Complex Emotion Detection

in Customer Support Messages

Advanced ML solution for multi-label emotion classification and intensity prediction in customer service communications

Anger

Frustration

Anxiety

Confusion

Disappointment

Satisfaction

Motivating Use Case

Why It Matters

Customer emotions strongly affect business performance:

- Likelihood of churn or order cancellation
- Negative reviews and public ratings
- Support workload and service efficiency
- Customer loyalty and future purchases

Key insight: Small and medium businesses especially lack resources for manual analysis

Why It's Challenging

Complex Emotions
Messages contain subtle, mixed emotions, not just positive/negative

No Public Dataset
Lacks publicly available datasets for complex emotion detection

Limited Tools
Existing tools are limited to Positive/Neutral/Negative only

Varied Message Styles
Slang, abbreviations, emojis, spelling errors, code-switching

How It's Solved Today

- Manual reading and subjective interpretation
- Basic sentiment analysis tools (limited detection)
- No widely adopted commercial solutions for multi-label emotion detection

Project Task Description

Formal Problem Statement

Develop a model that identifies multiple co-occurring emotions in customer support messages.

Input

A single customer message
(free-form text in English or Hebrew)

“I placed my order a week ago and the tracking hasn’t updated since Monday. I’m honestly starting to get worried something went wrong. I’ve messaged twice already and still no reply — this is really disappointing. I don’t want to cancel, but I’m getting pretty frustrated. Can someone please explain what’s happening?”

This message Contains:

- Multiple emotional layers
- Conflicting sentiments
- Escalation (worry → disappointment → frustration)
- Customer ambiguity (doesn’t want to cancel but is close)
- Multi-sentence mixed tone
- References to context (tracking, lack of response)

This demonstrates why simple sentiment analysis (“negative”) is not enough.

Output

Structured emotion vector with:

- Binary indicators for emotion presence
- Continuous intensity score (0-1)

```
{ "anger": 0.30,
  "frustration": 0.88,
  "anxiety": 0.76,
  "confusion": 0.42,
  "disappointment": 0.83,
  "satisfaction": 0.05 }
```

Project Novelty

- Multi-label complex emotion detection
- Emotion intensity prediction
- LLM-driven synthetic data generation

Models and Methods

Overall Pipeline

1. Synthetic Data
LLM-generated messages with controlled attributes

2. Preparation
Text normalization, train/validation/test split

3. Training
Multi-label classification & regression

4. Evaluation
Metrics + comparison with LLM baselines

5. Refinement
Analyze errors, regenerate targeted samples

Model Types & Techniques

Transformer-based Encoders:
BERT, RoBERTa, DistilBERT for multi-label classification

Joint Multi-task Learning:
Shared encoder with two output heads:

- - Multi-label classification head
- - Regression head for emotion intensity

Few-shot / Zero-shot Baselines:
GPT-4.1, GPT-5 for comparison

Fine-Tuning Strategy

Multi-label Sigmoid Output Layer (not softmax)

Weighted loss functions for unbalanced emotions

Hyperparameter tuning: learning rate, batch size, max sequence length

Synthetic data augmentation loops with style/domain variations

Cross-domain fine-tuning: train on multiple service domains

Implementation Approach
Joint model architecture with shared encoder and multi-modal outputs

Multi-label Classification

Intensity Regression

Data Specification and Generation

Data Requirements

- Diverse customer-support messages
- Multiple co-occurring emotions
- Emotion intensity scores (0-1)
- Multiple service domains
- Balanced emotion representation

No manual labeling required - labels produced during synthetic generation

Dataset Overview

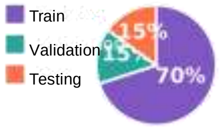
- 4,000-8,000 synthetic messages
- Structured JSON format
- "text" - customer message
- Emotion labels (anger, frustration, anxiety, confusion, disappointment, satisfaction)
- Binary presence + intensity score for each emotion

Example: "frustration": 0.85, "anxiety": 0.7

Generation Strategy

- Attribute-Driven Prompting
- Generation Diversity Controls
- Consistency Validation

Data Splits



Example Dataset Entry:

```
{ "text": "I've been waiting for my package all week. This is so frustrating and I'm starting to get worried.",
"anger": 0.1, "frustration": 0.85, "anxiety": 0.7, "confusion": 0.2, "disappointment": 0.6, "satisfaction": 0.0 }
```

Metrics and KPIs

The project evaluates two core prediction tasks: multi-label emotion classification and emotion intensity regression.

Multi-Label Classification

Micro F1-score
Best for unbalanced emotion distributions

Macro F1-score
Measures performance per emotion equally

Precision & Recall per emotion
Identifies which emotions the model confuses

Emotion Intensity Regression

MAE (Mean Absolute Error)
Average absolute difference in intensities

RMSE (Root Mean Square Error)
Penalizes larger errors more strongly

Pearson Correlation
Alignment between predicted and true patterns

Ground Truth Protocol

- Train/validation/test split with synthetic ground truth
- Manual sanity-checking for a small subset
- Compare models vs. zero/few-shot LLM baselines

End-to-End Quality Measures

Emotion Detection Accuracy @ 0.5
Threshold for binary classification

Consistency Score
Dominant emotion match with text meaning

Cross-Domain Generalization
Performance across domains (e-commerce, delivery, etc.)

Evaluation combines classification and regression metrics

Comprehensive framework for model assessment