NATURAL LANGUAGE PROCESSING

KORNELIA FLIZIK
FEDOR CHURSIN
MATEY NEDYALKOV
PANNA BLANKA PFANDLER

TABLE OF CONTENTS

PROJECT PIPELINE

DATA

- Training data collection
- Training data processing

EMOTION CLASSIFICATION

- Model training and evaluation

SPEECH TO TEXT

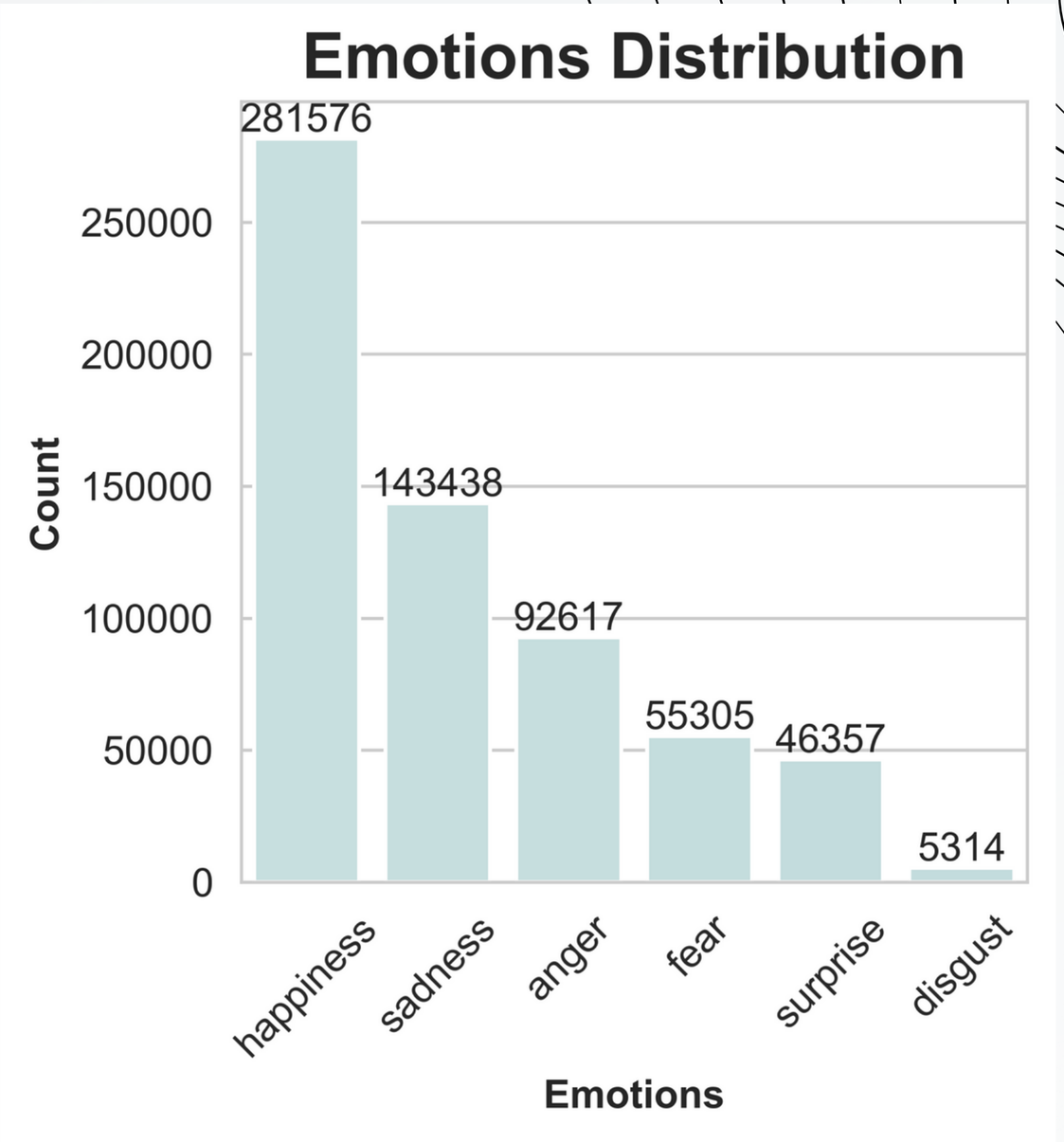
- Client data processing
- Transcribing audio with pretrained model

