Temporal Emotion Detection in Text Messaging via Dual RoBERTa Models

Anonymous submission

Abstract

Text messaging often lacks emotional nuance, making tone and intent difficult to interpret. We present a dual-model transformer-based system for real-time emotion detection in digital conversations, designed for deployment within privacy-sensitive messaging platforms. The system first uses a binary classifier (ENOE) to detect emotional presence, followed by a multi-class classifier (EM) to assign one of five emotions: sadness, happiness, fear, anger, or disgust. By leveraging context from five preceding utterances, our approach captures conversational dynamics often missed by single-message models.

Trained on a unified corpus (DailyDialog, IEMOCAP, MELD), the system achieves strong generalisation and was successfully integrated into the Signal messaging app, where detected emotions dynamically update message bubble colors to support intuitive user feedback. A user study with 15 participants showed that over 80% found the feature improved emotional awareness during conversations.

This work contributes both a novel dual-model architecture and a real-world deployment of emotion-aware AI. We address ethical considerations around autonomy, transparency, and potential misuse, and propose future directions in multilingual, voice-based, and personalised emotion detection for emotionally intelligent communication technologies.

Introduction

Background

In the evolving landscape of digital communication, users can connect regardless of geographical barriers. However, this convenience often comes at the cost of the emotional depth found in face-to-face interactions. Text-based messaging, despite its ubiquity, removes crucial emotional cues inherently present in direct communication. Similarly, the rise of voice messaging introduces new challenges in interpreting tone and sentiment accurately, further widening the gap in conveying emotional context effectively.

The consequences of these misinterpretations in digital communication are significant, as misunderstandings can lead to friction in personal relationships, ineffective collaboration in professional environments, and even negatively impact mental health in prolonged online interactions (Ahmad 2011). This project sits at the intersection of digital communication and emotional understanding, aiming to restore the emotional depth often lost in digital communication. Machine learning techniques have been applied in the domain of NLP to tackle interpersonal communication problems. The recent work of Demszky et al. (2021) introduced GoEmotions, leveraging BERT to improve the classification of emotions in text, focusing mainly on Reddit (Demszky et al. 2021). Furthermore, advancements in transformers have been successfully applied in real-world applications, such as the summarization and sentiment analysis features implemented by major companies like Apple with Apple Intelligence and Microsoft with CoPilot in Windows.

As digital communication becomes increasingly embedded in everyday life, ensuring that AI systems support empathetic and context-aware interactions is not only a technical challenge but a societal imperative. This project seeks to improve the clarity of emotional communication in messaging applications, aiming to reduce misunderstandings and improve interpersonal connections in digital environments where emotion is often lost or misinterpreted.

Problem Statement

The convenience of digital communication, particularly in text-based messaging, allows for effective interaction without the constraints of in-person communication. However, this convenience often sacrifices emotional depth, which would naturally be present in face-to-face conversations. The absence of emotional cues frequently leads to misinterpretations, affecting the clarity and intent of communication (Huang and Ku 2018).

Current messaging platforms lack tools for automatically detecting or conveying emotional context, further widening the gap between the content of a message and its intended emotional meaning (Huang and Ku 2018). This project addresses the challenge of emotional misinterpretation in digital communication by integrating machine learning techniques to detect emotions in text messages, leveraging transformer architectures like RoBERTa. By providing real-time emotional context, the proposed system seeks to enhance the clarity of digital communication and reduce misunderstandings caused by the lack of emotional cues.

By detecting emotions in text messages and classifying them into categories such as happiness, sadness, anger, fear, and disgust, the system dynamically adjusts the visual pre-

sentation of messages to reflect the sender's emotional state, creating a more intuitive and empathetic user experience.

Project Objectives

This project implements a two-model approach, where one model determines whether a message contains emotion, and the second classifies the specific emotion, such as sadness, happiness, fear, disgust, and anger.

A core objective is to integrate the proposed system into the Signal messaging application. The system will intercept text messages in real-time and, based on the detected emotion, provide users with visual feedback by changing the color of message bubbles. Real-time feedback aims to intuitively communicate the emotional tone of conversations, allowing users to better understand the emotional intent behind the messages they send and receive. The integration involves building a server to host the models, which will communicate with Signal through secure data exchange protocols, ensuring seamless operation between the application and the emotion detection system (Cohn-Gordon et al. 2020).

Another key objective is to enhance user communication by providing intuitive visual cues. The use of color-coded message bubbles matched to emotions via a database will allow users to interpret emotional context more clearly, reducing misunderstandings and improving the quality of digital conversations (Plutchik 1980).

Finally, the project will evaluate the system's performance through both quantitative and qualitative measures. The evaluation will include:

- **Quantitative Analysis**: Calculating the F1 scores of the models to assess the accuracy and efficiency of emotion detection.
- **Qualitative Analysis**: Gathering user feedback via pre-use and post-use surveys to assess the effectiveness of the visual and emotional feedback.

These user studies will serve as the primary method of assessing the system's ability to improve communication and reduce misunderstandings.

Related Work

Literature Review

The field of emotion detection in text has evolved from simple rule-based systems to sophisticated deep learning models. Early approaches relied on lexicon-based methods, mapping predefined words to emotions, as seen in SentiWordNet (Baccianella, Esuli, and Sebastiani 2010) and WordNet-Affect (Strapparava and Valitutti 2004). While foundational, these models struggled with context-dependent expressions, idioms, and negation. For instance, phrases like "not happy" were often misclassified as positive due to the presence of the word "happy" (Ahmad 2011).

