SemEval 2025-Task 11: Bridging the Gap in Text-Based Emotion Detection

Shamsuddeen Hassan Muhammad*, Nedjma Ousidhoum*, Idris Abdulmumin, Seid Muhie Yimam, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine De Kock, Tadesse Destaw Belay, Ibrahim Said Ahmad, Nirmal Surange, Daniela Teodorescu, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino Ali, Vladimir Araujo, Abinew Ali Ayele, Oana Ignat, Alexander Panchenko, Yi Zhou, Saif M. Mohammad

https://github.com/emotion-analysis-project/SemEval2025-Task11

Motivation

- **Human communication**
  is deeply emotional

- **Multilingual and cultural challenges**
  Emotional expression varies across **languages, cultures,** and **contexts** (perceived subjectively).

- **Our task**
  Text-based emotion detection across cultures and language

SemEval 2025 Task 11: Text-Based Emotion Detection

Focuses on **perceived emotions**

Predict… *what emotion most people will think the speaker may be feeling, given a sentence or a short text snippet uttered by the speaker.*

Task Setup

- **Track A (Multi-label Emotion Detection)**
  Classes: *joy, sadness, fear, anger, surprise, and disgust*

- **Track B (Emotion Intensity Detection)**
  Classes: 0, 1, 2, or 3

- **Track C (Cross-lingual Emotion Detection)**
  Classes: *joy, sadness, fear, anger, surprise, and disgust*

Evaluation Metrics

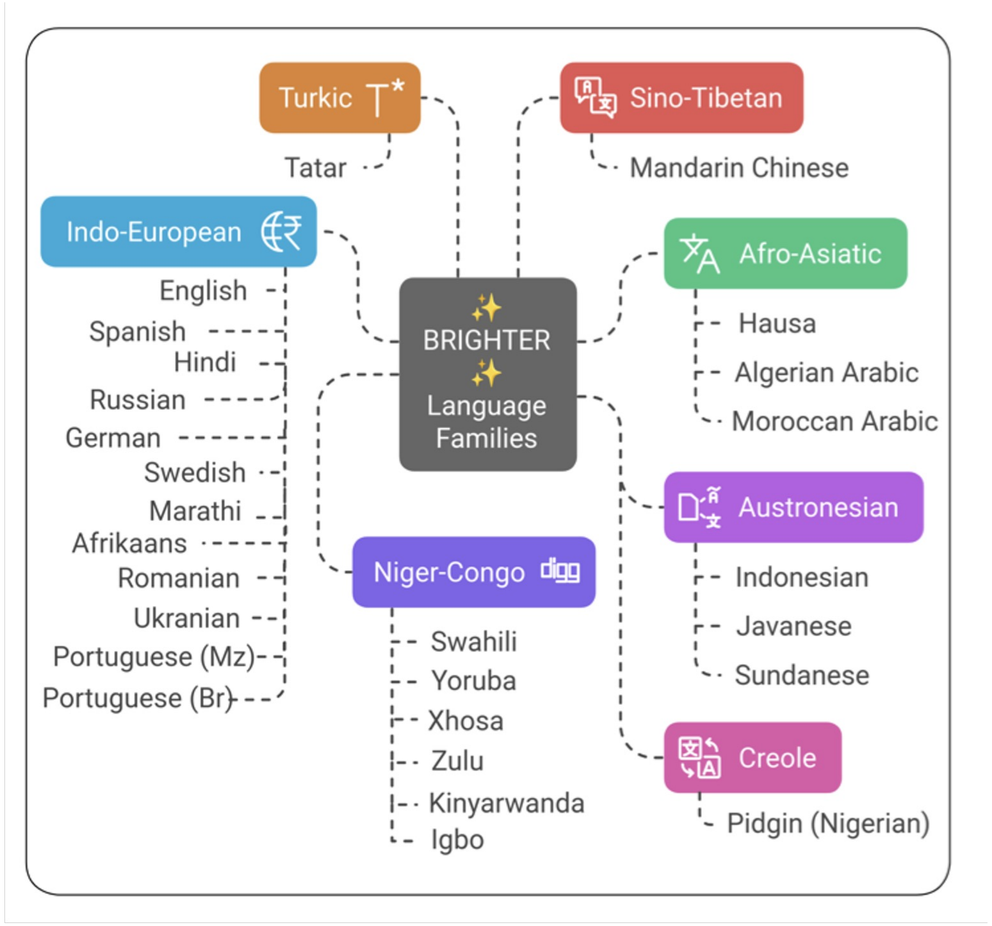
1. Average macro F1 score
2. Pearson correlation coefficient

