# Emotion Detection Pipeline for Polish Reality Television

INSIGHTS FROM MASTERCHEF POLAND

OCTOBER 2025
ZOFIA PILITOWSKA, KINGA MARCHLEWSKA, ARON WOJCIECHOWICZ

# The Challenge: Understanding Emotional Narratives

1

**Domain: Polish Reality Television**

High-stakes competition programs like Masterchef generate intense emotional moments that drive viewer engagement

2

**Use Case**

Automatically detect and analyze emotions in spoken dialogue to understand:

- Emotional peaks that correlate with viewer retention
- Character emotional arcs throughout episodes
- Tension patterns that predict viral moments

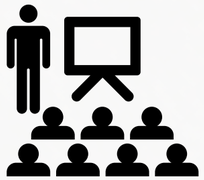
# Business Value: Why This Matters

**Efficiency Gains**

Reduce manual content tagging time through automation

**Targeted Marketing**

Create clips highlighting peak emotional moments for social media promotion

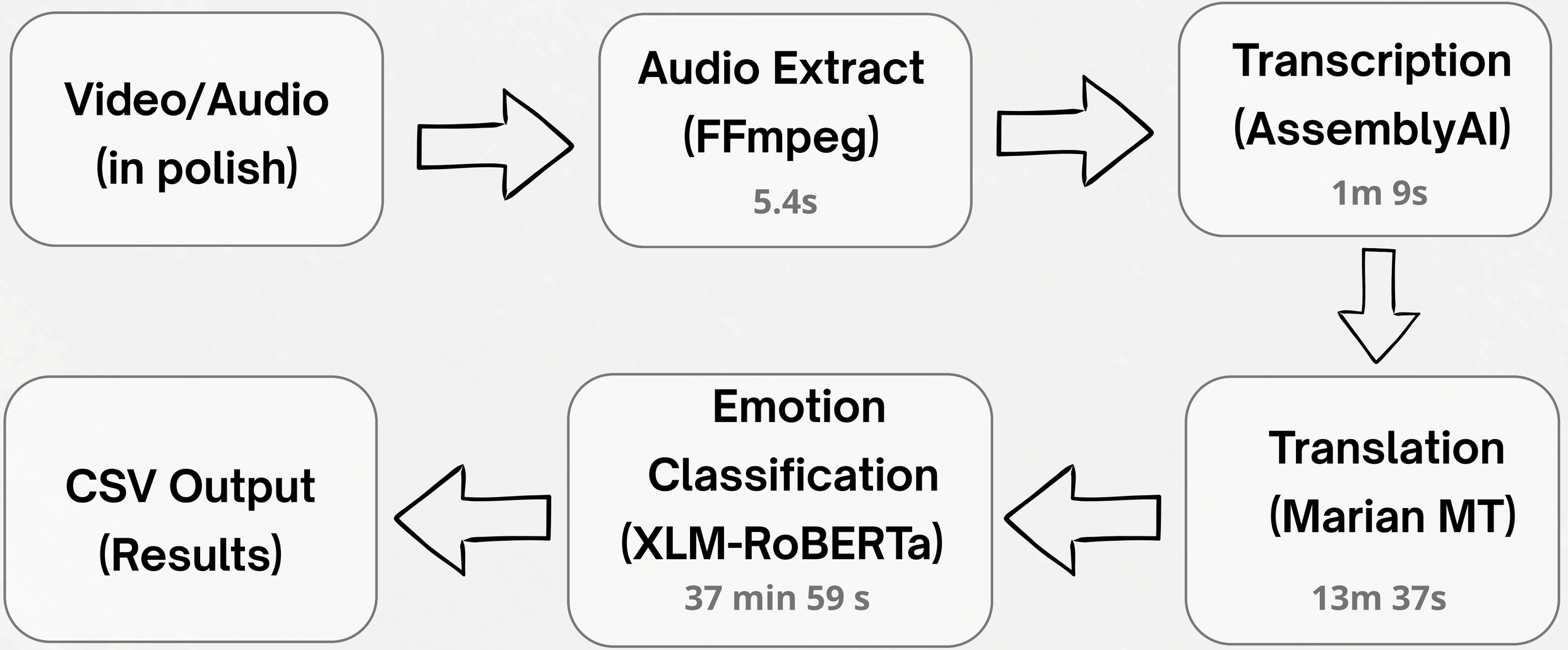


**Content Optimization**

Identify which emotional moments drive engagement and optimize editing accordingly

# — Pipeline overview

Test Case: 45-minute MasterChef Poland Episode



5.

