CLINICAL EMOTISUPPORT

Lior Zvieli

Yuval Reznik

Yoni Libman

UNDERSTANDING EMOTIONS

Telemedicine platforms face a high volume of patient messages, making it difficult to manually prioritize cases based on emotional distress or urgency.

OUR SOLUTION

Clinical EmotiSupport

An NLP based model which classifies the 6 major emotions most important in telemedicine requests.

This allows the responders of the requests to respond accordingly and effectively.

Target users

For healthcare: Doctors, nurses, medical secretaries...

For Administrative (for example: app support): medical secretaries, IT support team...

The 6 Emotions

Anger, Anxiety, Confusion, Disappointment, Frustration, Satisfaction

THE MODEL

INPUT

The telemedicine request of a patient.

(currently only in English)

OUTPUT

A Json object containing a multi-label classification of which of the 6 emotions occur in the input, each are represented by 1 or 0.

SUCCESS METRICS

F1 score, Precision, Recall

DATASET

Data:

- 2,000 synthetic, controlled examples
- Each example is a prompt which represents a patient request, attached with an array which shows which of the 6 emotions appear in the prompt.

Examples:

- {"text": "The appointment was supposed to be last week, but it keeps getting pushed back.", "domain": "administrative", "language": "en", "emotions": {"anxiety": 0, "confusion": 0, "frustration": 0, "anger": 1, "disappointment": 0, "satisfaction": 0}, "id": 1831}
- {"text": "Following up: We're waiting for the test results for Mom, and it's been over a week now. This is getting frustrating, and I'm not sure what to do.", "domain": "clinical", "language": "en", "emotions": {"anxiety": 0, "confusion": 1, "frustration": 1, "anger": 0, "disappointment": 0, "satisfaction": 0}, "id": 62}

DATA GENERATION PIPELINE

Attributed Prompting

Generation Engine: Local Ollama

Model: Deepseek r1-8b

Parameters that go into the data creation prompt:

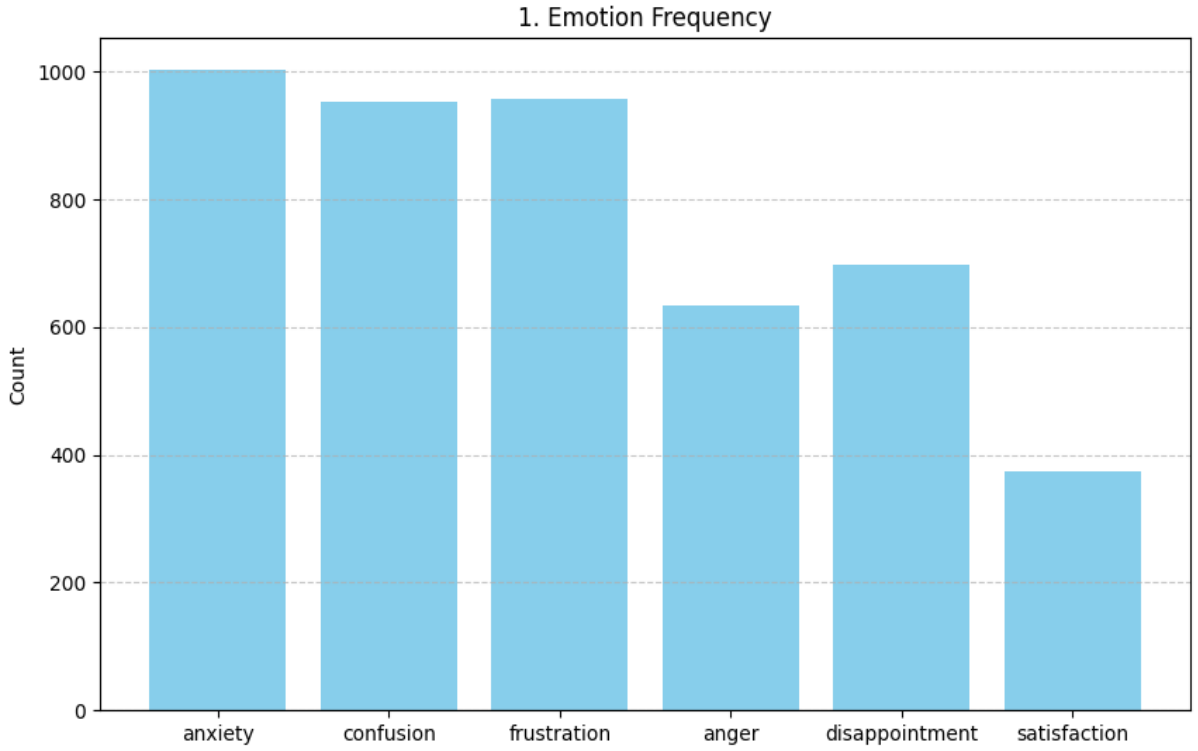
- Domain random selection (Clinical / Administrative)
- Style random selection (portal message, dialogue snipper, third person caregiver...)
- Channel random selection (phone call, email, in person...)
- Random selection of: age group, message length, urgency...
- Random selection of 0-3 of the emotions