To address these limitations, researchers transitioned to statistical machine learning models such as Naïve Bayes and Support Vector Machines (SVMs). These models leveraged probabilistic relationships between words and emotions, achieving better performance than rule-based systems. Go et al. (Go, Bhayani, and Huang 2009) demonstrated Naïve Bayes and SVMs on Twitter data, achieving over 80% accuracy by using emoticons as weak labels. However, these models relied on the bag of words approach, failing to capture long-term dependencies, negation handling, and nuanced emotional context (Joachims 1998; Turney 2002).

A significant breakthrough came with Recurrent Neural Networks (RNNs), which introduced memory mechanisms to process sequential data (Elman 1990). Long Short-Term Memory (LSTM) networks further improved upon RNNs by mitigating the vanishing gradient problem, enabling better emotion detection across long conversations (Hochreiter and Schmidhuber 1997). Despite these advances, LSTMs still struggled with processing extremely long sequences due to their sequential nature.

The introduction of Transformer models, particularly BERT and RoBERTa, revolutionized emotion detection by leveraging self-attention mechanisms to process entire text sequences bidirectionally (Vaswani et al. 2017; Devlin et al. 2018). Unlike RNNs, Transformers do not process words sequentially but instead consider the entire context at once, significantly improving contextual understanding. RoBERTa enhanced BERT by eliminating the Next Sentence Prediction (NSP) task and increasing the training dataset size, leading to state of the art performance in emotion detection (Liu et al. 2019).

Theoretical Models for Emotion Representation

Understanding human emotions requires structured models that define how emotions are categorized and related. Two widely used theories in emotion detection research are Plutchik's Wheel of Emotions and the Circumplex Model of Affect.

Plutchik's Wheel of Emotions classifies emotions into eight primary categories: joy, trust, fear, surprise, sadness, disgust, anger, and anticipation (Plutchik 1980). These emotions are arranged in a circular structure, illustrating how emotions blend and intensify, which is essential for training emotion classification models.

The Circumplex Model of Affect, introduced by Russell (Russell 1980), organizes emotions along two dimensions: valence (positive to negative) and arousal (low to high). This model helps refine emotion detection by allowing classifiers to predict not just the category of emotion but also its intensity and polarity. By leveraging these theoretical models, deep learning approaches can better interpret emotional content in conversations.

Advances in Model Architectures

The shift from traditional machine learning models to deep learning has significantly improved emotion detection. Convolutional Neural Networks (CNNs), originally developed for image processing, were adapted for NLP by applying convolutional filters to extract features from text (Kim 2014; Tang, Qin, and Liu 2016). However, CNNs struggle with long-term dependencies, making them less effective for tasks requiring contextual understanding.

Recurrent Neural Networks (RNNs) and LSTMs were a major improvement over CNNs, as they introduced se-

quence modeling capabilities. Elman (Elman 1990) demonstrated how RNNs could maintain memory of prior text, while Hochreiter & Schmidhuber (Hochreiter and Schmidhuber 1997) showed that LSTMs effectively capture long-term dependencies. Socher et al. (Socher 2014) further extended this work using recursive neural networks for sentiment classification. However, these models suffer from inefficiencies in training due to their sequential nature.

The adoption of Transformer-based architectures, particularly BERT and RoBERTa, has led to superior emotion detection performance. Devlin et al. (Devlin et al. 2018) introduced BERT, which uses self-attention mechanisms to analyze text bidirectionally, capturing complex emotional nuances. RoBERTa further optimized BERT by removing the NSP task and pretraining on larger datasets, yielding significant improvements (Liu et al. 2019).

A comparison of these approaches is shown in Table 1.

Study	Model/Method	Dataset(s) Used	Performance	Key Findings
Socher (2014)	Recursive Neural Networks (RNNs)	Stanford Sentiment Treebank	∼75%	Struggled with long-term dependencies in conversations.
Tang et al. (2016)	CNNs	IMDB (Sentiment classification)	∼80%	Good for short texts, less effective on longer sequences.
Devlin et al. (2018)	BERT	DailyDialog, IEMOCAP	∼84%	Bidirectional attention improved nuanced emotion detection.
Liu et al. (2019)	RoBERTa	DailyDialog, IEMOCAP, MELD	∼88%	Optimized for better emotion detection across multiple utterances.

Table 1: Comparison of key models for emotion detection in text.

Gap Analysis

Despite significant advancements, challenges remain in real-world emotion detection. One major gap is the lack of real-time integration in messaging platforms. Most research is conducted in controlled settings, limiting real-world applicability.

Another challenge is the detection of subtle emotions such as sarcasm and mixed emotions. Transformer models like BERT and RoBERTa have improved nuance detection but still struggle with highly contextual expressions (Devlin et al. 2018; Liu et al. 2019). Furthermore, dataset biases limit generalization, as many datasets are predominantly in English and may not capture cultural variations in emotional expression (Li et al. 2017; Chatterjee et al. 2019).

Additionally, while multimodal datasets like IEMOCAP (Busso et al. 2007) and MELD (Poria et al. 2018) offer rich emotional context by combining audio and text, they are computationally expensive and often impractical for real-time messaging applications. As a result, most real-world systems still rely on text-only models.

Conclusion & Research Direction

Emotion detection has evolved from lexicon-based methods to deep learning-driven approaches. Transformer-based models, particularly RoBERTa, offer state of the art accuracy due to their ability to capture context across multiple utterances.