Baseline

1. Majority class
2. Fine-tuned RoBERTa

Languages

32



Dataset Example

32

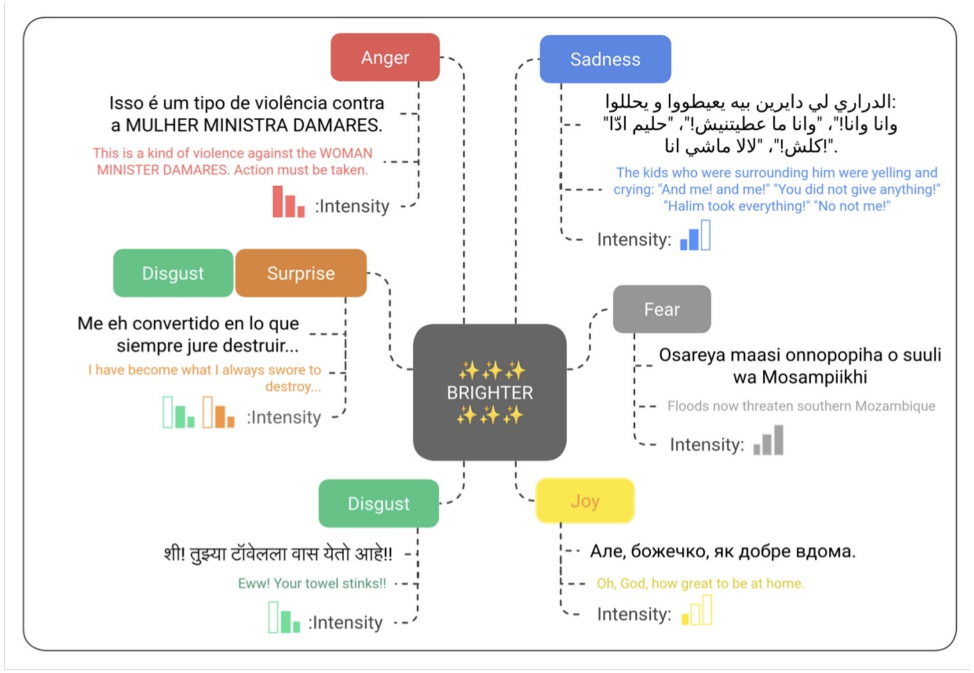
Emotion labels

Anger, Disgust, Sadness,
Joy, Fear, Surprise, Neutral

12

Emotion intensity

0 → no emotion
1 → low intensity
2 → moderate intensity
3 → high intensity



BRIGHTER: BRIdging the Gap in Human-Annotated Textual Emotion Recognition Datasets for 28 Languages
Evaluating the Capabilities of Large Language Models for Multi-label Emotion Understanding

Dataset Quality

Inter-Annotator **Agreement** (IAA) vs Annotation **Reliability**

- Split-Half Class Match Percentage (SHCMP)

  ↳ extends the concept **of Split-Half Reliability** (SHR), traditionally used for continuous scores to discrete categories (i.e., intensity scores).

