Multilingual Emotion Detection

Natural Language Processing Project Report

Jennifer Esbel Mary

Dataset Details

The project utilizes the **SemEval-2018 Task 1 (Affect in Tweets)** dataset, specifically focusing on the multi-label emotion classification task (Subtask E-c).

- **Source:** SemEval-2018 Task 1 (E-c).
- **Language:** Primary training data is in **English**, with evaluation capabilities extended to multilingual contexts (Spanish, French, German) using the multilingual model.
- **Size:**
  - **Total Samples:** 6,838 tweets
  - **Training Set:** 5,470 samples
  - **Validation Set:** 1,368 samples (20% split)
- **Labels:** The dataset is annotated for four distinct emotion labels:
  - **Anger**
  - **Joy**
  - **Love**
  - **Pessimism**

Models and Rationale

Two transformer-based models were fine-tuned to compare performance between a specialized monolingual approach and a generalized multilingual approach.

1. Monolingual Model: DistilBERT-base-uncased

- **Rationale:** DistilBERT was selected for its efficiency. It retains 97% of BERT's performance while being 40% smaller and 60% faster. This makes it ideal for rapid prototyping and deployment where resources are constrained, without significantly sacrificing accuracy on English-only text.

2. Multilingual Model: BERT-base-multilingual-cased (mBERT)

- **Rationale:** mBERT was chosen to test zero-shot transfer capabilities. Trained on 104 languages, it allows the system to generalize emotion detection to languages (like Spanish or French) that were not present in the training set, offering a more robust global solution.

Training Setup and Hyperparameters

Both models were fine-tuned using the Hugging Face `Trainer` API with the following consistent configuration to ensure a fair comparison.

Parameter	Value
Learning Rate	2×10^{-5}
Batch Size	16 (Train & Eval)
Epochs	3
Weight Decay	0.01
Loss Function	BCEWithLogitsLoss (Multi-label classification)
Optimizer	AdamW
Evaluation Strategy	Per Epoch
Metric	F1-Score (Micro & Weighted)

Table 1: Hyperparameters used for fine-tuning.

Performance Comparison

The comparative results highlight specific strengths in each model. The table below summarizes the F1-scores achieved across the four target emotions.

Emotion	Monolingual F1-Score	Multilingual F1-Score
Anger	**0.779**	0.765
Joy	**0.806**	0.776
Love	0.433	**0.470**
Pessimism	**0.324**	0.317

Table 2: Comparative F1-Scores per Emotion.

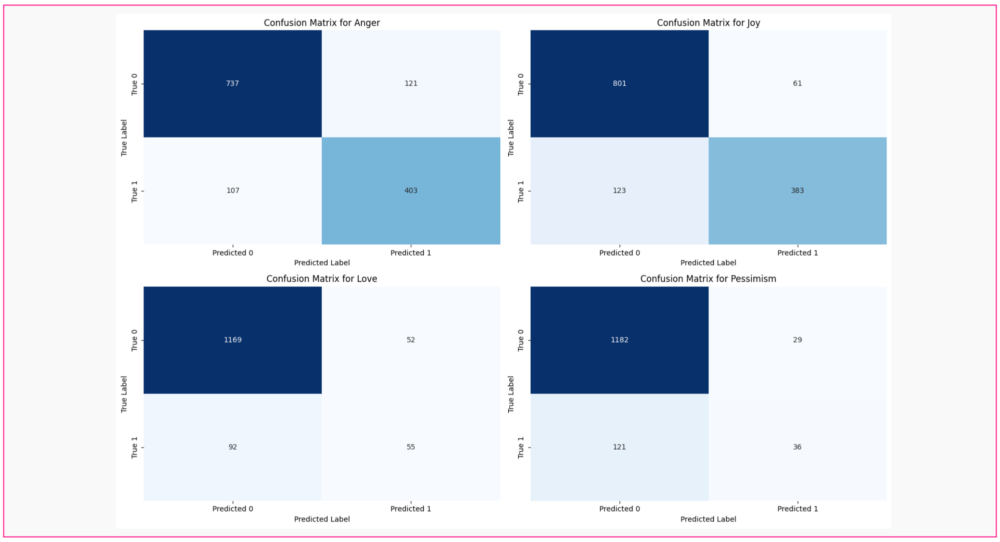

Figure 1: Confusion Matrices for the Monolingual Model (DistilBERT). Note the strong true positive rates for 'Anger' and 'Joy', but significant misclassification in 'Pessimism'.