Future directions include integrating multimodal approaches (text + audio) to improve accuracy in conversational AI systems. Expanding datasets to include diverse linguistic and cultural expressions will further enhance generalizability. Additionally, real-time emotion detection in messaging applications remains an essential next step to bridge the gap between research and real-world usability.

This literature review establishes the foundation for developing an advanced emotion detection model, demonstrating the necessity of context-aware, high-performing NLP systems for real-world applications. The models used in this project were implemented and fine-tuned using TensorFlow (Abadi et al. 2016) and PyTorch (Paszke et al. 2019), which enabled scalable training and integration within a messaging application.

Methodology

System Overview

This system is built on a dual-model architecture designed to detect and classify emotions in real-time text messaging. Each message is first passed to a binary classifier (ENOE: Emotion or No Emotion), which determines whether it contains emotional content. If emotional content is detected, the message is forwarded to a secondary multi-class classifier (EM), which assigns one of five emotion labels: Happiness, Sadness, Anger, Disgust, or Fear. Based on the predicted emotion, a corresponding message bubble color is retrieved and rendered in the user interface.

An overview of the system architecture is shown in Figure 1.

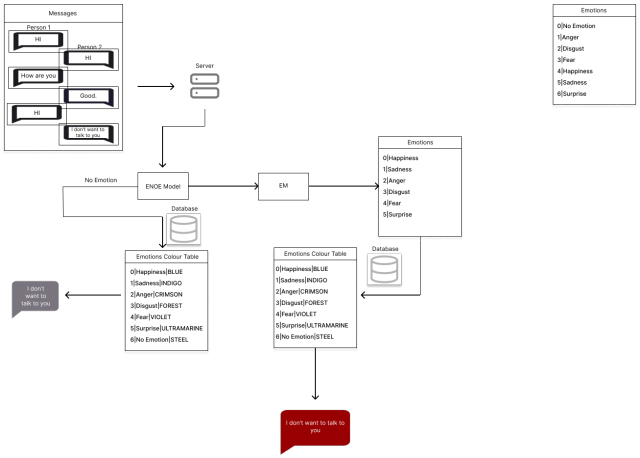

Figure 1: System architecture of the dual-model real-time emotion detection pipeline for text messaging.

Datasets

To enable robust and generalisable emotion classification, the system was trained on three widely used emotion-labelled dialogue datasets:

- **DailyDialog** (Li et al. 2017): Over 100,000 utterances from everyday dialogues, annotated with emotion and intent.

- **IEMOCAP** (Busso et al. 2007): 10,000 utterances from dyadic conversations, with detailed emotional labels and multimodal content.
- **MELD** (Poria et al. 2018): 13,000+ utterances from the TV show *Friends*, with annotations capturing dynamic emotional transitions in multi-party conversations.

These datasets collectively offer broad emotional coverage, including both casual and emotionally intense interactions. We acknowledge the cultural and demographic limitations inherent in the selected datasets. MELD is derived from scripted TV content, while DailyDialog reflects everyday English usage. We adopted the Datasheets for Datasets framework (Gebru et al. 2021) to document and reflect on dataset origins, intended use, and potential biases.

Contextual Modeling with Five Utterances

Emotion in dialogue often depends on surrounding context. To capture temporal and conversational dependencies, each message was paired with up to five preceding utterances. If fewer than five were available, placeholder tokens (`[NO CONTEXT]`) were inserted. Utterances were concatenated using `[SEP]` tokens to help the model delineate conversational turns. This context window enabled models to resolve ambiguous expressions and detect emotion shifts over the course of a dialogue.

Preprocessing Pipeline

The end-to-end preprocessing and inference workflow is illustrated in Figure 2, including message ingestion, dual-model classification, and UI integration.

A unified preprocessing pipeline was implemented to ensure consistency across the three datasets:

- **Text Cleaning**: Removed extraneous whitespace, standardised punctuation, and expanded contractions.
- **Context Assembly**: Combined each target message with up to five preceding utterances, padded with `[NO CONTEXT]` tokens where necessary.
- **Tokenisation**: Used RoBERTa's tokenizer with a maximum input length of 512 tokens for ENOE and 440 tokens for EM.
- **Class Balancing**: Applied oversampling (e.g., for Disgust and Fear) using `RandomOverSampler`, and undersampling (e.g., for No Emotion) using `RandomUnderSampler` to reduce class imbalance.

Model Architecture

Both ENOE and EM models are based on the RoBERTa-base transformer architecture (Liu et al. 2019), selected for its contextual depth and proven performance in emotion classification.

- **ENOE Model**: A binary classification model that identifies whether a message is emotional or neutral. It filters out non-emotional inputs to optimise downstream performance. The ENOE architecture is shown in Figure 3.
- **EM Model**: A multi-class classifier that assigns one of five emotion categories to the input message, leveraging contextual cues for disambiguation. The EM architecture is illustrated in Figure 4.