SOLUTION

- Emotion classification of client data

Data collection

Dataset name	Source
GoEmotions	Reddit comments
Smile Twitter Emotion	Twitter mentions
Friends Emotion-Labeled Dialogues	"Friends" TV show utterances
MELD dataset	"Friends" TV show utterances from dialogues
Carer dataset	English tweets



EDA

- Annotation
- Lemmatization
- Merging all the data
- Tokenization
- Label Encoding

Emotion	Label
happiness	0
anger	1
disgust	2
sadness	3
surprise	4
fear	5

MODEL SELECTION

MACHINE LEARNING	NEUTRAL NETWORKS	TRANSFORMERS
NAIVE BAYES	MULTI-LAYER PERCEPTRON	BERT
LOGISTIC REGRESSION	LSTM	ALBERT
SUPPORT VECTOR MACHINE		ROBERTA

ROBERTA MODEL

Transformer model with 24 layers

Pretrained on a combination of five massive datasets resulting in a total of 160 GB of text data.

Employs a dynamic masking strategy during training for more reliable and adaptable word representations.

EMOTION CLASSIFICATION MODEL

Input

X: tokens using Roberta Tokenizer

Y: one of 6 classes

Compiling

Sparse categorical cross-entropy

Adam with learning rate: 5e-5

Training

Training time: 4 epochs

Batch size of 128

Early stopping monitoring validation loss

WHISPER MODEL

Created by OpenAI and first released as open-source software in September 2022.

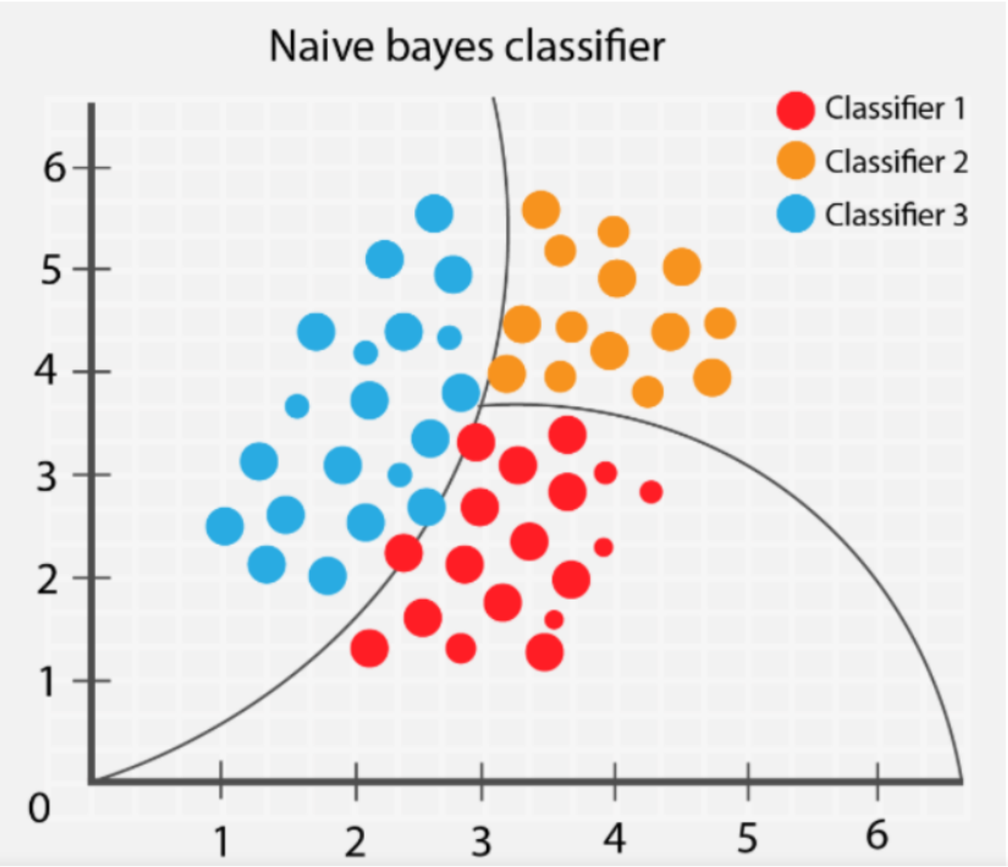
Transformer model that processes audio chunks into log-Mel spectrograms.

It is capable of transcribing speech in several languages, and translating speech into English.

