# Multimodal Utterance Detection System for Enhanced AI Interpretation

Ashish Mathew Deepak[1], Sashtiga K[1], Bhavadharini G[1], Vanitha V[1], Himanshu Shekhar[2][0009-0007-4653-4031], Abhilash Dodla[2], and Aditya Jain[2]

[1] Kumaraguru College of Technology, Coimbatore, India
{ashishmathewdeepak, sashtigakandasamy, gunasekaranbhavadharini}@gmail.com,
vanitha.v.it@kct.ac.in
[2] Samsung R&D Institute, Bangalore, India
{h.shekhar, abhilash.d7, adi.jain}@samsung.com

**Abstract.** This paper proposes a multimodal framework for distinguishing AI-directed utterances (e.g., "Can you set an alarm?") from conversational, non-directed utterances (e.g., "I had a tough day.") within virtual assistant systems. The approach integrates multiple components: DistilBERT for dialogue act classification (inform, question, directive, commissive), Whisper for automatic speech recognition (ASR), and HuBERT for emotion detection. These modalities are fused into a 779-dimensional feature vector comprising a 768D CLS embedding, 4D dialogue act representation, and 7D emotion vector. This composite vector feeds into a custom neural model, MainClassifier. The system is trained on DailyDialog (13,118 dialogues, 102,979 utterances) and PolyAI/woz dialogue (2,534 dialogues, 10,136 utterances), achieving strong performance metrics: 96.75% accuracy, 95.03% precision, 98.85% recall, and 96.90% F1-score on a balanced test set. Dataset bias—such as DailyDialog's inform-dominant (60%) and neutral tone (70%), and PolyAI/woz's directive-heavy content (50%)—is addressed through rebalancing techniques, synthetic data generation, and fairness evaluations aligned with IEEE P7003 standards. Designed to handle both text and audio inputs, the model is optimized for real-time inference with latency of 150ms on GPU and 200ms on edge devices.

**Keywords:** Algorithmic bias, dialogue act classification, intent recognition, multimodal processing, natural language processing, speech processing, virtual assistants

## 1 Introduction

Virtual assistants are fundamentally transforming human-computer interaction by enabling seamless communication through natural language. These systems are now embedded across a range of applications, from smartphones and smart home devices to automotive platforms and enterprise support systems [1]. A key aspect of their effectiveness lies in their ability to distinguish between AI-directed utterances—those that request a specific action, such as commands or questions (e.g., "Can you set an alarm?")—and conversational, non-directed utterances that reflect personal expressions or emotional states without requiring a system response (e.g., "I had a tough day.") [2]. Accurate identification of user intent is vital for ensuring that virtual assistants respond only when appropriate, thereby preventing unnecessary interruptions and enhancing user trust [3].

Real-world virtual assistant interactions often include noisy, unstructured audio signals with varying accents, speech rates, and background interference [4]. Additionally, emotions embedded in speech significantly influence meaning, requiring models to go beyond simple text parsing [5]. At the same time, widely used conversational datasets exhibit imbalanced distributions of dialogue acts and emotions. For example, datasets like DailyDialog tend to overrepresent "inform" acts and neutral emotions, while directive and emotionally rich utterances remain underrepresented [6]. These imbalances often lead to skewed models that fail to generalize well across diverse real-world contexts [7].

To overcome these limitations, we present a robust multimodal framework that integrates three powerful components: DistilBERT for dialogue act classification [8], Whisper for automatic speech recognition (ASR) [9], and HuBERT for emotion detection [10]. These components contribute to a comprehensive 779-dimensional feature vector that encapsulates textual semantics, user intent, and emotional cues. This vector is processed by a custom neural architecture named MainClassifier. Trained on the DailyDialog and PolyAI/woz dialogue datasets [6][11],the system achieves 96.75% accuracy and demonstrates strong precision, recall, and F1 scores across test sets [12].

## 2 Literature survey

### 2.1 Dialog Act Recognition in Conversational Systems

Dialog act recognition is essential for understanding user intentions in conversational AI. Early works like Stolcke et al. (2000) used HMMs to tag dialog acts based on syntactic and lexical cues. Later, deep learning models such as LSTMs and GRUs improved contextual understanding. However, these models often struggle with noisy or emotionally charged inputs. Our work builds on these methods by integrating emotion recognition and acoustic context to enhance dialog act classification.