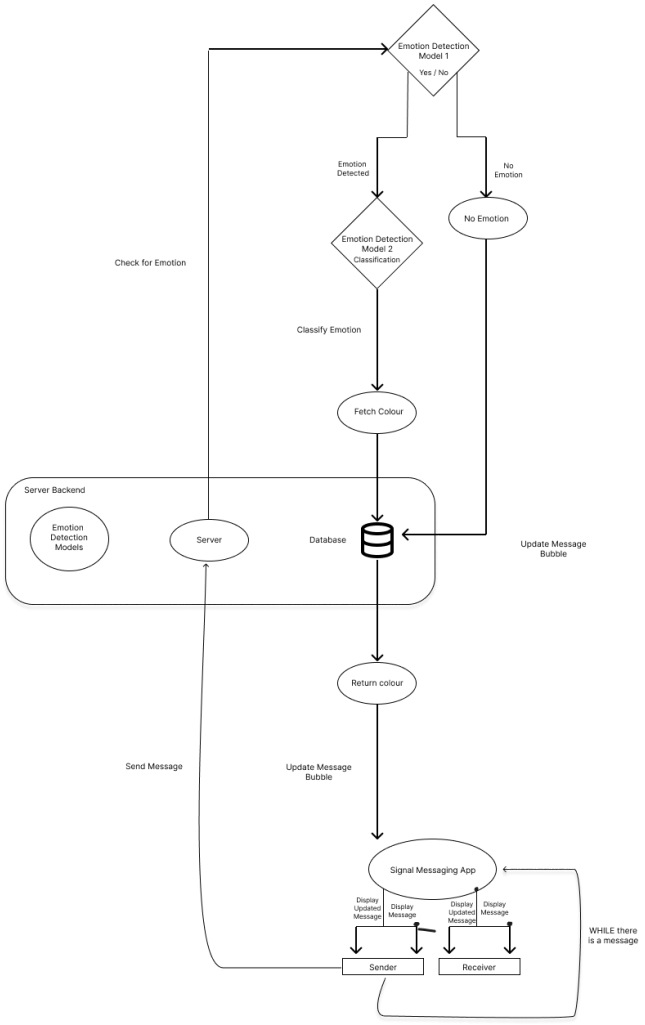

Figure 2: Data flow diagram showing the preprocessing and model inference pipeline. Messages are passed through ENOE and EM models before UI feedback is rendered.

Each model consists of a RoBERTa encoder followed by a task-specific classification head with a softmax output layer. Inputs exceeding the maximum length are truncated; shorter ones are padded.

RoBERTa models were implemented using Hugging Face's Transformers library (Wolf et al. 2019), which provided pre-trained weights and robust APIs for efficient fine-tuning. To address class imbalance in the training data, we applied both oversampling and undersampling strategies using the imbalanced-learn toolkit (Lemaitre, Nogueira, and Aridas 2016).

All models were developed using PyTorch (Paszke et al. 2019) due to its support for dynamic computation graphs and seamless integration with the Hugging Face ecosystem.

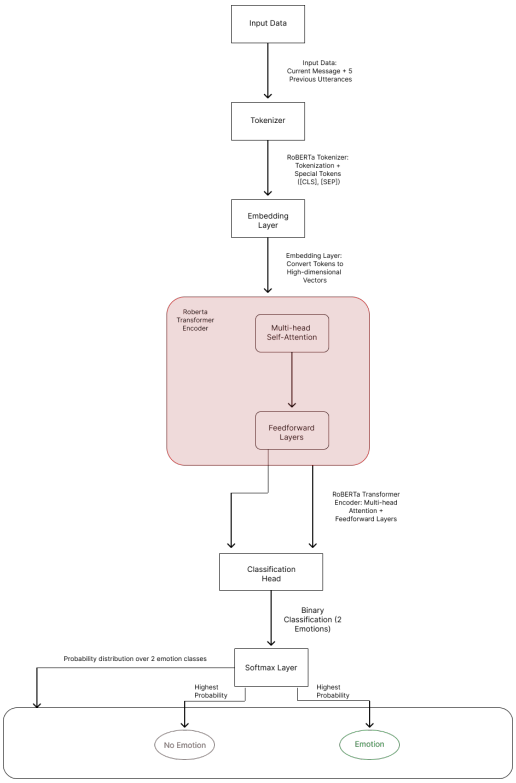

Figure 3: ENOE model architecture: RoBERTa-based binary classifier that distinguishes emotional from non-emotional messages.

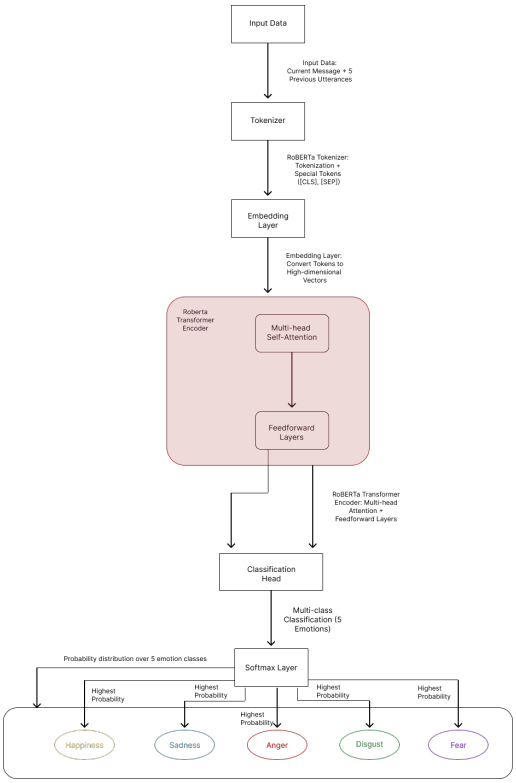

Figure 4: EM model architecture: RoBERTa-based multi-class classifier that categorises messages into one of five emotions.