TRAINING RESULTS

Model Type	Loss	Accuracy	F1 Score
Naive Bayes	0.62	80%	0.565
Logistic Regression	0.8	65%	0.521
Multilayer Perceptron	1.09	76%	0.459
LSTM	1.07	63%	0.573
roBERTa	0.34	87%	0.733

TRAINING RESULTS - NAIVE BAYES



Model Type	Loss	Accuracy	F1 Score
Naive Bayes	0.62	80%	0.565

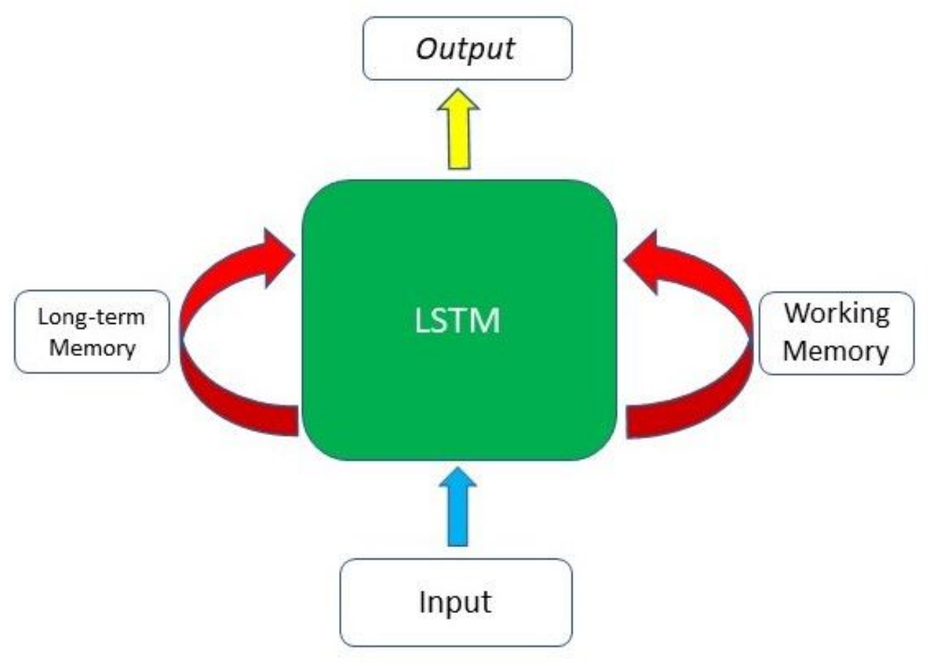
Strengths:

- Simplicity & Speed
- Baseline performance

Limitations:

