NLP: Emotions Classification Model for Arabic Language

Group 5 Arabic

Ashraf Nahlouse *Musaed Alfareh*

234669 225739

Table of contents:

Topic, Domain and Use Case

Topic

Emotion classification from spoken dialogue using Natural Language Processing (NLP)

Domain

Transcribed television show data across Arabic language

Focus

Analyzing emotional content in spoken media for insight and automation

Use Case

Addressing limitations in Arabic NLP by improving model generalization across diverse dialects

The Purpose

Manual annotation is time-consuming, subjective, and unscalable

Goal

Build an end-to-end pipeline that outputs emotion labels from raw video along side the translation and transcription

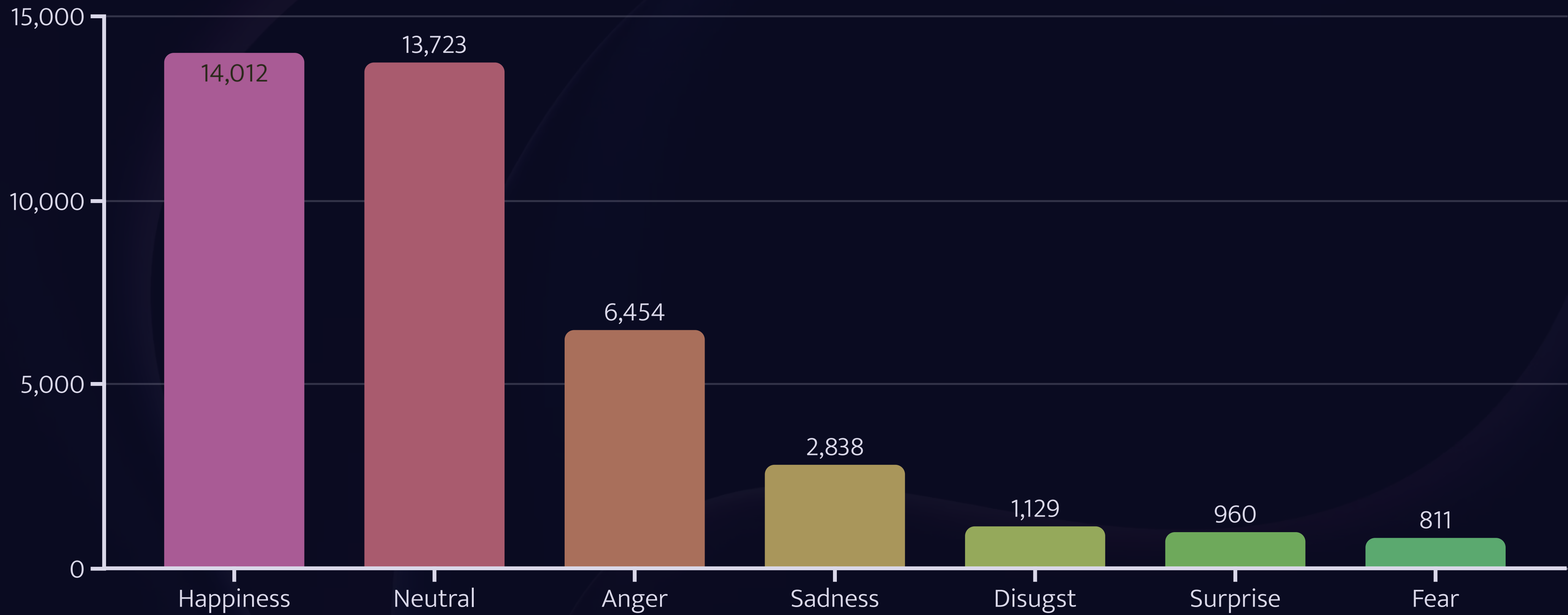
1. Input Video
2. Translate to English
3. Tag 6 core emotions + Neutral
4. Output file with Transcriptions, Translations and Emotions

The Value for the Client

- Explored limitations of Arabic processing
- Suggested actionable improvements for Arabic NLP
- Provided direction for future pipelines

Data & Limitations

- Training dataset: GoEmotions with an Arabic translation containing 43,000 sentences.
- Cleaned dataset: GoEmotions containing over 39,000 sentences.
- Test set : The agency dataset with more than 465 sentences



Model Overview & Selection Rationale

Logistic Regression & Naive Bayes

Simple baseline models to establish a performance reference

Fast to train, but struggled with complex language

LSTM & BiLSTM (RNNs)

Chosen for their ability to handle sequential, spoken text

Performed better than traditional models, but still affected by class imbalance

AraBERT & MARBERT (Transformers)

