Emotion Detection Across Various Languages

Mudit Arora
UC Santa Cruz
muarora@ucsc.edu

Yousuf Golding
UC Santa Cruz
ygolding@ucsc.edu

Abstract

This paper presents our approach to the SemEval-2025 Task 11: Bridging the Gap in Text-Based Emotion Detection. We focus on Track A (Multi-label Emotion Detection) and Track C (Cross-lingual Emotion Detection), implementing and comparing LSTM and BiLSTM architectures. For Track A, we develop a 2-layer LSTM model with trainable embeddings, achieving notable performance variations across different emotions (F1 scores ranging from 0.05 to 0.77) and an overall F1 score of 0.35 on the development set. For Track C, we implement a 4-layer BiLSTM with language-aware attention and FastText embeddings, achieving an F1 score of 0.33 when training on German and testing on Russian data. Our analysis reveals significant challenges in both scenarios: class imbalance heavily impacting single-language emotion detection, and linguistic-cultural differences affecting cross-lingual performance. We explore additional experiments using machine translation for cross-lingual detection, finding that direct emotion detection outperforms translation-based approaches. These findings contribute to understanding the complexities of emotion detection across languages and highlight the importance of addressing both linguistic and cultural aspects in future approaches.

1 Introducing The Problem

Emotion detection in text (and other data types) presents a unique challenge that goes beyond traditional sentiment analysis. Typically, the topic focuses on labeling text as positive or negative.

Our task however is more subtle and complex in comparison, as emotion detection instead requires an understanding of the subtle distinctions that represent different emotional states. A model designed for this task must be able to go through the text, identifying key words that respect specific emotional states, and then recognize how words are being utilized together.

The complexity of this task is further increased with the inclusion of cross-lingual scenarios, where we intend to train the model on text from one language, and then predict on text from another language. This complicates the task considerably more as the way emotions are expressed can heavily differ linguistically and culturally, due to nuances specific to native speakers of the language.

The SemEval task 11 is designed to address these challenges by providing a framework for evaluating emotion detection across a variety of languages by providing datasets composed of varying amounts of samples text (labeled and unlabeled).

Specifically, the task focuses on detecting perceived emotions, essentially how the emotions of the text are perceived rather than what the speaker is actually trying to convey. This is a crucial distinction because perceived emotions can heavily different from intent of the speaker due to multiple reasons including cultural context, individual differences in emotional expression as well as limits of text based communication.

The SemEval task includes three tracks for participants to follow. Our work focuses on Tracks A and C, which are multi-label emotion detection for a singular language and detection across multiple languages respectively.

2 Related Work in The Field

Previous work in emotion detection has explored varying methodologies in approaching the task. While the task scope expands into other data types such as visual and audible data, text based approaches still remain the most popular focus when approaching the problem. There are several approaches worth analyzing regarding what worked and what needs refinement.

(Noktehdan Esfahani and Adda, 2024) compared machine learning models with modern large language models for emotion recognition. The comparison showcased how older architectures suffer from a lack of comprehension regarding context, while larger architectures traded computational cost for better understanding of the emotional content. Their work provided the insight that choosing simpler models requires significant preprocessing to achieve similar contextual understanding.

(Gupta et al., 2023) took a unique approach by combining sentiment and semantic-based features in their SS-BED model. Their work emphasized how emotion detection performance varies significantly based on the distribution of emotional labels in the training data. This proved particularly relevant for our work, as we observed similar patterns where emotions with more training examples showed better performance than those with fewer examples.

(Abas et al., 2022) proposed a hybrid approach combining BERT with CNN and LSTM architectures. Their model processes BERT embeddings through LSTM layers to capture sequential patterns, while utilizing CNN layers for feature extraction. The use of BiLSTM with attention mechanisms and 1D convolutional layers enabled better capture of both temporal dependencies and local features in text. This multi-architecture approach achieved significant accuracy improvements on benchmark datasets.

The importance of feature engineering was further highlighted by (Machov'a et al., 2023). Their work analyzed how varying levels of preprocessing and feature extraction impacted model performance. They found that many deep learning architectures struggled with raw emotion detection, but showed significant improvement with proper engineering. This insight guided our choice in implementing thorough preprocessing.