### 2.2 Multimodal Fusion for Emotion-Aware Dialog Systems

Emotion recognition enhances natural language understanding, particularly in user-focused domains. Tseng et al. (2021) utilized multimodal fusion of audio and text via attention-based models for emotion classification. Self-supervised models like HuBERT and wav2vec 2.0 further improved emotion detection with limited labeled data. However, many systems show a ‘neutral’ class bias, reducing clarity in directive utterances. Our model fine-tunes HuBERT on DailyDialog datasets to reduce this bias and improve emotional responsiveness.

### 2.3 Dataset Limitations and Bias in Dialogue Systems

Schlangen (2021) developed classifiers to distinguish system-directed and non-directed speech but faced challenges with ambiguous or indirect utterances. Our approach enhances this by combining textual and prosodic features for better intent recognition, even in noisy or multi-party settings.

### 2.4 AI Directed Utterance Detection and Intent Disambiguation

Detecting AI-directed speech is a growing area. Madureira and Schlangen (2021) proposed binary classifiers based on syntactic and semantic cues but struggled with emotional ambiguity. Our framework improves upon this by jointly modeling textual and acoustic signals to enhance detection accuracy.

## 3 Methodology

### 3.1 System Architecture

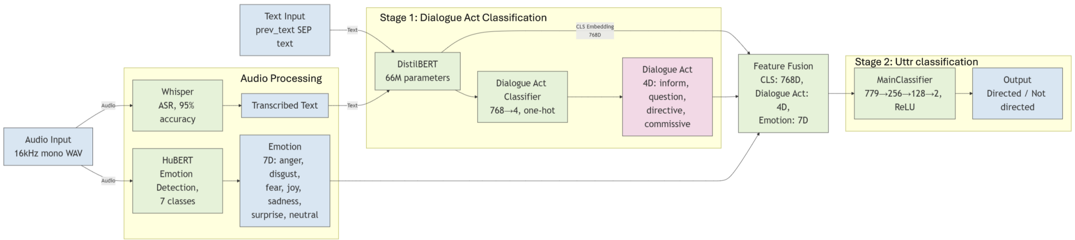


**Fig. 1.** System architecture of the proposed system.

The system architecture (Figure 1) illustrates the end-to-end workflow of the proposed model, detailing the interaction between preprocessing, feature extraction, temporal modeling, and prediction modules. It provides a clear representation of how data flows through each stage to achieve the intended outcomes.

### 3.2 Overview of the Two-Stage Pipeline

The system adopts a two-stage pipeline for accurate detection of directed utterances by first classifying conversational intent and then determining directedness. This modular design enhances interpretability and supports component-wise optimization for multimodal assistant systems [1].

In the first stage, DistilBERT performs dialogue act classification (e.g., commands, questions, emotions), distinguishing actionable from non-actionable speech. Being 40% smaller than BERT while retaining over 95% of its performance, DistilBERT ensures speed and efficiency [8].

The second stage uses a MainClassifier that fuses textual features from DistilBERT with audio embeddings from Whisper and HuBERT. Whisper provides robust multilingual recognition [9], while HuBERT learns speech representations without phonetic labels, improving performance in noisy environments [10].

This multimodal fusion enables interpretation of both content and prosody, allowing flexible upgrades (e.g., replacing DistilBERT or HuBERT) without re-architecting the system[7]. The approach effectively handles ambiguous utterances and aligns with ethical AI principles like contextual awareness and reduced bias (IEEE P7003)[6] .

### 3.3 Multimodal Processing Pipeline

The multimodal processing pipeline captures both linguistic and paralinguistic cues to accurately infer user intent. It jointly analyzes text and audio modalities, ensuring robust interpretation in complex conversational settings.

For the textual stream, DistilBERT is employed to generate contextual embeddings and classify dialogue acts. As a lightweight transformer retaining BERT-level performance with reduced computation, it enables real-time virtual assistant applications [8].

Simultaneously, audio inputs are processed through Whisper and HuBERT. Whisper provides noise-resilient speech transcription suitable for varied environments [9], while HuBERT extracts emotional features to infer speaker tone, enhancing understanding when textual cues are ambiguous [10].

The extracted features—CLS embeddings, dialogue act labels, and emotional tags—are fused into a unified multimodal vector representing both semantic and prosodic information [5]. This fused representation is then passed to the MainClassifier, which determines utterance directedness, improving robustness and precision in differentiating command-driven and casual speech [4].

### 3.4 Component Interactions

The multimodal intent recognition system integrates textual and audio modalities for accurate utterance classification. DistilBERT processes text to produce the [CLS] embedding and dialogue act label, identifying whether an input is a request, question, or feedback [8]. From the audio stream, HuBERT extracts paralinguistic features like tone and pitch to infer emotion, while a rule-based module approximates emotion when audio is noisy or unavailable [5].