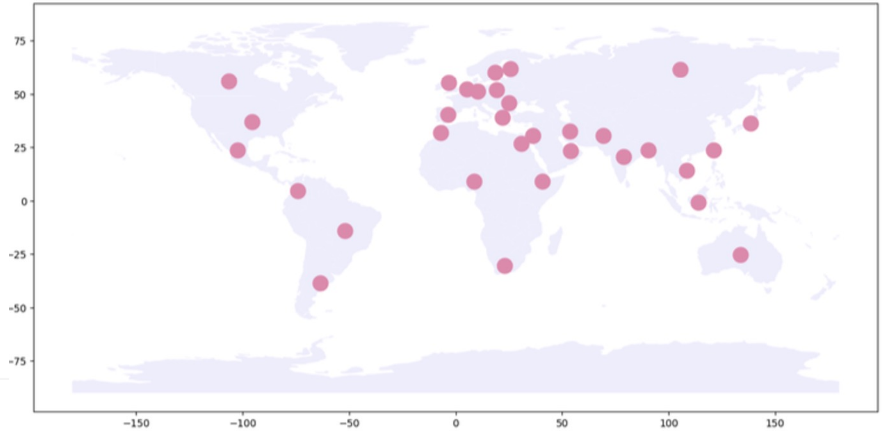
- SHMP scores vary from **60%** to more than **90%**, indicating that our datasets are of high quality.

WorryWords: Norms of Anxiety Association for over 44k English Words (Mohammad, EMNLP 2024)

Tasks Summary

Our task was the most popular
competition on Codabench in 2024

Codabench Newsletter 2024



700+
Registered Participants

362
Submitted system during evaluation phase

93
Submitted system description paper

220
Task A: Multi-label Emotion Detection

96
Task B: Emotion Intensity Detection

46
Task C: Cross-lingual Emotion Detection

Top Systems

	Team	# Languages
Task A Multi-Label Emotion Detection	**Pi** Ping An Life Insurance Company of China	20 out of 28 languages
Task B Emotion Intensity Detection	**Pi** Ping An Life Insurance Company of China	10 out of 11 languages
Task C Cross-lingual Emotion Detection	**DeepWave** Tomorrow Advancing Life (China)	25 out of 32 languages

Best System (Tracks A and B)

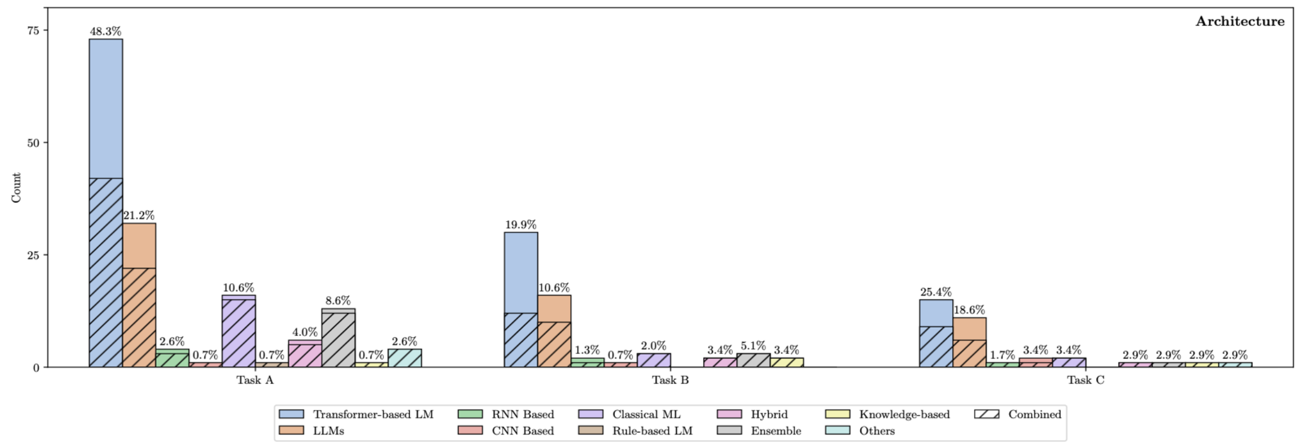
Track A & B

PAI at SemEval-2025 Task 11: A Large Language Model Ensemble Strategy for Text-Based Emotion Detection

Ping An Life Insurance Company of China

Approach

*CSECU-Learners ranked at the top in **Amharic** by fine-tuning language-specific transformers for Amharic with a classification layer and multi-sample dropout.*