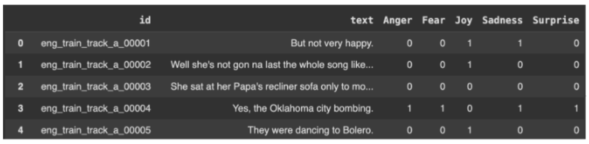
Moving towards multi-lingual applications, (Gao et al., 2022) explored emotion shift detection in conversations using a multi-task learning framework. Their work emphasized how traditional methods struggled with contextual understanding across conversations. Instead, they proposed methods using a shared BERT-based encoder that could work effectively across multiple utterances. This concept influenced our approach to Track C, where we focused on creating a model capable of direct cross-lingual detection.

3 The Dataset

The datasets provided consist of different text data entries for the languages. For the training datasets, this includes both the text and associated emotion labels. For the development datasets the text is all unlabeled.

There will be actual text datasets released in January with associated "true" labels, but currently our only option was to utilize the development datasets for testing.

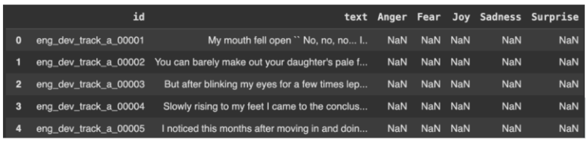
- For Track A, English dataset consisted of only 5 emotions as the labels; Anger, Fear, Joy, Sadness, Surprise.



	id	text	Anger	Fear	Joy	Sadness	Surprise
0	eng_train_track_a_00001	But not very happy.	0	0	1	1	0
1	eng_train_track_a_00002	Well she's not gon na last the whole song like...	0	0	1	0	0
2	eng_train_track_a_00003	She sat at her Papa's recliner sofa only to mo...	0	0	0	0	0
3	eng_train_track_a_00004	Yes, the Oklahoma city bombing.	1	1	0	1	1
4	eng_train_track_a_00005	They were dancing to Bolero.	0	0	1	0	0

Figure 1: Training Dataset (English)

- Vocabulary size of the dataset: 2768

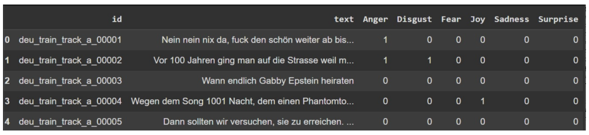


	id	text	Anger	Fear	Joy	Sadness	Surprise
0	eng_dev_track_a_00001	My mouth fell open `` No, no, no... I..	NaN	NaN	NaN	NaN	NaN
1	eng_dev_track_a_00002	You can barely make out your daughter's pale f...	NaN	NaN	NaN	NaN	NaN
2	eng_dev_track_a_00003	But after blinking my eyes for a few times lep...	NaN	NaN	NaN	NaN	NaN
3	eng_dev_track_a_00004	Slowly rising to my feet I came to the conclus...	NaN	NaN	NaN	NaN	NaN
4	eng_dev_track_a_00005	I noticed this months after moving in and doin...	NaN	NaN	NaN	NaN	NaN

Figure 2: Development Dataset (English)

- Vocabulary size of the dataset: 116

- For Track C, Cross-lingual languages have one extra emotion as label; Disgust. For our approach, we worked on using both Russian and German dataset.



	id	text	Anger	Disgust	Fear	Joy	Sadness	Surprise
0	deu_train_track_a_00001	Nein nein nix da, fuck den schön weiter ab bis...	1	0	0	0	0	0
1	deu_train_track_a_00002	Vor 100 Jahren ging man auf die Strasse weil m...	1	1	0	0	0	0
2	deu_train_track_a_00003	Wann endlich Gabby Epstein heiraten	0	0	0	0	0	0
3	deu_train_track_a_00004	Wegen dem Song 1001 Nacht, dem einen Phantomto...	0	0	0	1	0	0
4	deu_train_track_a_00005	Dann sollten wir versuchen, sie zu erreichen. ...	0	0	0	0	0	0

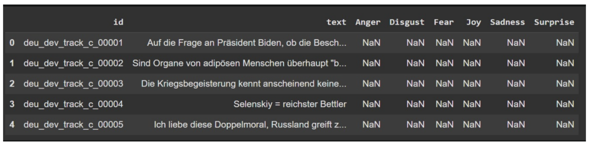
Figure 3: Training Dataset (German)

- Vocabulary size of the dataset: 2605

Note that for the sake of avoiding too many figures at once in the dataset section, we have included the Russian datasets in the appendix.