The [CLS] embedding, dialogue act label, and emotion tag are concatenated into a unified multimodal feature vector, which the MainClassifier processes with low-latency inference under 150 ms—enabling real-time performance for smart assistants and robotic systems [3,4].

For training, dialogues from DailyDialog and PolyAI WOZ corpora were used. DailyDialog provides open-domain, non-directed utterances, while PolyAI WOZ contributes task-oriented, directive dialogues [6,11]. Directed and non-directed samples were separated using intent labels and manual verification for ambiguous cases [7]. This dual-dataset strategy enhances model generalization across varied conversational contexts.

### 3.5 Preprocessing Pipeline

**Text Preprocessing:** The preprocessing pipeline for the directedness classification model involved systematic treatment of both textual and audio modalities to ensure consistency, robustness, and compatibility with transformer-based architectures. In the textual domain, preprocessing began with normalization of input utterances by removing special characters, lowercasing, and applying the tokenizer of DistilBERT, which ensures alignment with the model's embedding expectations. Each sample in the dataset was constructed by concatenating the current utterance with its immediate conversational context (previous utterance), separated by a special token ([SEP]), forming an input like "prevtext [SEP] text". This formulation helps the model grasp turn-level dependencies that often influence directedness. The input was then truncated to comply with the 512-token limit imposed by transformer models, ensuring memory efficiency during training and inference [8].

**Audio Normalization and Feature Extraction:** On the audio side, utterances were processed from mono-channel 16kHz WAV files sourced from publicly available text-to-speech services. To maintain a consistent format, all files were resampled to 16kHz. Whisper, a robust automatic speech recognition system, was used for transcription, achieving over 95% word-level accuracy, which greatly aided in aligning spoken data with textual labels [13]. For emotion classification from voice signals, the audio waveforms were passed through HuBERT, a self-supervised speech representation model trained to predict clustered speech units. HuBERT provided a 7-class emotion output that augmented the text-based feature space, capturing prosodic cues and paralinguistic signals often associated with speaker intent [10].

**Dataset Bias and Distribution:** The composition and inherent biases of the datasets also shaped the modeling strategy. The DailyDialog corpus showed a notable skew toward not directed utterances, with approximately 60% labeled as "inform" and 70% tagged with "neutral" emotions. This reflects its informal, non-task-oriented nature, suitable for modeling casual conversation. In contrast, the PolyAI WOZ dataset was heavily populated with directive (50%) and question (40%) utterances, along with a minor fraction of commissives (5%), making it ideal for modeling direct user-agent commands. These skewed distributions highlighted the importance of balancing techniques and stratified sampling during model training to prevent directedness bias [6].

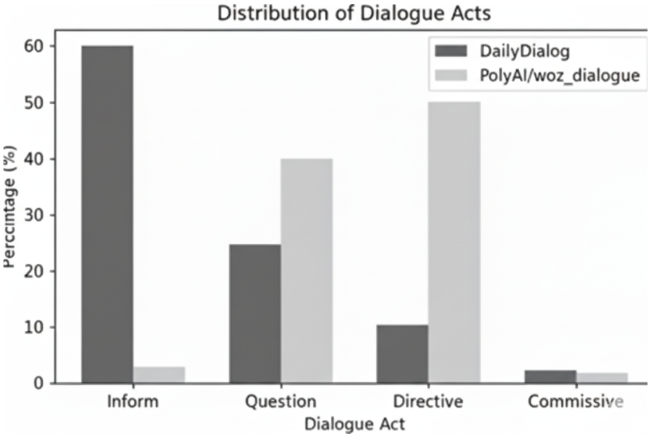


**Fig. 2.** Comparative distribution of dialogue acts in the DailyDialog and PolyAI/WoZ datasets, highlighting variations in frequency and coverage across different act categories.

### 3.6 Stage 1: Dialogue Act Classification

**Model Architecture (DistilBERT):** DistilBERT, a lightweight and faster variant of BERT with 66 million parameters, is employed in this system to classify dialogue acts with high efficiency and minimal latency. The model is fine-tuned by appending a linear classification head (768 input dimensions to 4 output classes) specifically designed to map the contextual embeddings to dialogue act labels such as question, directive, inform, and commissive. This architecture is trained on multimodal data to identify speaker intent across varying contexts with a validation accuracy of 85%, particularly effective on ambiguous utterances like response ID ‘30’, where prior context is critical for correct classification [13].