Pretrained on large Arabic datasets

Captured deeper meaning and context without manual features

MARBERT outperformed others on informal and dialectal data

Model Performance & Evaluation

We Applied 4 Iterations to Improve Each Model's Accuracy

The reliability score "F1" of the best models is

Logistic Regression

66

Combines **TF-IDF + simple features** (word/char/punctuation counts)

Uses **ensemble stacking** with multiple logistic models

Shows high **accuracy -F1 improvement** and, balance across classes

Naive Bayes

45

Uses **TF-IDF + sentiment + sentence length + avg word length**

α = 0.01 smoothing, **MaxAbsScaler** for numeric features

Adds linguistic cues, but still **fails on minority classes**

LSTM

69

Uses **larger vocab & sequence length** (7000 words, 200 tokens)

Higher **embedding dimension (256)** and **dropout (0.3)** for better generalization

Marginally better recall for minority classes vs basic LSTM

Bert

RNN

The reliability score of our two best models is

70

67

Translated and merged datasets to **include all emotion categories**, especially "disgust"

Preprocessed using `arabert.preprocess` for **linguistic consistency across datasets**

Increased **vocab size (7000)** and **sequence length (200)** allowed for better language representation

Used **embedding dim = 256** and higher **dropout (0.3)** for better regularization

Achieved **marginal improvement in recall**, especially for underrepresented emotions

MARBERT

fine-tuned on stratified data with large batch size

73

- MARBERT is **specifically optimized for Arabic dialects and social media**, making it ideal for informal data
- Used **batch size of 128**, **sequence length = 128**, and **efficient GPU utilization**
- Fine-tuned using **early stopping and F1 as the selection metric**
- Strong overall performance, but **macro F1 still low** — indicates difficulty in minority class prediction

Key Performance Insights

Performs **very well on "neutral" class** high precision and recall (F1 = 86%)

Model is **reliable for common emotions**, but needs improvement for rare ones

Weak on minority classes like **disgust, fear, and surprise** due to limited training data

Overall accuracy (~76%) is skewed by the dominance of the neutral class

Weighted F1-score (0.7373) better reflects the balance across all predictions

Emotion	Precision	Recall	F1-Score	Support
Neutral	0.8210	0.9042	0.8606	355
Anger	0.3333	0.2727	0.3000	22
Sadness	0.5625	0.3750	0.4500	24
Happiness	0.2581	0.2353	0.2462	34
Surprise	1.0000	0.3636	0.5333	11
Fear	1.0000	0.1538	0.2667	13
Disgust	0.0000	0.0000	0.0000	3

Strengths

- **High precision and recall for neutral class**:

Precision: **0.82** Recall: **0.90** F1: **0.86**

- **Strong performance in majority class**
- **Surprise, fear, sadness** classes show **some precision**, even if recall is low

Weaknesses

- **Very low performance on minority classes**, especially:

Disgust: F1 = **0.00**

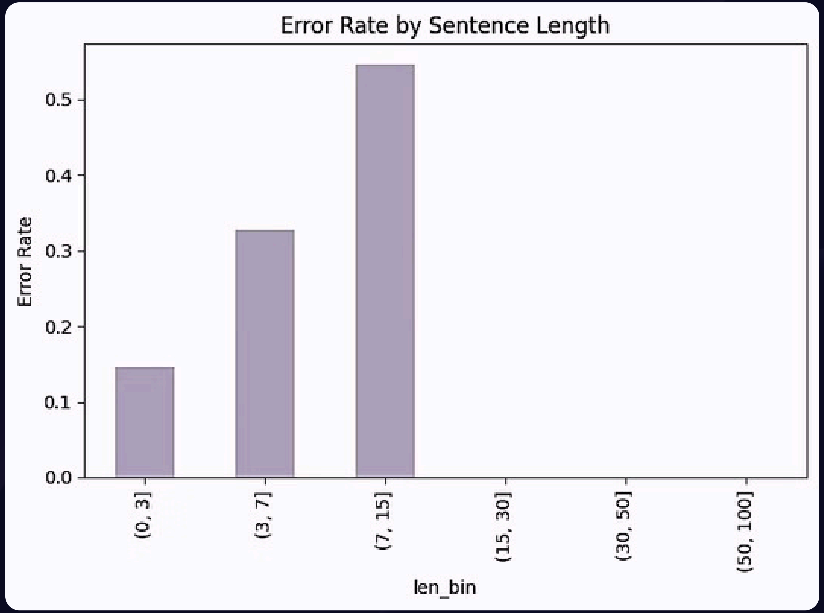
Fear: Recall = **0.15** F1 = **0.26**

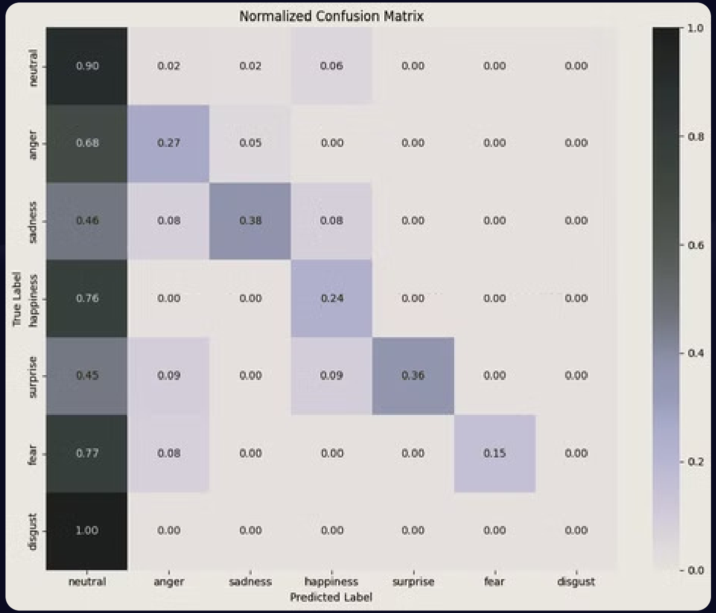
Surprise: Recall = **0.36**, F1 = **0.53**

- **Imbalance sensitivity**

Error Analysis

- The model often defaults to **“neutral”** when uncertain — even when the second-best prediction is correct
- **Longer sentences (7–15 tokens)** have the **highest error rate**, while shorter ones are more accurately classified
- **Vague or emotionally overlapping phrases** cause confusion (e.g., “أكبر من كده”, “العفو عند المقدرة”)
- Many incorrect predictions had **high confidence** (0.5–0.7), indicating **overconfidence** in wrong decisions





Made with Gamma

Explainable AI

What we did

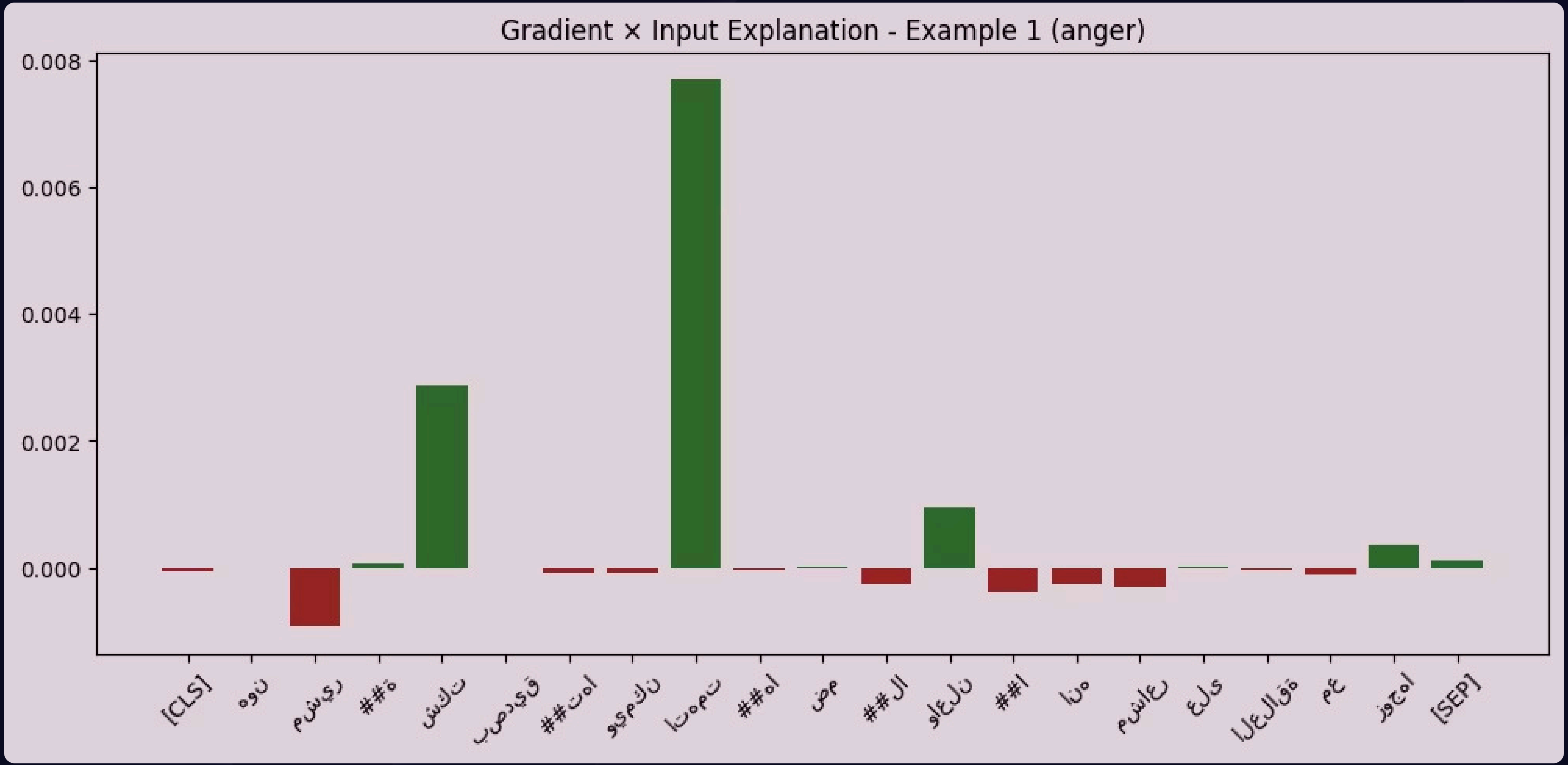
Applied XAI on our best Transformer model (MARBERT)