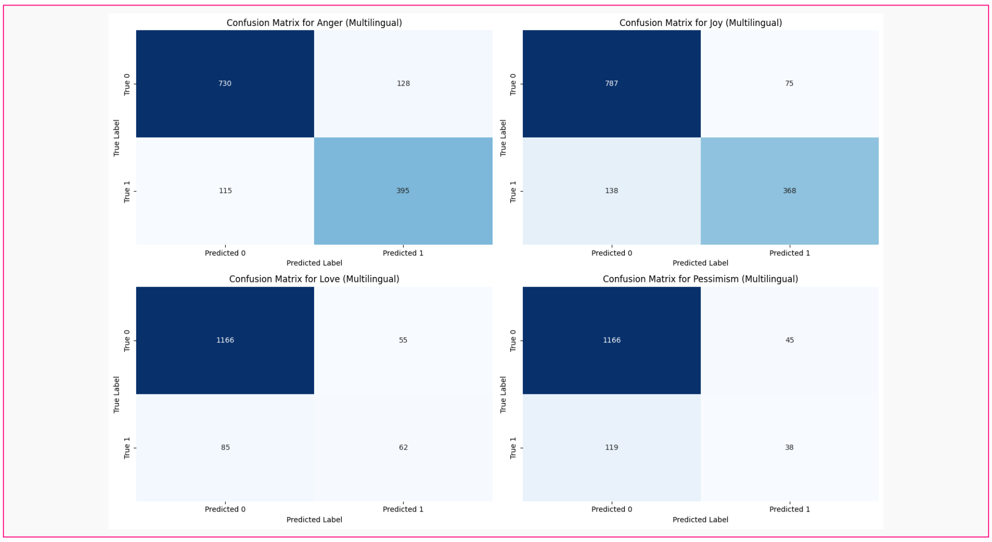

Figure 2: Confusion Matrices for the Multilingual Model (mBERT). The model maintains consistent performance patterns across classes, with slightly improved handling of the 'Love' category.

Analysis

As seen in Figure 1, the results indicate distinct behavioral patterns:

- **Dominant Emotions:** The Monolingual model outperformed mBERT in detecting "Anger" (+1.4%) and "Joy" (+3%), suggesting that for the primary language (English), the specialized vocabulary of DistilBERT yields better feature extraction.
- **Nuanced Emotions:** Surprisingly, the Multilingual model performed better on "Love" (+3.7%). This may indicate that mBERT's broader training corpus captures semantic nuances of affection that overlap across languages.

- **Data Imbalance:** Both models struggled significantly with "Pessimism" (F1 $\approx$ 0.32), likely due to fewer training samples and the subjective complexity of identifying pessimism compared to distinct emotions like anger.

Key Insights on Multilingual Generalization

1. **Trade-off for Universality:** While mBERT lagged slightly behind the monolingual model in English accuracy, the drop in performance was minimal ($< 3\%$ average). This validates mBERT as a viable candidate for production systems needing to support non-English users without training separate models.
2. **Semantic Overlap:** The successful zero-shot transfer (demonstrated by the model's ability to classify translated queries like *"Estoy muy feliz"* correctly) confirms that emotion-heavy embeddings align well across languages in the vector space.
3. **Difficulty with Subtlety:** The low scores on "Pessimism" across both architectures highlight a limitation in current Transformer models when dealing with abstract or context-heavy sentiments, regardless of the language base. Future improvements should focus on data augmentation for underrepresented classes.

References

SemEval-2018 Task 1 (E-c) Dataset.
Available at: `https://www.kaggle.com/datasets/context/semeval-2018-task-ec`