4 Evaluation Metric

For both track A and track C, the F1-score was chosen as the primary evaluation metric. The F1-score is calculated as the harmonic mean of precision and recall:



	id	text	Anger	Disgust	Fear	Joy	Sadness	Surprise
0	deu_dev_track_c_00001	Auf die Frage an Präsident Biden, ob die Besch...	NaN	NaN	NaN	NaN	NaN	NaN
1	deu_dev_track_c_00002	Sind Organe von adipösen Menschen überhaupt "b...	NaN	NaN	NaN	NaN	NaN	NaN
2	deu_dev_track_c_00003	Die Kriegsbegeisterung kennt anscheinend keine...	NaN	NaN	NaN	NaN	NaN	NaN
3	deu_dev_track_c_00004	Selenskiy = reichster Bettler	NaN	NaN	NaN	NaN	NaN	NaN
4	deu_dev_track_c_00005	Ich liebe diese Doppelmoral, Russland greift z...	NaN	NaN	NaN	NaN	NaN	NaN

Figure 4: Development Dataset (German)

- Vocabulary size of the dataset: 201

$$\text{F1} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{1}$$

This metric is particularly well-suited for emotion detection tasks due to several key characteristics:

- **Balance**: F1-score provides a balanced assessment by combining:
  - Precision: The accuracy of positive predictions (how many predicted emotions are correct)
  - Recall: The completeness of positive predictions (how many actual emotions are captured)
- **Class Imbalance Handling**: Given that emotional labels often have uneven distribution in the dataset, F1-score helps:
  - Prevent bias towards majority classes
  - Ensure fair evaluation for emotions with fewer samples
  - Maintain model accountability for all emotion categories

This metric choice ensures that the model's performance is evaluated comprehensively, considering both its ability to avoid false positives and its capability to detect all relevant emotions, regardless of their frequency in the dataset.

5 Proposed Approach

5.1 Proposed Approach - LSTM (Track A)

This approach focused on Track A (Multi-label Emotion Detection). The architecture incorporates several key components designed for effective text processing and emotion detection, with particular attention to handling varying emotional expressions in text.

5.1.1 Text Preprocessing and Embedding

The input processing pipeline consists of the following steps:

- Input texts are tokenized into words and converted to lowercase
- A vocabulary is constructed from the training data with minimum frequency threshold of 2
- Special tokens <PAD> and <UNK> are used for padding and unknown words
- Words are embedded into 128-dimensional vectors using trainable embeddings

5.1.2 Model Architecture

The network architecture comprises three main components:

Embedding Layer:

$$E : R^V \to R^d \tag{2}$$

where:

- V = Vocabulary size (2,741 tokens)
- d = Embedding dimension (128)
- Includes padding index for efficient batch processing

LSTM Layer: The LSTM layer is defined as:

$$h_t = \text{LSTM}(x_t, h_{t-1}, c_{t-1}) \tag{3}$$

with specifications:

- 2-layer LSTM architecture
- Hidden dimension: 256 units per layer
- Dropout rate: 0.5 for regularization
- Forward sequence processing

Output Layer:

$$\hat{y} = \sigma(W[h_{n-1}, h_n] + b) \tag{4}$$

where:

- $W \in R^{5\times512}$ (concatenated outputs from last two layers)
- σ is the sigmoid activation function
- Output dimension: 5 (one per emotion)
- Output range: [0,1] for each emotion class

5.1.3 Training Configuration

The model is trained with the following parameters:

- Loss Function: Binary Cross-Entropy
- Optimizer: Adam with learning rate $\alpha = 0.001$
- Batch size: 32
- Number of epochs: 5
- Device: CPU/GPU based on availability

5.2 Proposed Approach - BiLSTM (Track C)

This approach focused on Track C (Cross-lingual emotion detection), extending the LSTM architecture to handle the complexities of multiple languages while maintaining emotion detection accuracy.