- Feature independence
- Unseen features

TRAINING RESULTS - LSTM



Model Type	Loss	Accuracy	F1 Score
LSTM	1.07	63%	0.573

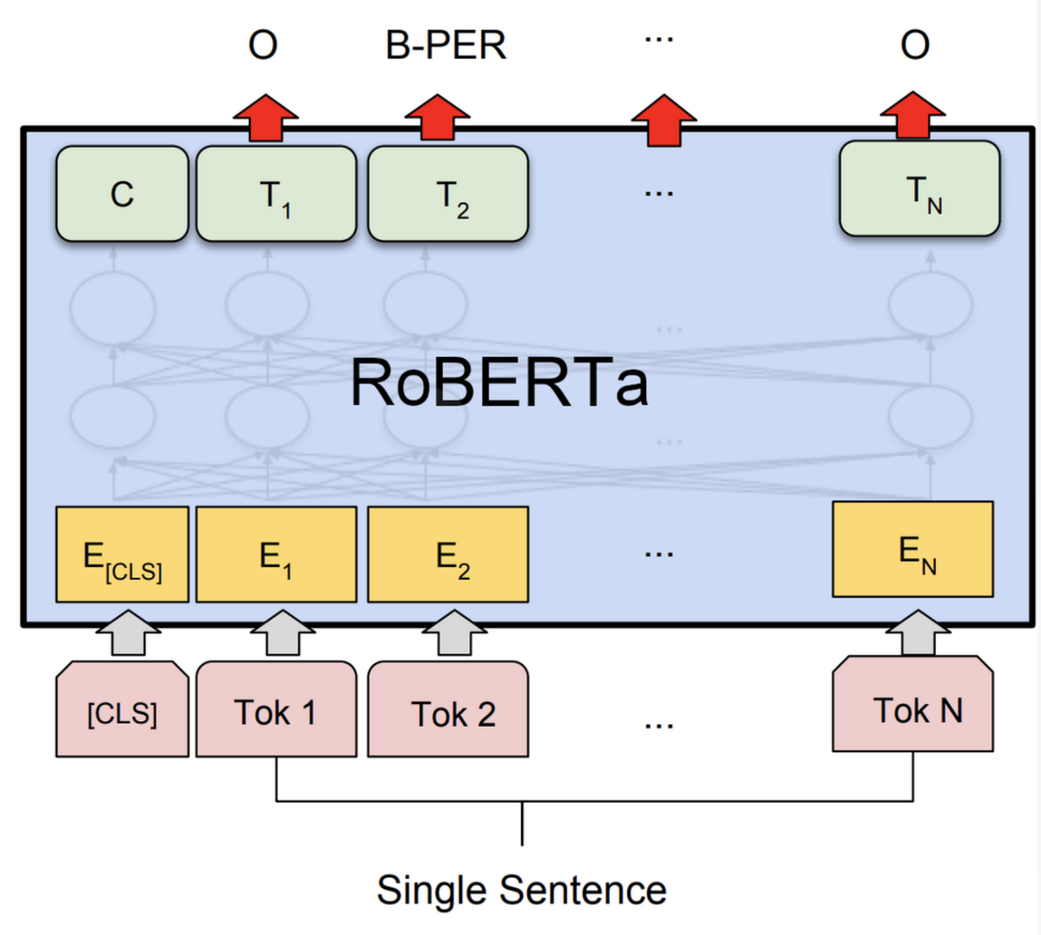
Strengths:

- Temporal dynamics