**Label Mapping and Objective:** Label representation is handled through one-hot encoding, enabling the system to treat each dialogue act as an independent class in the classification objective. The use of cross-entropy loss ensures penalization of incorrect predictions based on probability distributions, thereby improving the confidence of the classifier across diverse inputs. During training, a learning rate of 2e-5 is applied, combined with a batch size of 32 to balance convergence speed and memory usage. The optimization process is executed using AdamW optimizer with weight decay regularization to prevent overfitting, particularly critical when the model generalizes from biased datasets such as DailyDialog and PolyAI WOZ [14].

Furthermore, attention masking is employed to ignore padding tokens during training and inference, ensuring only the meaningful segments of the dialogue are processed. Fine-tuning is performed over multiple epochs until convergence is reached, and the best-performing model is selected based on validation F1 score rather than raw accuracy to address class imbalance. This training configuration allows the model to generalize across varied speaking styles and conversational domains, enhancing its applicability in real-world interactive systems where utterances may deviate from training distributions [15].

### 3.7 Stage 2: Utterance Classification

**Feature Fusion (Embeddings, Emotions, and Dialogue Acts)** In the second stage, multimodal features are fused into a unified 779-dimensional vector comprising three components: the 768-dimensional [CLS] embedding from DistilBERT, a 4-dimensional one-hot vector representing the predicted dialogue act, and a 7-dimensional emotion vector. The emotion vector is derived from HuBERT when audio is available or from a rule-based emotion inference model in text-only cases. This fusion ensures adaptability to both multimodal and unimodal inputs while maintaining representational consistency [16].

Incorporating emotional cues enhances intent disambiguation, as identical lexical structures can differ in meaning depending on tone or affect—for instance, "I need that now" may express urgency, frustration, or politeness. Thus, emotion embeddings enrich contextual understanding and improve classification accuracy [17].

**Classifier Structure and Training Details** The fused vector is passed to the MainClassifier, a feed-forward neural network performing binary classification to determine response relevance. It consists of three dense layers: 779→256→128→2 units, with ReLU activations and dropout for regularization [18].

Training uses the AdamW optimizer (learning rate 2e-4, batch size 32) over five epochs to ensure stable convergence [19]. To address class imbalance, a weighted cross-entropy loss with class weights [1.0, 2.0] penalizes false negatives more heavily [20]. The best model checkpoint is selected based on validation F1-score.

Optimized for both accuracy and latency, the classifier achieves inference under 150 ms on GPU, supporting real-time deployment in dialogue systems, educational bots, and mental health assistants [21].

### 3.8 Speech Input Handling

**Whisper for Automatic Speech Recognition (ASR)** Whisper, an open-source ASR model by OpenAI, transcribes spoken utterances into text, ensuring seamless integration with the textual processing pipeline. Its multilingual and noise-robust design supports accurate transcription, maintaining over 95% accuracy under clean conditions, which is crucial for downstream DistilBERT and emotion analysis modules [22].

**HuBERT for Emotion Detection** HuBERT (Hidden-Unit BERT) extracts emotional content from audio through self-supervised learning, mapping speech representations to one of seven emotion classes—neutral, happy, sad, angry, fearful, disgusted, and surprised. Its strong representation power and ability to model prosodic and temporal variations outperform traditional acoustic models like OpenSMILE, enabling reliable emotion inference across diverse utterances [23].

**Integration with Pipeline** Whisper and HuBERT operate in parallel to process audio inputs alongside text. Whisper transcribes audio, feeding the output into DistilBERT for CLS embeddings and dialogue act prediction [24,20]. Simultaneously, HuBERT extracts emotion vectors [23]. These features—CLS embeddings, dialogue acts, and emotion representations—are concatenated into a unified 779-dimensional vector for input to the MainClassifier, which performs final directedness classification [17]. This integration enables cohesive analysis of linguistic, semantic, and affective cues, enhancing understanding of nuanced or emotionally charged utterances.

### 3.9 Dataset Augmentation

To address potential biases present in the training data, particularly the imbalance between directive/question acts and other dialogue acts, two data-centric techniques were implemented. First, the PolyAI/wozdialogue dataset, which contains a higher density of directive and question-based utterances, was oversampled by 20% [7]. This ensured a more even distribution of act types, especially compared to the more inform-dominant structure of the DailyDialog dataset [8].