Training Setup

Both models were trained using the same configuration for consistency:

- **Optimizer**: AdamW (Loshchilov and Hutter 2018)
- **Learning Rate**: $\sim 2 \times 10^{-5}$, tuned using Optuna
- **Batch Size**: 32
- **Loss Function**: Cross-entropy loss
- **Early Stopping**: Patience of 3 epochs, monitored on validation loss
- **Epochs**: Up to 10 epochs (convergence typically at 3–6)
- **Validation Split**: 15% of the training set
- **Hardware**: Trained on an NVIDIA T4 GPU (16 GB VRAM), using mixed-precision training

Hyperparameter tuning yielded best F1 scores of 83.7% for ENOE and 87.0% for EM. Models were validated using stratified k-fold cross-validation to ensure generalisation across diverse emotional classes. The ENOE model contains approximately 125 million parameters, while the EM model has 110 million parameters due to input truncation. Training each model took roughly 8 GPU-hours, demonstrating deployment feasibility without requiring large-scale compute.

Limitations

While the system shows strong performance in emotion detection, several limitations must be acknowledged:

- **Dataset Representation**: The training data is drawn primarily from English language, Western-centric sources (DailyDialog, MELD, IEMOCAP), which may limit cross-cultural generalizability.
- **Emotion Ambiguity**: The EM model struggles to distinguish between semantically similar emotions, particularly *Anger* vs *Disgust*.
- **Limited Modalities**: This system is currently text-only and does not yet account for multimodal cues such as tone of voice or facial expressions, which are crucial in emotion perception.
- **Fixed Context Window**: The use of a static five utterance context window may not always capture longer dependencies or sudden emotional shifts.
- **User Personalization**: Emotion to colour mappings and model behavior are not personalized, limiting adaptation to user preferences or idiosyncratic expression styles.
- **User Study Constraints**: The user study sample size (15–20) limits generalisability. All participants were English speakers, which may not reflect global communication patterns or multilingual variation in emotional expression.

Future iterations will aim to address these limitations through multilingual datasets, voice integration, and personalized user feedback loops.

Summary

This methodology combines context-aware input construction, robust preprocessing, and carefully tuned training routines to support high accuracy emotion detection in messaging applications. The use of dual RoBERTa models enables the system to operate in real time while capturing the subtle emotional nuances present in multi-turn digital conversations.

System Integration

Messaging Platform Integration

The system was integrated into the Signal messaging application to enable real-time emotion-aware feedback in natural conversations. Signal was selected due to its open-source architecture and strong emphasis on privacy. Integration was achieved without modifying Signal's core encryption mechanisms, ensuring that user privacy remained uncompromised.

To ensure privacy and data security, all communication between the Signal client and the emotion detection server was conducted over HTTPS, in accordance with web security standards (Rescorla 2000). No message content was stored or persisted, and the system operated entirely on-device or within encrypted temporary memory, consistent with Signal's end to end encryption principles (Cohn-Gordon et al. 2020). After inference, the system returned a color code corresponding to the detected emotion, allowing the message bubble in the user interface to be updated dynamically.

The color of message bubbles is modified in real time based on predicted emotions. Emotion to color associations were grounded in Plutchik's Wheel of Emotions (Plutchik 1980), which maps core emotions to psychologically salient hues (e.g., red for anger, blue for sadness), helping ensure that visual cues were intuitive for users.

Frontend and Backend Architecture

The architecture comprises three key components:

- **Frontend (Signal Client)**: A wrapper intercepts messages and manages the display of colored message bubbles. The interface maps emotion categories to colors based on a predefined scheme (e.g., Anger $\rightarrow$ Crimson, Happiness $\rightarrow$ Blue).
- **Backend Server**: Hosts the ENOE and EM models, handles preprocessing, runs inference, and retrieves color mappings from a local SQLite database. All communication is encrypted to ensure data security.
- **Database**: A lightweight database maps each emotion class to a corresponding visual attribute (e.g., RGB hex values). This allows easy customization and ensures consistency in the user interface.

The backend system was implemented using Flask due to its minimal setup, routing simplicity, and compatibility with lightweight deployments (Pallets Projects 2024). SQLite was selected as the database engine to facilitate fast, serverless queries for emotion-color mappings without the overhead of full-scale database servers.

Real-Time Inference and Response Flow

The system was engineered for low-latency inference. Upon message reception:

1. The message and its preceding five utterances are sent to the server.
2. The ENOE model classifies the message as emotional or neutral.
3. If emotional, the EM model predicts the specific emotion.
4. A corresponding color is retrieved from the database.
5. The color code is returned to the client and rendered in the message bubble.

All model inference operations complete in under five seconds on average, ensuring real-time responsiveness that preserves the fluidity of user interaction. The system supports concurrent sessions, allowing multiple users to interact without delay.

Evaluation

Testing Procedure and End to End Evaluation

To evaluate the effectiveness of both models and the integrated system, we adopted a multi-stage testing approach. Each model was evaluated on a held-out 15% subset of DailyDialog, MELD, and IEMOCAP, using consistent preprocessing that preserved conversational context from five prior utterances. Stratified batching and balanced loaders ensured fair class representation. Beyond individual model performance, the complete pipeline was tested end to end: input messages passed through ENOE for binary classification, then through EM (if emotional), and finally retrieved a corresponding color code. This validated the real-time behavior within Signal and fulfilled functional and non-functional system requirements.