- Memory

Limitations:

- Computational cost
- Tuning complexity

TRAINING RESULTS - ROBERTA



Model Type	Loss	Accuracy	F1 Score
roBERTa	0.34	83%	0.733

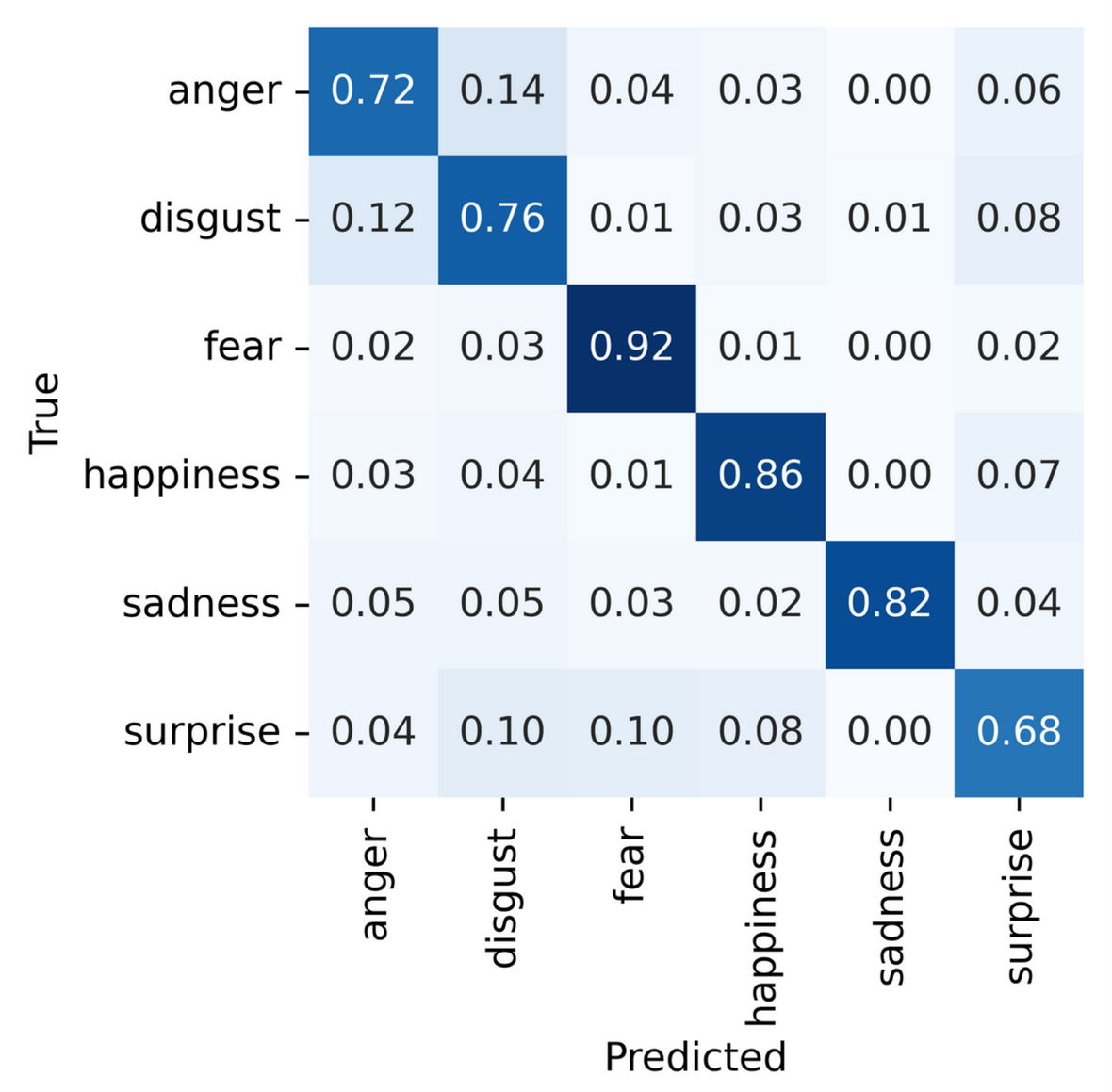
Strengths:

- State-of-the-art performance
- Pre-trained knowledge

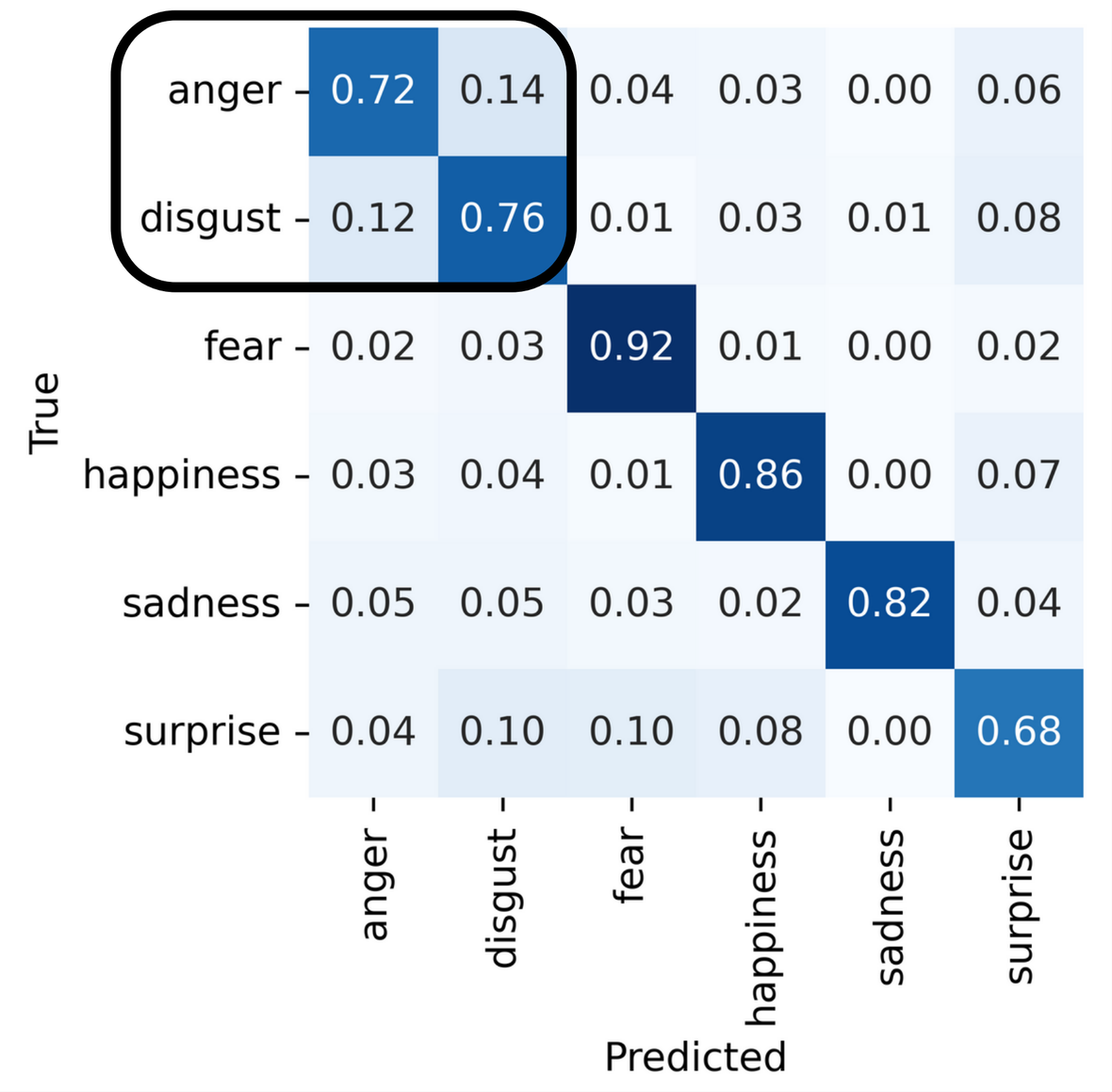
Limitations:

- Computational resources
- Black box nature

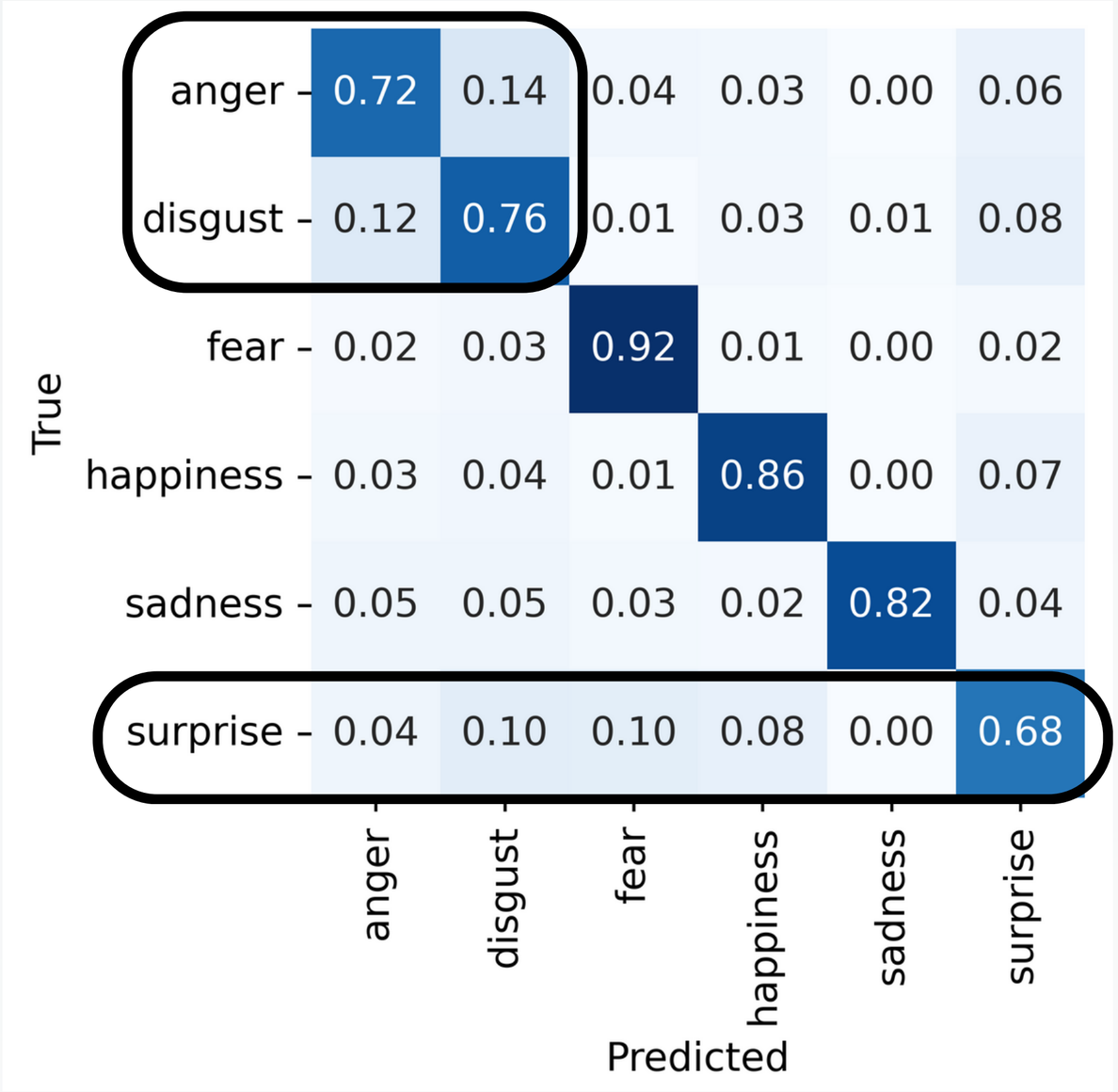
Error Analysis



Error Analysis



Error Analysis



Error Analysis

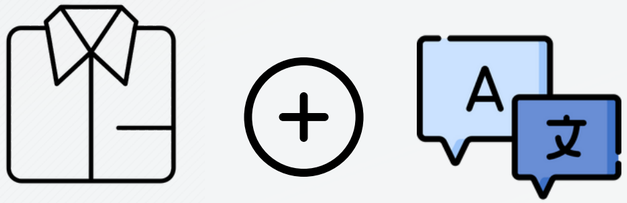
Slang-heavy dataset

Hallucinations

Model Type	Loss	Accuracy	F1 Score
roBERTa	0.34	87%	0.733

Error Analysis

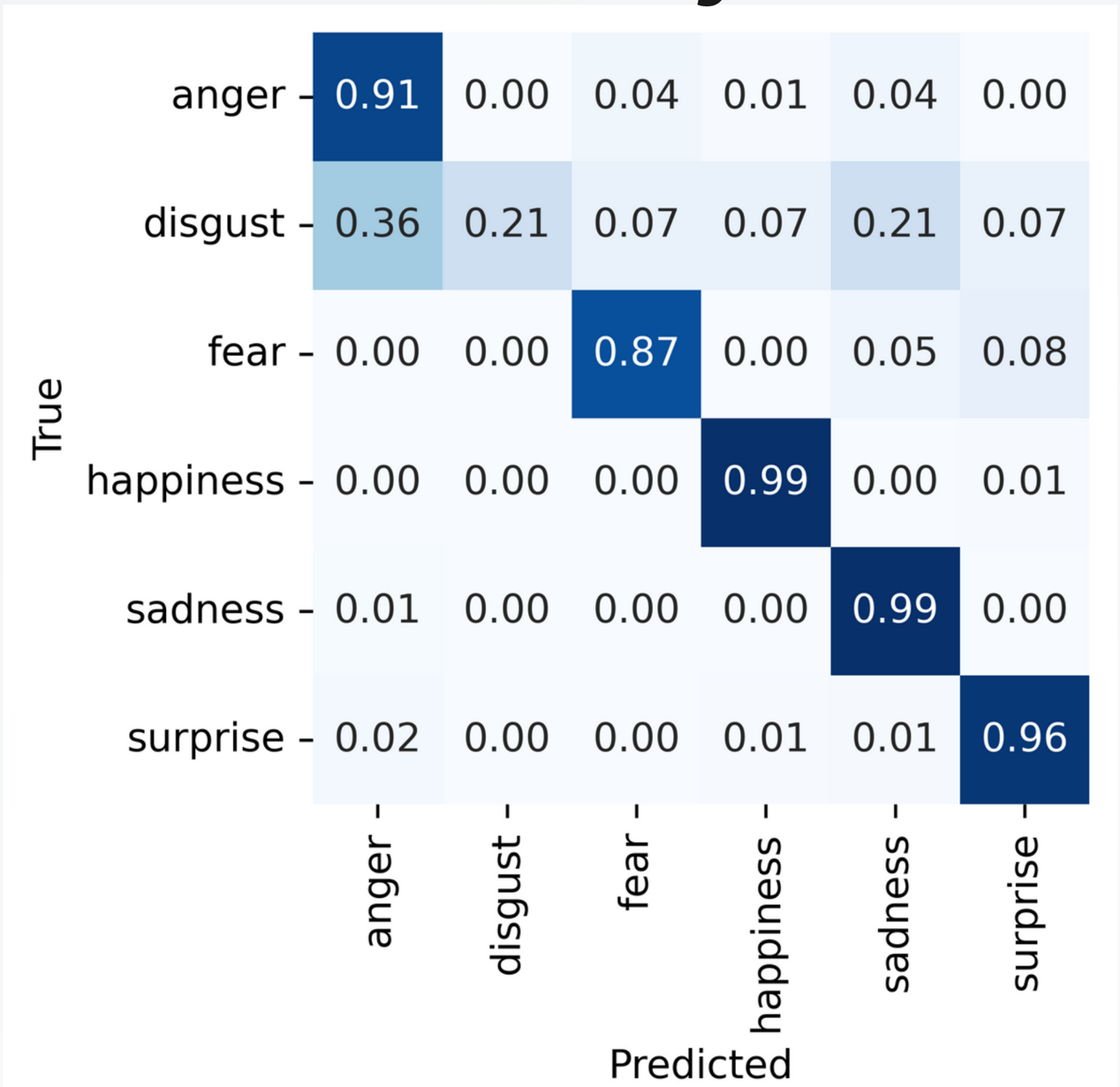
Formal language dataset