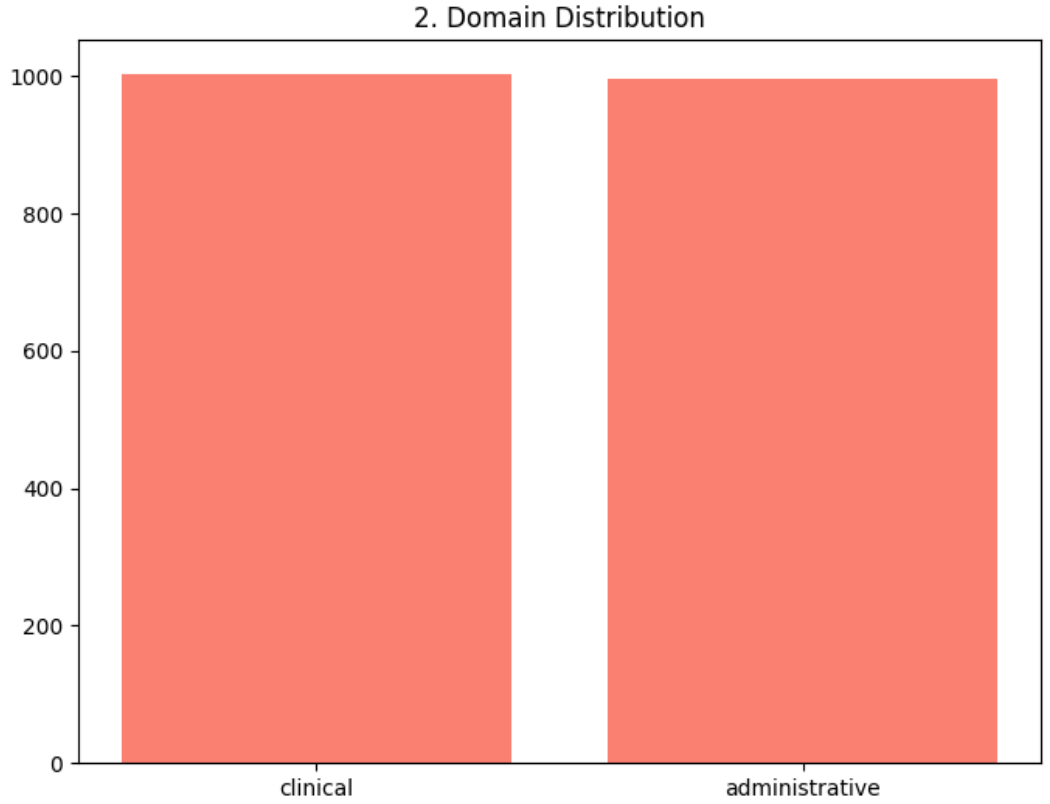
Targeted Augmentation – Filling the Gaps

After viewing the generated dataset and seeing where which emotions lack representation, we used an augmentation script which fills the gaps:

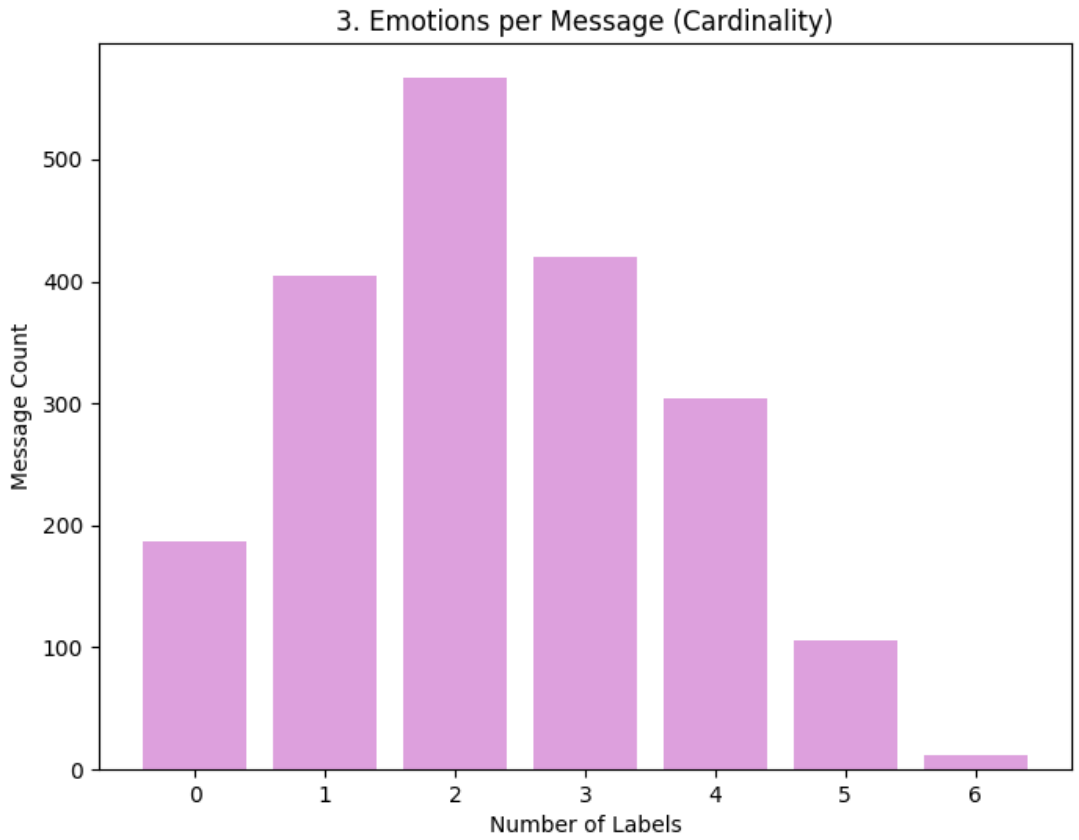
- Neutral messages (emotion-less)
    - Adding more neutral messages – improves false-positive control)
- Satisfaction and Anger
    - Lacked examples
- Mixed/resolved examples
    - More examples of multi-emotion messages