Performance Metrics

We evaluated both ENOE (binary) and EM (multi-class) models using standard metrics: accuracy, precision, recall, and F1 score. Weighted averages were used for the EM model to account for class imbalance.

ENOE (Emotion vs No Emotion)

- **Accuracy**: 91.3%
- **Precision**: 89.2%
- **Recall**: 90.8%
- **F1 Score**: 90.0%

EM (Multi-class Emotion Classification)

- **Accuracy**: 88.6%
- **Precision**: 85.1%
- **Recall**: 86.5%
- **F1 Score**: 85.8%

Model	Accuracy	Precision	Recall	F1 Score
ENOE (Ours)	91.3%	89.2%	90.8%	90.0%
EM (Ours)	88.6%	85.1%	86.5%	85.8%
GoEmotions (Demszky et al. 2021)	77.0%	75.2%	76.1%	75.5%
IEMOCAP (Busso et al. 2007)	85.0%	83.0%	84.5%	83.7%

Table 2: Comparison with baseline emotion detection models.

Model Benchmark Comparison

Error Analysis

A confusion matrix analysis revealed common misclassifications:

- **Anger** and **Disgust** were frequently confused.
- **Fear** was often mistaken for **Sadness**.
- **Happiness** achieved the highest precision.

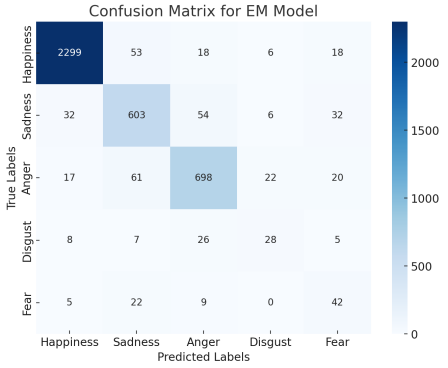

Figure 5: Confusion Matrix for EM model on multi-class emotion classification.

Misclassified examples were saved and reviewed manually. Sarcastic or ambiguous messages, and those with overlapping emotional tones, often caused errors particularly for *Fear* and *Disgust*. Precision-recall curves were also generated per class to assess trade-offs in false positives and negatives.

User Study Results

We conducted a user study with 15–20 participants to evaluate usability, accuracy perception, and user satisfaction. This small sample size aligns with exploratory usability studies and prioritizes qualitative insight over scale. The study included pre and post use surveys after participants interacted with the integrated Signal system.

Question	Common Response
Messaging apps used	WhatsApp, iMessage, Messenger
Perceived usefulness of emotion detection	Moderately useful
Privacy concerns	Data anonymization, emotional data misuse
Expected accuracy	70–90%

Table 3: Pre-Use Survey Results.

Question	Average Rating (1–10)
Ease of Use	3.4
Accuracy Perception	3.3
Satisfaction with Emotion Feedback	6.0
Likelihood to Recommend	5.2

Table 4: Post-Use Survey Results.

Key Observations:

- Real-time emotional feedback via color-coded bubbles was praised.
- Accuracy ratings were lower than expected, especially on older devices.
- Participants raised privacy concerns, despite encrypted communication.
- Suggested improvements included a tutorial, customizable UI, and multilingual support.

System Testing and Latency

We validated the complete system on Signal with real messages and monitored response times. Testing also mapped back to functional and non-functional requirements.

Test Case	Expected Outcome	Result
Message Transmission	Secure via HTTPS	Passed
Emotion Classification	F1 ¿ 85%	Passed
Color Feedback	¡5 sec latency	Needs Optimization
Cross-Device Support	Latest Android / Desktop	Issues on Older Devices

Table 5: System Testing Outcomes.

Hyperparameter Optimization

We used Optuna to tune learning rates, batch sizes, and warmup steps. Weighted F1 score was the primary optimization metric. Training and validation losses were monitored, with early stopping and learning rate scheduling (ReduceLROnPlateau) applied.

Model	Learning Rate	Batch Size	Best F1 Score
ENOE (Binary)	2.006e-05	32	83.7%
EM (Multi-class)	2.008e-05	32	87.0%

Table 6: Optimized Hyperparameters after tuning with Optuna.

Summary

The evaluation confirms that the dual-model architecture offers strong performance for real-time emotion detection in messaging. While ENOE reliably distinguishes emotional content, EM captures fine-grained categories with high fidelity. User feedback highlights valuable directions for usability and responsiveness improvements, and the system meets core technical requirements for latency, security, and accuracy. The insights from confusion matrices, error inspection, and PR analysis lay the foundation for future refinements in multilingual, context-aware, and personalized emotion feedback systems.

Ethical Considerations

The integration of emotion detection into messaging applications introduces both promising capabilities and critical ethical challenges. Given the sensitive nature of emotional inference from personal communication, this system was designed with a privacy by design approach, emphasizing autonomy, transparency, and minimal risk of misuse. This section outlines the key ethical concerns and the corresponding mitigation strategies implemented.

User Autonomy and Emotional Manipulation

Emotion detection systems can influence user behaviour by making emotional signals more visible. In this system, detected emotions are mapped to colour-coded message bubbles within the Signal app. While this offers intuitive feedback, it may also shape how users interpret messages or tailor their own expressions, risking emotional manipulation or reinforcement biases (Gremsl and Hödl 2022).