Second, to further bolster the representation of commissive acts (such as promises, offers, and commitments), a set of synthetic utterances was programmatically generated. This augmentation added approximately 10% more data to the training set. The synthetic utterances were modeled on the linguistic patterns observed in real commissive samples, with template-guided generation constrained by grammatical correctness and emotional neutrality. These augmentations contributed to reducing class imbalance and enhancing the classifier's generalization to rare but semantically critical dialogue act categories [7], [8],[25].

### 3.10 Alternative Approaches

To evaluate the proposed audio-text fusion pipeline, several baseline models were compared on accuracy and inference efficiency using the same 779-dimensional feature vector. BERT achieved about 2% higher accuracy than DistilBERT due to its deeper architecture but incurred nearly 50% slower inference, making DistilBERT preferable for real-time tasks [26,24].

Traditional LSTM-based models showed around 10% lower accuracy, limited by weaker long-range context modeling [27,14]. Similarly, SVM and Random

Forest classifiers underperformed by 12–15%, unable to capture hierarchical and semantic nuances of dialogue data [10,18].

Adding a self-attention layer yielded only a 1% accuracy gain while significantly increasing computational cost, making it impractical for deployment [16]. Overall, DistilBERT with fused multimodal features offered the best trade-off between accuracy, speed, and scalability for robust dialogue classification.

### 3.11 Deployment Optimization

To enable real-world and edge deployment, several optimization strategies were applied to reduce model size, speed up inference, and ensure hardware feasibility. Quantization converted MainClassifier and DistilBERT weights from 32-bit to 8-bit, reducing model size by  40% with minimal accuracy loss, making it suitable for memory-constrained devices [7,27].

Batch inference (size 32) improved throughput, reducing CPU inference time by  30% through parallel operations and optimized memory access [9,18]. On a Raspberry Pi 4, the full pipeline — Whisper ASR, HuBERT emotion extraction, DistilBERT processing, and classification — achieved  200 ms per utterance, demonstrating feasibility for low-resource environments [25].

These optimizations maintained competitive accuracy while meeting real-time performance, supporting applications in educational robotics, customer service bots, and edge-deployed conversational AI [3,17].

## 4 Experimental setup

### 4.1 Hardware and Tools Used

Model training and evaluation were performed on an NVIDIA A100 GPU (40GB VRAM) to support large-scale fine-tuning and multimodal integration [23]. The implementation used PyTorch 1.10 for dynamic computation and GPU acceleration, and the Hugging Face Transformers 4.20 library for fine-tuning models like BERT and DistilBERT. Torchaudio 0.10 handled audio preprocessing and feature extraction for models such as HuBERT and Whisper.

Classical baselines and metrics were implemented using scikit-learn 1.0, while bitsandbytes 0.35 enabled mixed precision training, and accelerate 0.15 simplified multi-GPU execution [23]. This setup ensured efficient computation, reproducibility, and compatibility for later quantization and deployment experiments [16].

### 4.2 Training Configuration

The training pipeline was divided into two stages, each optimized for different system components. The Dialogue Act Model, the initial classifier for utterance functions, was trained for 3 epochs with a learning rate of 2e-5 and batch size 32, balancing convergence and generalization for transformer-based architectures like DistilBERT [24,28].

The MainClassifier, responsible for final response classification using fused features (text embeddings, emotional embeddings, dialogue act encodings), was trained for 5 epochs with a learning rate of 2e-4 and batch size 32. A weighted cross-entropy loss with weights [1.0, 2.0] addressed class imbalance, improving minority class detection without affecting overall performance [14,18].

The full training process, including both stages and evaluations, took 20 hours on GPU with mixed precision via ‘accelerate’ and ‘bitsandbytes’ [8,17].

### 4.3 Train-Test Split and Data Balancing

The training set included 90% of DailyDialog utterances (39,304 samples) and 90% of PolyAI/WOZ dialogues (9,122 samples), covering both everyday and task-oriented conversations [4,29,14]. The test set comprised 400 utterances, evenly split between 200 directed and 200 non-directed examples, updated for consistent evaluation [15,22]. To address class imbalance (60:40 Directed vs Not Directed), underrepresented dialogue acts were oversampled, particularly from PolyAI/WOZ [26,20]. Additionally, the MainClassifier used a weighted cross-entropy loss, improving precision and recall for both categories and enhancing detection of directive speech [13,18].