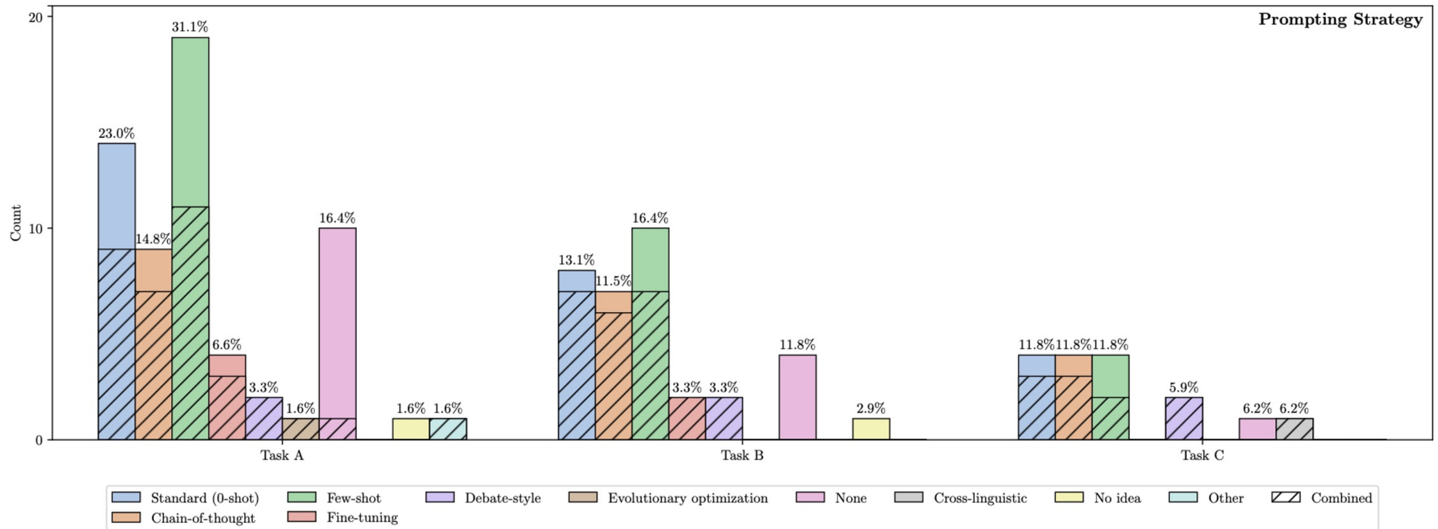
Best System (Track C)

Track C

PAI at SemEval-2025 Task 11: A Large Language Model Ensemble Strategy for Text-Based Emotion Detection

Ping An Life Insurance Company of China

- *Supervised fine-tuning (SFT) Google Gemma 2 large language model*
- *data augmentation, Chain-of-Thought (CoT) prompting, and model ensembling techniques.*



Takeaways: Popular Methods

- Most top-performing teams favored fine-tuning and prompting **LLMs**
- **Full fine-tuning** and **parameter-efficient fine-tuning** were the most commonly used strategies to enhance performance
- For prompting, **few-shot**, **zero-shot**, and **chain-of-thought** prompting were the most frequently used techniques.
- Traditional transformer-based models, particularly **XLM-RoBERTa**, **mBERT**, **DeBERTa**

Takeaways: Best Performing Systems

- **LLMs** achieve strong overall performance; however, their effectiveness is heavily dependent on **prompt engineering** techniques and **wording**.
- Performance varies significantly by language.
    - Better in **high-resource languages** such as English and Russian
    - Dropped on low-resource languages such as **Swahili** and **Emakhuwa**
- Most teams did not incorporate **additional datasets** to enhance performance, as **few-shot** and **zero-shot** approaches proved highly effective.

SemEval 2025 Task 11: Text-Based Emotion Detection

Thanks to all participants

Thanks to the SemEval Chairs

Shamsuddeen Hassan Muhammad*, Nedjma Ousidhoum*, Idris Abdulmumin, Seid Muhie Yimam, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine De Kock, Tadesse Destaw, Belay, Ibrahim Said Ahmad, Nirmal Surange, Daniela Teodorescu, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino Ali, Vladimir Araujo, Abinew Ali Ayele, Oana Ignat, Alexander Panchenko, Yi Zhou, Saif M. Mohammad

https://github.com/emotion-analysis-project/SemEval2025-Task11