5.2.1 Text Preprocessing and Embedding

The preprocessing pipeline builds upon the base LSTM approach with additional considerations for cross-lingual processing:

- Text length normalization (140 tokens for source, 80 for target language)
- Language-specific text cleaning and whitespace normalization
- Frequency threshold of 3 for vocabulary construction
- FastText embeddings (300-dimensional) for cross-lingual representation
- Separate preprocessing pipelines for German and Russian texts to preserve linguistic characteristics

5.2.2 Model Architecture

The architecture extends the base LSTM model (defined in equations 1-3) with bidirectional processing and attention mechanisms. Key components include:

Bidirectional Layer:

- 4-layer BiLSTM utilizing forward and backward processing
- Hidden dimension: 512 units per layer
- Dropout rate: 0.5 between layers for regularization

Attention Mechanism:

- Language-aware attention for capturing cross-lingual patterns
- Attention weights computed over concatenated bidirectional outputs
- Masked attention to handle varying sequence lengths

5.2.3 Training Configuration

The model employs specialized training parameters for cross-lingual learning:

- Weighted Focal Loss to handle class imbalance
- AdamW optimizer with weight decay 0.1
- Learning rate: 0.001 with ReduceLROnPlateau scheduling
- Batch size: 32
- Number of epochs: 10
- Gradient clipping with max norm 1.0

6 Results / Analysis

6.1 Results for LSTM Approach (Track A)

The LSTM model achieved the following performance metrics on the training set based on the hyperparameters discussed in 4.1 Proposed Approach - LSTM (Track A):

Emotion	F1 Score
Anger	0.0522
Fear	0.7690
Joy	0.0847
Sadness	0.3569
Surprise	0.5812

Table 1: F1 Scores per emotion category

The table below showcases how the training/validation loss and F1-score changed throughout the process of the model training:

- We were able to achieve an F1 score of 0.35 on the development/validation set by uploading our results on Codebench that is given by the sponsors for this SemEval.

Epoch	Loss
1/5	0.5746
2/5	0.5555
3/5	0.5421
4/5	0.5366
5/5	0.5260

Table 2: Training Loss per Epoch

6.1.1 Class Imbalance Impact

Analysis of the performance metrics shows:

- Superior performance on Fear (F1: 0.7690) correlating with highest class representation
- Poor performance on Anger (F1: 0.0522) and Joy (F1: 0.0847) corresponding to lower representation

6.1.2 Classification Metrics

Emotion	Precision	Recall	F1-score
Anger	0.75	0.03	0.0522
Fear	0.65	0.94	0.7690
Joy	0.88	0.04	0.0847
Sadness	0.59	0.26	0.3569
Surprise	0.68	0.51	0.5812

Table 3: Detailed classification metrics per emotion

6.1.3 Additional Experimentation

We conducted additional experiments to evaluate our model's performance on Track C, testing it both with and without using the (Kossen et al., 2020) M2M100 (machine translator) on the Russian Dataset. These experiments aimed to assess the model's adaptability to different languages and data conditions.

- Achieved F1 score of 0.09 using M2M100.
- Achieved F1 score of 0.05 without using M2M100.

- Our assumption for this low F1 score was that Russian language might have linguistic characteristics that posed unique challenges for emotion detection.

6.2 Results for BiLSTM Approach (Track C)

Listed below is a table showcasing the hyperparameters used for the most optimal results for the BiLSTM approach for cross-lingual emotion detection.

Hyperparameter	Value
Hidden size	512
Number of layers	4
Dropout	0.5
Batch size	32
Epochs	10

Table 4: Optimized Hyperparameters

With these hyperparemeters we were able to get a final result of 33% for the F1-score, this was utilizing German as the training dataset and Russian as the development dataset which the model predicted on.