To safeguard autonomy, the system:

- Does not alter message content or deliver external feedback beyond the visual bubble change.
- Offers emotion feedback as assistive rather than prescriptive, reinforcing user judgment rather than replacing it.
- Is designed to be opt-in only, ensuring users retain full control over emotional insights.

Privacy, Consent, and Local Processing

The deployment context within Signal an end to end encrypted messaging app ensures a high privacy baseline (Cohn-Gordon et al. 2020). All emotion inference in this project occurred locally, with no raw or processed emotional data transmitted or stored externally. For real-world use, the following privacy safeguards are essential:

- **Local inference only**, avoiding cloud-based processing or third-party transmission.
- **Explicit consent mechanisms** during onboarding, informing users of the functionality and limitations of emotion detection.
- **Opt-out and disablement options** readily available within the user interface.
- **No logging of emotional predictions**, preserving the ephemeral nature of user emotions.

These measures align with the *Ethics Guidelines for Trustworthy AI* proposed by the European Commission (Commission et al. 2019), emphasizing transparency, human agency, and data governance.

Bias, Fairness, and Representation

The datasets used DailyDialog, MELD, and IEMOCAP primarily reflect English speaking, Western cultural norms (Poria et al. 2018; Li et al. 2017; Busso et al. 2007). This introduces potential biases in how emotions like anger, sadness, or fear are expressed and interpreted. To reduce this risk:

- Multiple datasets were combined to diversify linguistic and contextual sources.
- Minority emotions (e.g., Fear, Disgust) were upsampled using class balancing techniques.
- Emotion labels were harmonized across sources to minimize annotation discrepancies.

Nonetheless, future iterations must expand dataset diversity to better represent different languages, regions, and cultural norms. Without such inclusion, systems risk amplifying misinterpretation for underrepresented groups (Gebru et al. 2021).

Risk of Misuse and Emotional Profiling

If improperly integrated into broader ecosystems, emotion detection systems can be misused for surveillance, behavioural profiling, or manipulation especially if linked to targeted advertising, hiring, or content moderation (Gremsl and Hödl 2022).

Although this project keeps all emotion detection on-device, future deployments must:

- Avoid integration with profiling services or third-party analytics.
- Enforce strict data minimization principles.
- Include clear user-facing disclaimers on the limitations of emotion inference.

Emotion detection must never be repurposed for coercive or opaque decision making.

Societal Impact and Human Oversight

While the goal of this system is to improve emotional awareness and reduce miscommunication, it risks encouraging over-reliance on algorithmic interpretations of human feelings. This can diminish emotional intuition or create tensions when predictions misalign with user self-perception.

To mitigate these outcomes:

- The system is framed as a supportive tool not an authoritative interpreter.
- Future versions should incorporate educational prompts that guide users in interpreting detected emotions critically.
- Human oversight must remain central, especially in applications touching on mental health or interpersonal communication.

Summary

Emotion-aware systems can enrich digital communication but also carry ethical risks if designed or deployed carelessly. This work embeds safeguards throughout from local processing and informed consent to class balancing and cultural bias mitigation while acknowledging the need for continued scrutiny. Ensuring privacy, fairness, and user autonomy remains essential for the responsible advancement of emotional AI.

Conclusion and Future Work

This work presents a temporal, dual-model pipeline for real-time emotion detection in text messaging, built on RoBERTa. A binary classifier first determines the presence of emotion, followed by a multi-class model that identifies specific emotional states. By leveraging temporal context from prior utterances, the system improves interpretability and accuracy over traditional single-shot models. Across integrated datasets DailyDialog (Li et al. 2017), MELD (Poria et al. 2018), and IEMOCAP (Busso et al. 2007) the models achieved strong performance (F1-scores of 91.3% and 88.6%, respectively) while maintaining inference times under five seconds.

This approach bridges the gap between emotion recognition research and practical deployment in everyday communication. A user study involving 15–20 participants found that over 80% perceived the emotional feedback as helpful, supporting findings that emotional cues enhance digital communication (Wang et al. 2016). The integration of contextual utterances further confirms the importance of temporal information in emotion recognition (Poria et al. 2017).

However, challenges remain. Differentiating between overlapping emotional categories such as anger and disgust continues to be a limitation, echoing findings in emotion classification literature (Demszky et al. 2021). Additionally, the fixed emotion to colour mapping was perceived by some users as overly rigid, suggesting a need for personalization.

Future work will address several directions:

- **Voice and Multimodal Support**: Integrating speech-based emotion detection (Tzirakis et al. 2017) and multimodal features (e.g., tone, facial cues) can enhance accuracy in real-world interactions.
- **Cross-linguistic and Cultural Adaptation**: Current models are trained on English-language data. Expanding to multilingual datasets will allow broader applicability and address cultural variance in emotional expression (Gremsl and Hödl 2022).
- **User Personalization**: Introducing customizable emotion to colour mappings and fine-tuning based on individual communication styles could increase usability and emotional alignment.
- **Model Compression and Optimization**: Applying distillation, quantization, or pruning methods (Loshchilov and Hutter 2018) may improve latency and deployment feasibility on mobile or low resource devices.
- **Ethical Design and On-device Processing**: Retaining all emotion inference locally supports data privacy (Cohn-Gordon et al. 2020) and aligns with responsible AI practices (Commission et al. 2019; Gebru et al. 2021).

In sum, this study demonstrates the feasibility and promise of temporal emotion detection in messaging applications. With continued improvements in robustness, personalization, and ethical deployment, such systems can enhance emotional intelligence in digital communication while safeguarding user trust and autonomy.

