394
Weighted Avg	0.9984	0.9984	0.9984	11394

The intent classifier demonstrated highly consistent and reliable performance across both classes, benefiting from a perfectly balanced dataset of 5,697 samples per class. Both the Legal and Emotional categories achieved identical precision, recall, and F1-scores of 0.9984, reflecting the model's ability to clearly

distinguish between the two types without bias toward either class. The macro and weighted averages are equivalently 0.9984, and the overall accuracy of 99.84% across 11,394 test samples further validates the model's effectiveness on this dataset. The symmetry in performance metrics across both classes suggests well-separated decision boundaries, consistent with the model's rapid convergence during training. However, these results should be interpreted with caution. The labels were assigned using rule-based automated annotation on legal text, which may introduce systematic patterns that simplify the classification task beyond what would be observed in organic user queries. Cross-validation on independently sourced conversational data is recommended to confirm these results.

Table 6: Precision, Recall and F1-Score of Emotion Classifier

Class	Precision	Recall	F1-Score	Support
Fear	0.55	0.69	0.61	86
Neutral	0.95	0.51	0.66	4033
Happy_Relief	0.28	0.73	0.40	200
Sad_Isolation	0.33	0.71	0.45	131
Angry_Boundary	0.34	0.84	0.48	717
Anxious_Gaslit	0.22	0.62	0.32	162
Shame_Self_Blame	0.57	0.68	0.62	97
Accuracy			0.57	5426
Macro Avg	0.46	0.68	0.51	5426
Weighted Avg	0.79	0.57	0.61	5426

The emotion classifier achieves an overall accuracy of 57% across seven fine-grained psychological categories. The model exhibits high recall on negative emotional states, such as Angry Boundary (0.84), and lower recall on Neutral (0.51). In a crisis support context, this asymmetry is intentional and desirable. It is far safer to flag a neutral message as potentially distressed than to overlook genuine distress as neutral. The model's intention toward negative class detection aligns with the safety-first design principle, where the cost of a false negative significantly outweighs the cost of a false positive. This result, therefore, reflects a conscious trade-off favoring user safety over raw accuracy.

6. Discussion

The proposed system demonstrates satisfactory performance in supporting users experiencing emotional distress and providing legal guidance. The intent classifier accurately distinguishes between legal and emotional queries, while the emotion recognition module effectively identifies key emotional states. The integration of the RAG framework allows the system to provide responses grounded in verified legal documents, reducing misinformation and improving trustworthiness.

The hybrid architecture combining BERT-based classification with RAG retrieval shows several advantages. It enables the system to quickly identify user needs, deliver empathetic responses, and provide factually accurate legal guidance and sympathetic emotional response. The modular design also supports future enhancements, such as adding more legal resources or improving multilingual support.

However, some limitations were observed. Ambiguous or complex queries occasionally reduce classification accuracy, and subtle mixed emotions may be misinterpreted. The effectiveness of RAG-based responses depends on the quality and completeness of the underlying document corpus. Additionally, embedding and retrieval computations may affect response times in resource-limited environments.

Although the classification models demonstrate high performance on the constructed dataset, these results should be interpreted with caution. The dataset includes synthetically labeled samples and shares structural similarity between training and testing splits, which may simplify the classification task. Therefore, the reported performance may not fully reflect real-world complexity. Future work should incorporate cross-validation, external datasets, and real-world user data to rigorously evaluate generalization capability.

Overall, the results indicate that a hybrid approach can balance empathy and factual correctness, offering meaningful support to users while minimizing risks associated with generative AI. Future work can focus on expanding the legal corpus, refining emotion detection, and optimizing system performance for real-time deployment.

Ethical and safety considerations are central to the design of chatbot given the sensitive nature of its target domain. The system handles disclosures related to domestic violence, which requires careful attention to user privacy, data security, and the avoidance of harm. All user interactions are processed locally without persistent storage of personal content beyond the session. The RAG module is constrained to retrieve only from curated legal sources, reducing the risk of generating harmful or unsupported legal advice. The emotion and crisis classifiers are designed with a safety-first bias, favouring high recall on distress categories to minimise the risk of failing to detect a user in crisis. The system does not attempt to replace professional legal counsel or mental health support, and responses consistently direct users towards relevant helplines and support organisations. Furthermore, the trauma-informed conversational prompts used during emotional interactions follow established principles of non-judgmental, empathetic communication to avoid re-traumatisation. Future deployments should include formal safety audits and user testing with domain experts such as social workers and legal professionals before any live release.

7. Conclusion

This paper presented a hybrid NLP-based chatbot designed to provide legal and emotional support to survivors of domestic violence. The system integrates a multi-stage classification pipeline comprising crisis, intent, and emotion classifiers built on IndicBERT and DistilBERT, coupled with a Retrieval-Augmented Generation module that grounds legal responses in curated statutory documents. The crisis and intent classifiers achieved high performance on the test set; however, these results were obtained on a synthetically annotated dataset derived from legal text, and further validation on real-world data is required before conclusions about generalizability can be drawn. The emotion classifier achieved 57% accuracy across seven fine-grained psychological categories, a result that reflects a deliberate safety-first bias toward high recall on negative emotional states rather than raw accuracy. The RAG system demonstrated sound retrieval and grounding performance, with a RAGAS overall score of 0.73. The modular architecture ensures that each component can be independently improved or replaced, supporting ongoing development. Future work will focus on expanding and diversifying the annotated dataset, incorporating human evaluation of response quality, and conducting real-world user studies with domain experts to validate safety and effectiveness prior to deployment.

Acknowledgement

We would like to thank Er.Nirajan Acharya, our supervisor for his guidance and support throughout the research.

References

[1] "The Importance of Language in Domestic," Asian Pacific Institute on Gender-Based Violence, 2017.

[2] "Violence against women: Intimate partner and sexual violence," World Health Organization, 2012.

[3] X. F. et al, "The Effectiveness of AI Chatbots in Alleviating Mental Distress and Promoting Health Behaviors Among Adolescents and Young Adults: SystematicReview and Meta-Analysis," 2025.

[4] T. et.al, "#MeTooMaastricht: Building a Chatbot to Assist Survivors of Sexual Harassment.," 2019.

[5] S. et.al, "Designing Chatbots to Support Victims and Survivors of Domestic Abuse," 27 Feb 2024.

[6] J.-W. et.al, "Recognizing the role of victim supports in building and maintaining healthy and safe communities," Community Safety and Well-Being, vol. 1(2), pp. 12-15, 2016.

[7] V. et.al, "LAW-U: Legal Guidance Through Artificial Intelligence Chatbot for Sexual Violence Victims and Survivors," vol. 9, pp. 131440-131431, 2021.

[8] "rAInbow:Chatbot to support Victims of Domestic Abuse," World Justice Project , 2021.

[9] "Spring Act," Sophia-The Chatbot, 2021.