We attempted a variety of hyperparameters but found that the results were largely the same. The graph below showcases how the training/validation loss and F1-score changed throughout the process of the model training.

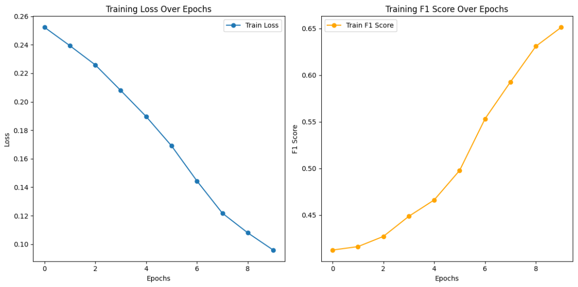


Figure 5: Training Loss and F1-score Change

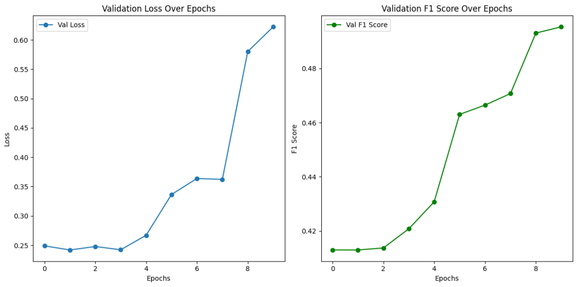


Figure 6: Validation Loss and F1-score Change

6.2.1 Analysis of Results

From analyzing the training / validation graphs, we can note how there are clear signs of overfitting showcased by the validation loss initially decreasing for the first five epochs, and then increasing after.

However, as mentioned, even with different hyperparameters these results were overall fairly consistent. This suggests that there is a more fundamental problem wrong with the approach beyond just overfitting. There are two suspicions for this, the first being that BiLSTM is simply not the right approach for multi-lingual emotion detection. More practical architecture approaches will be discussed in the next section.

Secondly, the data preprocessing approach is potentially not practical for this type of task. In particular many of the related works emphasize the importance of multi-lingual embeddings and that if done incorrectly, will constitute a considerable drop in the effectiveness of the model.

One additional point to also consider regarding this task for cross-lingual detection in general is the languages being utilized. Due to the limited amount of datasets available, we chose to use German and Russian as our training and development datasets.

However, they are considerably different languages both linguistically and culturally. This means that the results when testing these languages was never going to be that substantial in the first place.

7 Future Approach And Concluding Remarks

Regarding future approaches, a critical aspect of our current methodology was the use of LSTM and BiLSTM architectures. This choice stemmed primarily from a misunderstanding about whether transformers were permitted for the project. In retrospect, transformer-based models would be the more effective approach, as supported by numerous related works.

Specifically, transformer models designed for multilingual applications (such as mBart) would be particularly well-suited for this task. We intend to implement this architecture in preparation for the official SemEval testing phase in January. To enhance performance further, we plan to incorporate advanced preprocessing techniques including named entity recognition, sophisticated tokenization, and text normalization. These improvements are especially crucial for track C, where we must address the nuances specific to multiple languages and cultural contexts. Additionally, we can optimize the training process through the implementation of model checkpointing and early stopping.

A key question that emerged during this project was whether to use machine translation for track C or pursue direct emotion detection across languages. While our results showed that machine translation was less effective than initially anticipated, various related works demonstrated successful applications of translation-based approaches. Given this mixed evidence, both methodologies merit consideration, as each offers distinct advantages depending on the specific context and requirements.

8 What We Worked On

- Mudit implemented the LSTM model for Track A. He also worked on the initial exploration of the datasets to understand what methods of data pre-processing would be necessary for the task.
- Yousuf implemented the BiLSTM model for Track C. He also focused on researching related works to better understand the task for single-language and cross-lingual emotion detection.
- The report and presentation was a joint effort. We split the sections / slides based on what we worked on and researched.

References

Ahmed R Abas, Ibrahim Elhenawy, Mahinda Zidan, and Mahmoud Othman. 2022. Bert-cnn: A deep learning model for detecting emotions from text. *Computers, Materials Continua*, 71(2):2944–2961.