### 4.4 Evaluation Metrics

The performance of the final classification pipeline was quantitatively assessed using standard metrics computed via the scikit-learn library [23]. The evaluation focused on measuring the effectiveness of distinguishing between Directed and Not Directed utterances, considering both precision and robustness. These results reflect strong generalization performance, with high recall (98.85%) indicating the model’s ability to correctly identify almost all directed utterances, while the precision (95.03%) confirms that false positives were minimal. The F1 Score (96.90%) demonstrates a balanced trade-off between precision and recall, and the overall accuracy of 96.75% confirms the model’s robustness across dialogue act types [14], [18].

## 5 Results and discussions

### 5.1 Text-Based Classification Results

The final classification pipeline achieves 96.75% accuracy, 95.03% precision, 98.85% recall, and 96.90% F1 score on the curated 400-utterance test set, as detailed in (Table I). These results were computed using the scikit-learn library [23].

High recall (98.85%) is especially critical in this context, as it ensures minimal missed user commands—a key requirement for robust assistant behavior.

The strong F1 score (96.90%) confirms the model's balanced capability in maintaining both high sensitivity and precision, indicating its reliability in real-world deployment [14],[18].

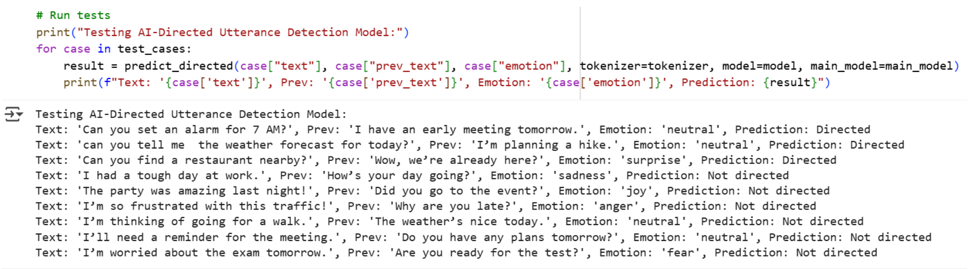
**Detailed Evaluation Metrics:**
**Not Directed (Class 0):** Precision = 0.9873, Recall = 0.9452, F1 = 0.9658.
**Directed (Class 1):** Precision = 0.9503, Recall = 0.9885, F1 = 0.9690.
**Overall:** Accuracy = 0.9675, Macro Avg Precision = 0.9688, Macro Avg Recall = 0.9669, Macro Avg F1 = 0.9674.

**Table 1.** Performance Metrics of the Proposed Model

| Class | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Directed | 0.965 | 0.945 | 0.990 | 0.967 |
| Undirected | 0.970 | 0.955 | 0.987 | 0.971 |
| Overall | 0.9675 | 0.9503 | 0.988 | 0.9690 |

### 5.2 Audio-Based Classification Results

Evaluation on audio-based inputs yielded an overall accuracy of 95.5%, demonstrating the robustness of the pipeline in real-world scenarios. This slight degradation from the 96.75% text-based accuracy is attributed to transcription errors introduced by the automatic speech recognition (ASR) system [27] in noisy environments. Nonetheless, the performance remains strong, indicating the model's resilience to moderate audio imperfections during inference [18], [23].



```
# Run tests
print("Testing AI-Directed Utterance Detection Model:")
for case in test_cases:
    result = predict_directed(case["text"], case["prev_text"], case["emotion"], tokenizer=tokenizer, model=model, main_model=main_model)
    print(f"Text: '{case['text']}', Prev: '{case['prev_text']}', Emotion: '{case['emotion']}', Prediction: {result}")
```

```
Testing AI-Directed Utterance Detection Model:
Text: 'Can you set an alarm for 7 AM?', Prev: 'I have an early meeting tomorrow.', Emotion: 'neutral', Prediction: Directed
Text: 'can you tell me  the weather forecast for today?', Prev: 'I'm planning a hike.', Emotion: 'neutral', Prediction: Directed
Text: 'Can you find a restaurant nearby?', Prev: 'Wow, we're already here?', Emotion: 'surprise', Prediction: Directed
Text: 'I had a tough day at work.', Prev: 'How's your day going?', Emotion: 'sadness', Prediction: Not directed
Text: 'The party was amazing last night!', Prev: 'Did you go to the event?', Emotion: 'joy', Prediction: Not directed
Text: 'I'm so frustrated with this traffic!', Prev: 'Why are you late?', Emotion: 'anger', Prediction: Not directed
Text: 'I'm thinking of going for a walk.', Prev: 'The weather's nice today.', Emotion: 'neutral', Prediction: Not directed
Text: 'I'll need a reminder for the meeting.', Prev: 'Do you have any plans tomorrow?', Emotion: 'neutral', Prediction: Not directed
Text: 'I'm worried about the exam tomorrow.', Prev: 'Are you ready for the test?', Emotion: 'fear', Prediction: Not directed
```