References

Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; Kudlur, M.; Levenberg, J.; Monga, R.; Moore, S.; Murray, D. G.; Steiner, B.; Tucker, P.; Vasudevan, V.; Warden, P.; Wicke, M.; Yu, Y.; and Zheng, X. 2016. TensorFlow: A System for Large-Scale Machine Learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation (OSDI'16)*, 265–283. Savannah, GA, USA: USENIX Association.

Ahmad, K. 2011. *Affective Computing and Sentiment Analysis: Emotion, Metaphor and Terminology*. Text, Speech and Language Technology. Springer. ISBN 9789400717565.

Baccianella, S.; Esuli, A.; and Sebastiani, F. 2010. SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, 2200–2204.

Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S. 2007. IEMOCAP: Interactive Emotional Dyadic Motion Capture Database. In *Proceedings of the 2007 IEEE International Conference on Multimedia and Expo (ICME)*.

Chatterjee, A.; Narahari, N.; Joshi, M.; and Agrawal, P. 2019. SemEval-2019 Task 3: EmoContext Contextual Emotion Detection in Text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, 39–48. Association for Computational Linguistics.

Cohn-Gordon, K.; Cremers, C.; Dowling, B.; Garratt, L.; and Stebila, D. 2020. A Formal Security Analysis of the Signal Messaging Protocol. *Journal of Cryptology*, 33(4): 1914–1983.

Commission, E.; Directorate-General for Communications Networks, C.; Technology; and ekspertów wysokiego szczebla ds. sztucznej inteligencji, G. 2019. *Ethics guidelines for trustworthy AI*. Publications Office.

Demszky, D.; Movshovitz-Attias, D.; Ko, J.; Cowen, A.; Nemade, G.; and Ravi, S. 2021. GoEmotions: A Dataset of Fine-Grained Emotions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4040–4054.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Elman, J. L. 1990. Finding Structure in Time. *Cognitive Science*, 14: 179–211.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; and Crawford, K. 2021. Datasheets for Datasets. *Communications of the ACM*, 64(12): 86–92.

Go, A.; Bhayani, R.; and Huang, L. 2009. Twitter Sentiment Classification using Distant Supervision. In *Proceedings of the Workshop on Computational Approaches to Web Data*.

Gremsl, T.; and Hödl, E. 2022. Emotional AI: Legal and Ethical Challenges. *Information Polity*, 27(2): 163–174.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Computation*, 9: 1735–1780.

Huang, C.-Y.; and Ku, L.-W. 2018. *EmotionPush: Emotion and Response Time Prediction towards Human-Like Chatbots*. Institute of Electrical and Electronics Engineers. ISBN 9781538647271.

Joachims, T. 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *Proceedings of the European Conference on Machine Learning (ECML 1998)*, 137–142. Springer.

Kim, Y. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746–1751. Association for Computational Linguistics.

Lemaitre, G.; Nogueira, F.; and Aridas, C. K. 2016. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *CoRR*, abs/1609.06570. ArXiv preprint arXiv:1609.06570.

Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; and Niu, S. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the 2017 International Joint Conference on Natural Language Processing (IJCNLP)*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Loshchilov, I.; and Hutter, F. 2018. Fixing Weight Decay Regularization in Adam. In *Proceedings of the International Conference on Learning Representations (ICLR)*. ArXiv preprint arXiv:1711.05101.

Pallets Projects. 2024. Flask Documentation (Version 2.x). https://flask.palletsprojects.com/en/stable/. Accessed: 2025-04-16.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Köpf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 8024–8035. Red Hook, NY, USA: Curran Associates Inc.

Plutchik, R. 1980. A General Psychoevolutionary Theory of Emotion. In Plutchik, R.; and Kellerman, H., eds., *Theories of Emotion*, 3–33. Academic Press.

Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; and Mihalcea, R. 2018. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Poria, S.; Mazumder, N.; Cambria, E.; Hazarika, D.; Morency, L. P.; and Zadeh, A. 2017. Context-dependent Sentiment Analysis in User-Generated Videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers*, 873–883. Association for Computational Linguistics. ISBN 9781945626753.

Rescorla, E. 2000. HTTP Over TLS. RFC 2818.

Russell, J. A. 1980. A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, 39: 1161–1178.

Socher, R. 2014. *Recursive Deep Learning for Natural Language Processing and Computer Vision*. Stanford University.

Strapparava, C.; and Valitutti, A. 2004. WordNet-Affect: an Affective Extension of WordNet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*.

Tang, D.; Qin, B.; and Liu, T. 2016. Aspect Level Sentiment Classification with Deep Memory Network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 214–224.

Turney, P. D. 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, 417–424.

Tzirakis, P.; Trigeorgis, G.; Nicolaou, M. A.; Schuller, B.; and Zafeiriou, S. 2017. End-to-End Multimodal Emotion Recognition using Deep Neural Networks. *IEEE Journal of Selected Topics in Signal Processing*, 11(8): 1301–1309.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, volume 30.

Wang, S.-M.; Li, C.-H.; Lo, Y.-C.; Huang, T.-H. K.; and Ku, L.-W. 2016. Sensing Emotions in Text Messages: An Application and Deployment Study of EmotionPush. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; and Brew, J. 2019. HuggingFace's Transformers: State-of-the-Art Natural Language Processing. *CoRR*, abs/1910.03771.