Ankush Chatterjee, Umang Gupta, Manoj Kumar Chinnakotla, Radhakrishnan Srikanth, Michel Galley, and Puneet Agrawal. 2019. Understanding emotions in text using deep learning and big data. *Computers in Human Behavior*, 93:309–317.

Qingqing Gao, Biwei Cao, Xin Guan, Tianyun Gu, Xing Bao, Junyan Wu, Bo Liu, and Jiuxin Cao. 2022. Emotion recognition in conversations with emotion shift detection based on multi-task learning. *Knowledge-Based Systems*, 248:108861.

Akriti Gupta, Ankit Kumar, and Vasudeva Varma. 2023. Emotion recognition in conversations using commonsense knowledge graph attention network. *arXiv preprint arXiv:2403.01222*.

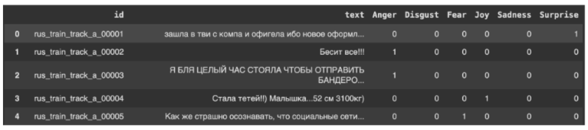
Vinay Kumar Jain, Shishir Kumar, and Steven Lawrence Fernandes. 2017. Extraction of emotions from multilingual text using intelligent text processing and computational linguistics. *Journal of Computational Science*, 21:316–326.

Jannik Kossen, Neil Band, Clare Lyle, Aidan N. Gomez, Tom Rainforth, and Yarin Gal. 2020. Self-attention between datapoints: Going beyond individual input-output pairs in deep learning. *arXiv preprint arXiv:2010.11125*.

Krist'ina Machov'a, Martina Szab'oov'a, J'an Paralič, and J'an Mičko. 2023. Detection of emotion by text analysis using machine learning. *Frontiers in Psychology*, 14.

Seyed Hamed Noktehdan Esfahani and Mehdi Adda. 2024. Classical machine learning and large models for text-based emotion recognition. *Procedia Computer Science*, 241:77–84.

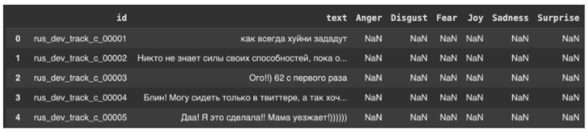
A Appendix



	id	text	Anger	Disgust	Fear	Joy	Sadness	Surprise
0	rus_train_track_a_00001	зашла в тви с компа и офигела ибо новое оформл...	0	0	0	0	0	1
1	rus_train_track_a_00002	Бесит все!!!	1	0	0	0	0	0
2	rus_train_track_a_00003	Я БЛЯ ЦЕЛЫЙ ЧАС СТОЯЛА ЧТОБЫ ОТПРАВИТЬ БАНДЕРО...	1	0	0	0	0	0
3	rus_train_track_a_00004	Стала тетей!!) Малышка...52 см 3100кг)	0	0	0	1	0	0
4	rus_train_track_a_00005	Как же страшно осознавать, что социальные сети...	0	0	1	0	0	0

Figure 7: Training Dataset (Russian)

- Vocabulary size of the dataset: 2679



	id	text	Anger	Disgust	Fear	Joy	Sadness	Surprise
0	rus_dev_track_c_00001	как всегда хуйни зададут	NaN	NaN	NaN	NaN	NaN	NaN
1	rus_dev_track_c_00002	Никто не знает силы своих способностей, пока о...	NaN	NaN	NaN	NaN	NaN	NaN
2	rus_dev_track_c_00003	Ого!!) 62 с первого раза	NaN	NaN	NaN	NaN	NaN	NaN
3	rus_dev_track_c_00004	Блин! Могу сидеть только в твиттере, а так хоч...	NaN	NaN	NaN	NaN	NaN	NaN
4	rus_dev_track_c_00005	Даа! Я это сделала!! Мама уезжает!))))))	NaN	NaN	NaN	NaN	NaN	NaN

Figure 8: Development Dataset (Russian)

- Vocabulary size of the dataset: 401