Efficient learning process

Model Type	Loss	Accuracy	F1 Score
roBERTa	0.07	96%	0.580

Error Analysis



Application

Show Episode

Whisper

Preprocessing

Application Limitations

Show Episode

Whisper

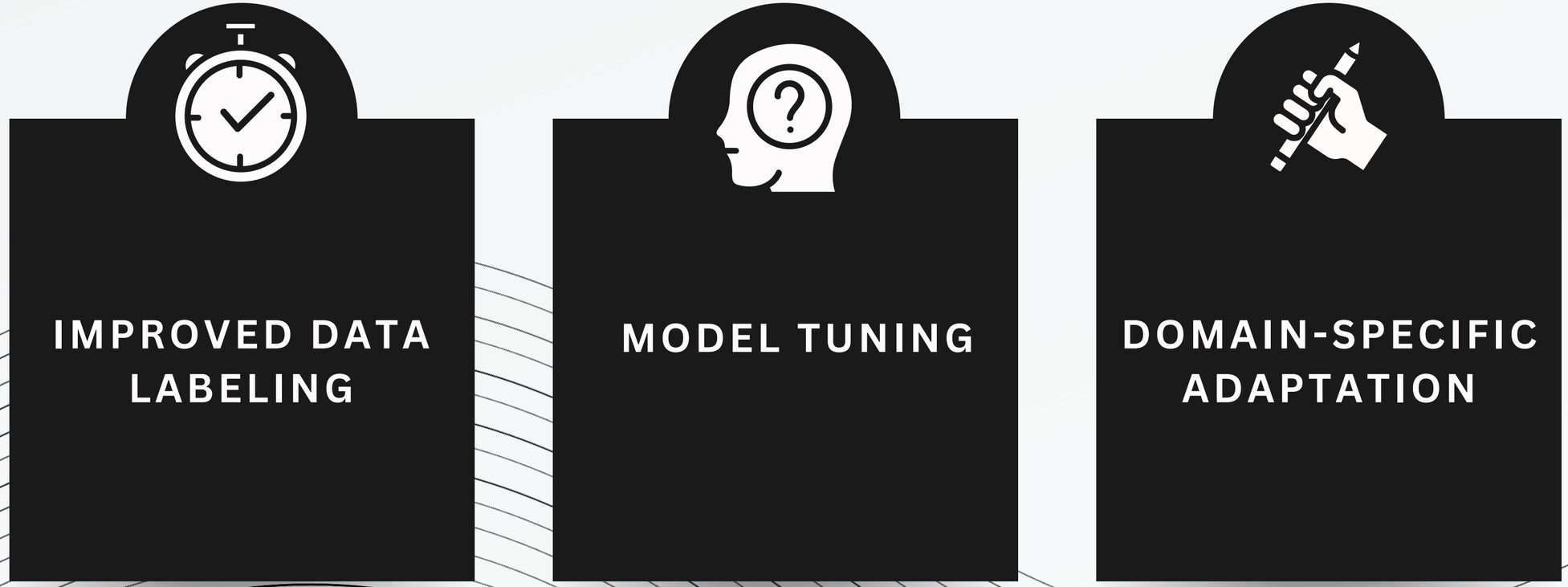
Preprocessing

Context Loss

Slow Performance

High Memory Load

RECOMMENDATIONS



THANK YOU FOR YOUR ATTENTION

Any questions?