# — Pipeline output

| Start Time | End Time | Sentence | Translation | Emotion |
|---|---|---|---|---|
| 00:00:00,042 | 00:00:01,762 | Kto ostatni puści ręce, wygrywa. | The last one to let go of his hands wins. | neutral |
| 00:00:01,762 | 00:00:02,382 | Czas, start! | Time to go! | neutral |
| 00:00:02,382 | 00:00:03,222 | Przepraszam, wolało. | I'm sorry.  I prefer that. | sadness |
| 00:00:03,222 | 00:00:05,002 | Ale to jest wspaniałe | It's wonderful! | happiness |

# — Data Landscape

### Training Data: TwitterEmo Dataset

- Large-scale Polish emotion corpus capturing informal, emotional language

- **Strengths:**

Natural emotional expression, diverse Polish language patterns

- **Size:**

Robust sample representing Polish emotional vocabulary competition format

### Test Data: MasterChef Transcripts

- Real spoken dialogue from Polish reality TV

- **Characteristics:**

Interruptions, fillers ('yyy', 'no'), dialect variation, fast-paced commands, emotional interviews

- **Context:**

High-pressure cooking environment with structured competition format

# — Data Challenges

### Emotion Imbalance

Neutral content dominates (50.9%), while crucial emotions are rare:
→ Fear: 0.9%
→ Trust: 4.5%
→ Sadness: 4.6%

### Domain Mismatch

The training data (Twitter) differs from the test data (reality show transcripts)

### Noisy Real-World Data

Spoken TV language includes interruptions, fillers, dialects and informal phrasing

# — Solution Strategy

### Class-Weighted Loss

Computed weights inversely proportional to class frequency, applied them in CrossEntropyLoss to give rare emotions higher importance.

### Data Cleaning, Preprocessing

Normalized transcripts and removed unnecessary fillers or symbols.

### Consistent Tokenization

Used AutoTokenizer from Hugging Face with bert-base-multilingual-cased to handle informal expressions, diacritics, multilingual content uniformly.

8.

# — Models we tested:

### Logistic Regression

Accuracy: 0.06

Precision: 0.196

Recall: 0.799

F1 Score: 0.311

### RNN

Accuracy: 0.481

Precision: 0.332

Recall: 0.337

F1 Score: 0.308

### LSTM

Accuracy: 0.385

Precision: 0.391

Recall: 0.344

F1 Score: 0.3826

### Naive Bayes

Accuracy: 0.474

Precision: 0.590

Recall: 0.474

F1 Score: 0.445

### SVM

Accuracy: 0.546

Precision: 0.510

Recall: 0.546

F1 Score: 0.453

### Transformers

Accuracy: 0.623

Precision: 0.666

Recall: 0.586

F1 Score: 0.6113

9.

# — Model Architecture: Why Transformers?

Modern neural networks designed specifically for language understanding

Captures long-range emotional context

Adapts to informal, messy spoken language

Understands context beyond individual words

Multi-label emotion detection

Handles interruptions and non-linear dialogue

# Model evaluation

### Misclassified samples tend to have:

- Longer sentences: 38 vs 29 characters (correct)
- More words: 7.4 vs 5.6 words (correct)
- More punctuation: Questions & exclamations
- Model struggles with complex emotional expressions

### Stats

Accuracy: 0.623

Precision: 0.666

Recall: 0.586

F1 Score: 0.6113

## per-emotion performance

| Emotion | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| happiness | 0.928 | 0.666 | 0.730 | 0.696 |
| sadness | 0.960 | 0.637 | 0.398 | 0.490 |
| anger | 0.883 | 0.694 | 0.583 | 0.634 |
| surprise | 0.938 | 0.517 | 0.541 | 0.529 |
| fear | 0.992 | 0.645 | 0.294 | 0.404 |
| disgust | 0.854 | 0.655 | 0.767 | 0.707 |
| neutral | 0.815 | 0.849 | 0.787 | 0.817 |

11.

# — Possible limitations

## Technical

Processing Speed: Real-time analysis requires heavy GPU infrastructure It takes time for pipeline to finish.

Transcription Dependency: Model accuracy relies on high-quality speech-to-text preprocessing

Language-Specific: Currently optimized only for Polish – requires retraining for other languages

## Operational

Domain Adaptation: Performance may vary on shows with different formats (talk shows vs. competitions)

Cold Start Problem: New show formats need initial human validation period

Model is Food show specific, so it needs special categorization before using.

# Ethical consideration

PRIVACY & FAIRNESS CONCERNS

- Consent gaps
- Training data bias
- Contestant wellbeing

TRANSPARENCYISSUES

- Black box model
- Interpretability
- Automated without human oversight

ENVIRONMENTAL RESPONSIBILITY

- Training: 6.89 kg $CO_2$ (Poland grid)
- Deployment: 58 kg $CO_2$/year at scale
- Mitigation: Renewable energy, compression

# Next Steps & Recommendations

DOMAIN ADAPTATION

Collecting 500-1000 polish reality TV transcripts with emotion labels

MODEL ARCHITECTURE IMPROVEMENTS

- More sophisticated fine-tuning strategies
- Ensemble methods
- Adding context windows
- Incorporating speaker information

EXPLORE ALTERNATIVE APPROACHES

→ for example GPT-4 few-shot prompting
→ fine-tuning on English TV, then adapt to Polish

# The End

THANK YOU FOR LISTENING