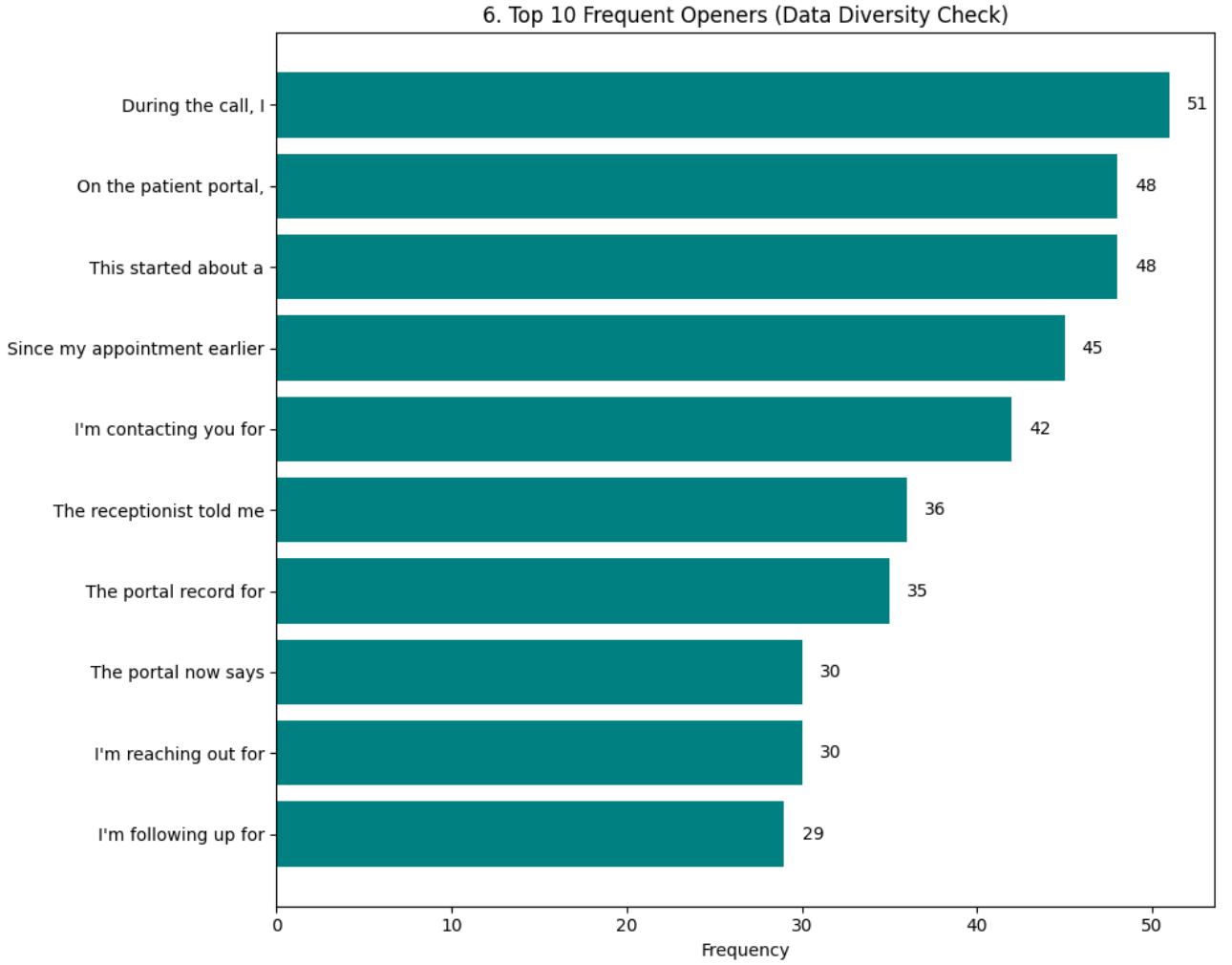
DATA EDA


1. Emotion Frequency
Count
1000
800
600
400
200
0
anxiety
confusion
frustration
anger
disappointment
satisfaction


2. Domain Distribution
1000
800
600
400
200
0
clinical
administrative

DATA EDA


3. Emotions per Message (Cardinality)
Message Count
Number of Labels
0
100
200
300
400
500
0
1
2
3
4
5
6


6. Top 10 Frequent Openers (Data Diversity Check)
During the call, I
On the patient portal,
This started about a
Since my appointment earlier
I'm contacting you for
The receptionist told me
The portal record for
The portal now says
I'm reaching out for
I'm following up for
51
48
48
45
42
36
35
30
30
29
0
10
20
30
40
50
Frequency

MODELS

BASELINE

Backbone: DistilBert

Data size: 1000 samples

FINAL MODEL

Backbone: ClinicalBert

Data Size: 2000 samples

COMPARING THE MODELS

ClinicalBert's specialized medical vocabulary allows representing complex clinical terms as single/fewer tokens, which means that ClinicalBert understands the clinical terminology better than the generic DistilBert.

This difference was significant enough to show better results after the model creation, which will be presented later on (Slide 12).

FINAL MODEL ARCHITECTURE

ENCODER - (TRANSFORMER)

In addition to the domain-specificity of ClinicalBert we talked about in the previous slide,

Unlike DistilBERT's lightweight 6-layer architecture designed for inference speed, ClinicalBERT utilizes a full 12-layer Transformer encoder with 110M parameters, allowing for significantly deeper hierarchical representation of complex clinical semantics.

SIGMOID VS SOFTMAX

We chose sigmoid over SoftMax because we are going for multi-label classification but we want to classify each emotion by itself in a scale of 0-1 and not spread the range between the emotions.

THRESHOLDS

We tuned a Neutral Threshold Guardrail (0.35), if none of the emotions manage to reach the threshold, the prompt is defined as neutral with no emotions.

At first we evaluated the model based on a single threshold for all the emotions (0.5).
Based on the metrics we got, we tuned each emotion's threshold accordingly.

METRICS

Percision

Of the emotions the model predicted, what fraction is correct.

Recall

Of the true emotions in the prompt, what fraction the model successfully predicted.

F1

A single score that balances Precision and Recall.

RESULTS

Final Model (ClinicalBert)

	Precision	Recall	F1-Score
Anger	0.66	0.59	0.62
Anxiety	0.97	0.87	0.92
Confusion	0.92	0.85	0.89
Disappointment	0.87	0.75	0.81
Frustration	0.95	0.82	0.88
Satisfaction	0.73	0.67	0.70
Micro-F1 Average	0.89	0.80	0.84
Macro-F1 Average	0.85	0.76	0.80

Baseline (DistilBert)

	Precision	Recall	F1-Score
Anger	0.76	0.70	0.73
Anxiety	0.81	0.57	0.67
Confusion	0.73	0.81	0.77
Disappointment	1.00	0.71	0.83
Frustration	0.94	0.82	0.87
Satisfaction	0.93	0.89	0.91
Micro-F1 Average	0.84	0.73	0.78
Macro-F1 Average	0.86	0.75	0.79

The Baseline overfitted Anger and Satisfaction because of a mixture of politeness in angry messages in the synthetic data.

For example: Ending the prompt with "thanks in advance" made him include satisfaction in the emotions even though it was just politeness and not satisfaction.

CONCLUSION

Achieved desired results:

- Improvement over the baseline.
- ClinicalBert fits better than DistilBert.

What we'd do differently?

- Expand the dataset beyond 2,000 prompt to increase coverage and diversity.

Future ideas:

- Evaluate on real telemedicine prompts to test the model against real life data.

THANK YOU