Used **3 techniques** to understand model decisions:

- Gradient × Input
- Layer-wise Relevance Propagation (LRP)
- Input Perturbation

What we found

1. Model **focuses on emotional keywords** (e.g. "حب", "شكت") but sometimes still predicts **neutral**
2. **Some irrelevant words** get high influence, while real emotion cues are missed
3. **When tokens are removed**, model confidence **drops sharply**, showing strong reliance on just a few words
4. XAI revealed **where the model works well**, and where it **fails to understand subtle emotions**


Gradient × Input Explanation - Example 1 (anger)
0.008
0.006
0.004
0.002
0.000
[CLS]
[SEP]

The Complete Pipeline

1. **Raw data (URL)**

2. **Video to Transcript**

   We have selected OpenAI Whisper over AI assembly

3. **Emotions predictions**

   MARBET

4. **Translated**

   Pre-trained model over the model from scratch

5. **Final csv file contains all the elements**

Why?

Translated "Pre-Trained"

- Fast
- High quality translation
- Better at generalization

Video to Transcript "Whisper"

- Lower error rate than assembly ai
- Clarity and coherence
- The freedom of model choice

Metric	Best From Scratch (Iteration 9)	Pretrained Model
BLEU	0.58	6.58
METEOR	0.0832	0.0191
TER	11.9997	1.1785

Start Time	End Time	Sentence	Translation	Emotion
00:00.000	00:03.000	بشيرة طمينة أموري كويسة وكل شيء تمام	I'm okay and everything's okay.	happiness
00:03.000	00:04.280	يا أهلا وسهلا	Welcome.	neutral

Ethical Considerations

Bias and Label Imbalance

The dataset was dominated by the *neutral* class, raising fairness concerns.

We prioritized the **weighted F1-score** over plain accuracy to better reflect minority class performance.

Cultural and Dialectal Sensitivity

Some Arabic expressions vary in meaning across dialects (e.g., Egyptian).

We used **XAI** to verify token-level decision patterns in order To avoid misclassification due to cultural nuances

Transparency and Interpretability

We applied **Gradient × Input**, **LRP**, and **Input Perturbation** to understand how predictions were made and to ensure accountability.

Overconfidence in Misclassifications

Our model often made incorrect predictions with high confidence, especially on minority emotions.

This informed our decision to assess **confidence distribution** alongside standard metrics like F1 and recall.

Possible Limitations

1

Dialect Limitation

The model was trained on only one Arabic dialect.

2

Dialect Diversity

There are over 30 Arabic dialects, which makes generalization difficult.

3

Reddit Source

Training data came from Reddit, which is primarily in English.

4

Real-World Testing

To address this, we tested the model on spoken Arabic dialects from video content.

Next Steps

1

Dialect Data Collection

Gather diverse spoken and written Arabic dialect datasets (e.g., Gulf, Egyptian, Maghrebi)

2

Dialect Adaptation Techniques

- Contrastive learning across dialects
- Few-shot or zero-shot transfer between dialects

3

Evaluation on Real-World Dialects

4

Community or User Feedback

- Involve native speakers or annotators to validate model outputs.
- Use their insights to refine the model or label edge cases.

Thank you for Watching