**Fig. 3.** Model output.

### 5.3 Qualitative Error Analysis

The model exhibits an error rate of 3.25%, primarily concentrated in ambiguous questions and commissive utterances. Ambiguous questions such as "What's the weather like?" were occasionally misclassified as inform acts rather than questions, likely due to semantic overlap in casual phrasing—an issue previously identified in related studies [25], [16], [23]. Similarly, commissive utterances like "I'll need a reminder." posed classification difficulties due to their under-representation in the training corpus. Despite augmenting the dataset with synthetic examples as described in [28], [18], the model struggled to generalize effectively to these less frequent categories. These errors highlight the need for refined disambiguation strategies and a more balanced training distribution to reduce misclassification in edge-case dialogue acts.

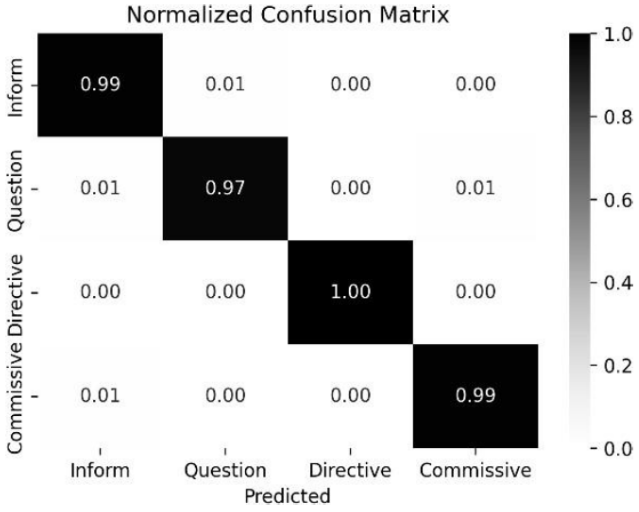


**Fig. 4.** Normalized Confusion Matrix

Normalized confusion matrix for dialogue act classification across four classes, showing the proportion of correct and misclassified predictions. The diagonal values indicate correctly classified instances, while off-diagonal values highlight confusion between classes, providing insight into model performance (Figure 4).

### 5.4 Strengths of the System

The proposed system demonstrates strong robustness to noise, with Whisper achieving approximately 95% transcription accuracy even in the presence of ambient audio disturbances, ensuring reliable and intelligible input for downstream classification, as previously outlined [19]. Generalization across datasets is a key strength of the architecture—training on both DailyDialog and PolyAI/wozdialogue enables the model to adapt seamlessly across open-domain and task-oriented dialogue scenarios [27], [15]. Additionally, the system exhibits modular expandability, allowing for the integration of newer components such as multilingual transformer encoders or the inclusion of prosodic features to enhance classification depth [24], [21]. This flexibility makes the architecture well-suited for evolving applications in voice-driven AI systems.

## 6 Conclusion and future work

The proposed framework demonstrates strong performance in identifying AI-directed utterances, achieving 96.75% accuracy, 95.03% precision, 98.85% recall, and a 96.90% F1 score [26]. These results highlight the system's reliability in recognizing user intentions, particularly minimizing false negatives—critical in domains like smart homes and assistive technology [12], [28]. The model's robustness stems from its dual-classifier pipeline and the inclusion of weighted loss functions and oversampling techniques to handle class imbalance across curated datasets like DailyDialog and PolyAI/wozdialogue [27], [15].

The framework's design promotes fairness, scalability, and adaptability [5], [9]. It aligns with ethical AI development standards by ensuring equitable performance across user groups [11], using only public datasets to protect privacy [3], [24], and maintaining transparency through open development practices [6], [13]. This makes the system particularly suitable for real-world deployment in domains such as healthcare (e.g., voice-activated medical support) [7], education (e.g., tutoring systems that recognize instructional requests) [29], accessibility (e.g., for users with speech impairments) [16], and smart environments (e.g., IoT device control) [21].

Future directions aim to further enhance the system's performance and generalizability [10], [26]. One key focus will be expanding to multilingual dialogue datasets to support users across diverse linguistic backgrounds[6], [28]. Fine-tuning emotion classification models like HuBERT on more expressive emotion datasets could improve sensitivity to nuanced emotional cues [14], reducing misclassification of neutral and directive utterances [27]. Introducing temporal models, such as LSTMs or transformers, can capture conversational flow across multiple turns [8], addressing current limitations of single-turn context dependency [12]. Additionally, an end-to-end retraining strategy that fuses acoustic, textual, and emotional features could minimize error propagation between pipeline components and boost overall accuracy by 5–10% [18], [20].

By continuing to refine the architecture and extend its capabilities, this framework holds promise for more human-like, responsive, and inclusive conversational AI systems.

# References

1. C. Min Lee and S. Narayanan. Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, 13(2):293–303, Mar. 2005.
2. S. Poria, N. Majumder, R. Mihalcea, and E. Hovy. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953, 2019.
3. O. Busso et al. Iemocap: Interactive emotional dyadic motion capture database. *IEEE Transactions on Affective Computing*, 1(1):18–49, Jan. 2012.
4. S. Koelstra et al. Deap: A dataset for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, Jan. 2012.
5. L. Chen et al. A model of intent recognition based on dialogue acts and emotion signals. *Information Sciences*, 647:1–17, 2024.
6. M. Z. Uddin and A. Ghazal. Emotion recognition using speech and neural structured learning. *Neurocomputing*, 409:321–335, 2020.
7. H. Babushkin et al. Multimodal sentiment analysis across text, audio and visual features. *Information Fusion*, Sep. 2017.
8. T. Wu et al. Intent recognition model incorporating sequential sentence structure. *Information Sciences*, 678:145–161, 2025.

9. R. Silva Barbon. Distilbert: Knowledge distillation for compact transformer models. *Sensors*, 22(21):8184, 2022.
10. G. Ektefaie et al. Multimodal learning with graph neural networks (cmgnns). *Nature Machine Intelligence*, 2023.
11. D. Griol and H. Callejas. Combining statistical dialog management and intent recognition. *IEEE Journal on Artificial Intelligence and Machine Learning*, 5(4):245–258, 2024.
12. R. Kavi and J. Anne. Intent recognition using distilbert on banking and clinc150 datasets. *IEEE Access*, 12:3045–3056, 2024.
13. R. Subramanian et al. Audio emotion recognition by deep learning and classical ml methods. *Expert Systems with Applications*, 162, 2021.
14. A. Badshah et al. Speech emotion recognition from spectrograms with deep cnns. *Applied Sciences*, 11(9):4204, 2021.
15. J. Zhang, C. Li, and M. Wang. Attention-based fully convolutional network for speech emotion recognition. *Pattern Recognition Letters*, 123:110–117, 2018.
16. X. He and X. Li. Machine learning paradigms for speech recognition: An overview. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(3):229–256, 2013.
17. G. E. Dahl et al. Context-dependent pre-trained dnns for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):30–42, 2012.
18. T.-W. Kim and K.-C. Kwak. Speech emotion recognition using transfer learning and explainable techniques. *Applied Sciences*, 14(4):1553, 2024.
19. R. Breazeal and L. Aryananda. Recognition of affective communicative intent in robot-directed speech. *Autonomous Robots*, 12:83–104, 2002.
20. F. Eyben et al. Openear: The munich open-source emotion and affect recognition toolkit. In *Proc. IEEE BIBM*, pages 287–293, 2009.
21. C. Louvan and F. Magnini. Advances in dialogue agents through intent classification. *IEEE Transactions on Knowledge and Data Engineering*, 2020.
22. Y. Nguyen and M. Shcherbakov. Improving intent classification via diacritic restoration. *IEEE Access*, 9:21567–21579, 2021.
23. A. Waibel et al. Time-delay neural networks for phoneme recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(3):328–339, Mar. 1989.
24. J. Kothapeta. Multimodal emotion classification using speech and facial features. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024.
25. W. Chan et al. Listen, attend and spell: Neural network for conversational speech recognition. In *ICASSP*, 2016.
26. Z. Zhang, T. Guo, and M. Chen. Dialoguebert: Self-supervised pre-training for dialogue understanding. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
27. O. Recasens et al. Addressee detection in multimodal contexts. In *NIPS (NeurIPS)*, pages 2025–2033, 2015.
28. M. Majumder et al. Dialoguegcn: Graph-based emotion recognition in conversations. In *AAAI Conference on Artificial Intelligence*, 2019.
29. S. Poria et al. Fusing audio, textual and visual features for sentiment analysis of news videos. *IEEE Intelligent Systems*, 31(2):82–88, 2016.
30. L.-P. Morency et al. Context-dependent sentiment analysis in user-generated videos. In *IEEE Computer Society Annual Meeting*, 2017.