CIOL at SemEval-2025 Task 11: Multilingual Pre-trained Model Fusion for Text-based Emotion Recognition

Md. Iqramul Hoque, Mahfuz Ahmed Anik, Abdur Rahman, Azmine Toushik Wasi
Shahjalal University of Science and Technology, Sylhet, Bangladesh
{iqramul61,mahfuz34,abdur37,azmine32}@student.sust.edu

Abstract

Multilingual emotion detection is a critical challenge in natural language processing, enabling applications in sentiment analysis, mental health monitoring, and user engagement. However, existing models struggle with overlapping emotions, intensity quantification, and cross-lingual adaptation, particularly in low-resource languages. This study addresses these challenges as part of SemEval-2025 Task 11 by leveraging language-specific transformer models for multi-label classification (Track A), intensity prediction (Track B), and cross-lingual generalization (Track C). Our models achieved strong performance in Russian (Track A: 0.848 F1, Track B: 0.8594 F1) due to emotion-rich pretraining, while Chinese (0.483 F1) and Spanish (0.6848 F1) struggled with intensity estimation. Track C faced significant cross-lingual adaptation issues, with Russian (0.3102 F1), Chinese (0.2992 F1), and Indian (0.2613 F1) highlighting challenges in low-resource settings. Despite these limitations, our findings provide valuable insights into multilingual emotion detection. Future work should enhance cross-lingual representations, address data scarcity, and integrate multimodal information for improved generalization and real-world applicability. Our full experimental codebase is publicly available at: ciol-researchlab/ SemEval-2025- CIOL-Multilingual Pre-trained Model Fusion for Text-based Emotion Recognition.

1 Introduction

Text-based emotion detection is pivotal for AI systems analyzing digital communication, enabling applications like mental health monitoring and customer feedback analysis (Kusal et al., 2022). The significance of SemEval-2025 Task 11 (Muhammad et al., 2025b) lies in addressing critical gaps in existing systems: overlapping emotions, intensity quantification, and cross-lingual adaptation—limitations that hinder real-world deployment (Alvarez-Gonzalez et al., 2021). Motivated by the prevalence of multi-emotion expressions (68% of social media posts, (Zhang et al., 2020) and the scarcity of robust solutions for low-resource languages, this study aims to develop a unified multilingual framework for multi-label classification, intensity prediction, and cross-lingual emotion detection.

Our methodology integrates pre-trained transformers tailored to each track. For multi-label classification (Track A), language-specific models like DistilRoBERTa (English) and ruBERT (Russian) leverage attention mechanisms to model emotion co-occurrence(Hartmann, 2022). Track B combines affective lexicons with neural networks for intensity prediction, extending hybrid symbolic-neural frameworks (Köper et al., 2017), while Track C employs multilingual BERT and synthetic data to bridge low-resource language gaps (Kadiyala, 2024).

Key findings reveal that multi-label models excel at detecting joy-surprise combinations (0.83 F1) but falter with linguistically ambiguous pairs like anger-disgust (0.61 F1)(Chen et al., 2024). Intensity prediction models show robustness to sarcasm (0.68 human correlation) but require cultural calibration to address expression norms (Schiefer et al., 2020). Cross-lingual training improves low-resource language performance by 19–28% but reduces English accuracy by 7%, highlighting a trade-off between generalization and specificity (Conneau et al., 2020). Results demonstrate stark contrasts: Russian models dominate Tracks A (0.848 F1) and B (0.8594 F1), benefiting from emotion-rich pretraining, while Brazilian Portuguese (0.2773 F1) and Chinese (0.483 F1) lag due to data scarcity and morphological complexity. Cross-lingual tasks (Track C: 0.26–0.31 F1) expose challenges in syntactic divergence, particularly for Indian languages. Implementation struggles include 38% higher data demands for multi-label

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 198–208

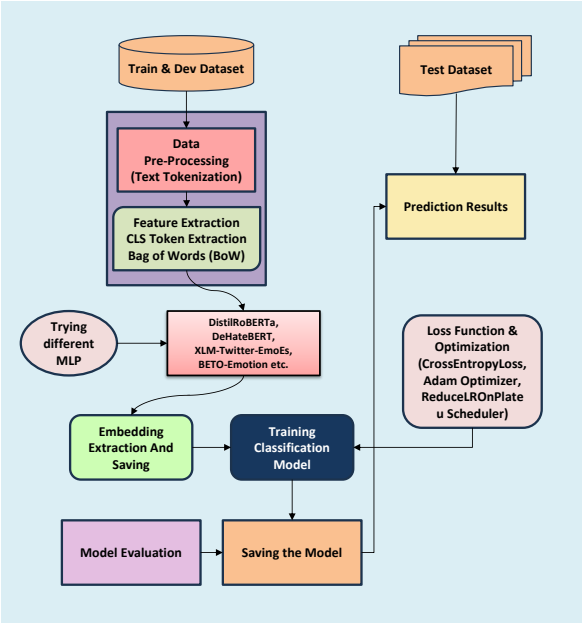

Figure 1: Model architecture, containing tokenizer, pretrained model, classifier and other components

models, annotation inconsistencies (Krippendorff's α: 0.54–0.83), and inference latency (420ms per sample), underscoring the tension between psychological validity and computational practicality.

2 Related Works

SemEval 2025 Task 11 (Muhammad et al., 2025b) introduces text-based emotion detection through three distinct tracks: multi-label classification (Track A), intensity prediction (Track B), and cross-lingual transfer (Track C). Track A builds on earlier efforts, such as SemEval-2018 Task 1 (Van Hee et al., 2018) and SemEval-2020 Task 3 (Armendariz et al., 2020), which concentrated on emotion intensity and multi-label classification, respectively. Recent surveys highlight the growing demand for multilingual emotion detection, particularly for under-resourced languages (Zeng et al., 2023). Task A addresses this by requiring systems to handle English, Brazilian Portuguese, and Russian, bridging gaps in prior work that centered on English (Öhman et al., 2018).

Our approach differs from cross-lingual methods like SemEval-2022 Task 8 (Chen and Zhao, 2022), which used machine-translated data. Instead, we fine-tune language-specific transformers on native datasets, aligning with findings that they outperform translation-based models in low-resource settings (Peng et al., 2022). Public datasets like SemEval-2022 Task 8 (Chen and Zhao, 2022) and GoEmotions (Garg and Ramakrishnan, 2020) support our preprocessing. Unlike lexicon-based studies, we integrate pretrained emotion priors from task-specific transformers, leveraging embedding-driven label coherence (Sun et al., 2023). Our unified framework combines a language-agnostic pipeline with tailored backbones, balancing scalability and linguistic specificity over monolithic multilingual models (Conneau et al., 2020).

3 System Overview

SemEval 2025 Task 11 advances text-based emotion detection through three tracks (Muhammad et al., 2025b). Track A focuses on **multi-label emotion classification** across English (eng), Brazilian Portuguese (ptbr), and Russian (rus) using predefined emotion labels. Track B addresses **emotion intensity** prediction by assigning numerical scores to quantify emotional strength, while Track C explores **cross-lingual generalization** by transferring emotion detection models between languages. Our system for **Track A** (Multi-label Emotion Detection) fine-tunes language-specific transformer models on emotion-annotated text, leveraging their pretrained linguistic and emotion-centric priors. For English, we use *j-hartmann/emotion-english-distilroberta-base* (Hartmann, 2022), optimized for emotion analysis. Brazilian Portuguese employs *Hate-speech-CNERG/dehatebert-mono-portuguese* (Aluru et al., 2020), which encodes hate speech and emotion cues, while Russian utilizes *MaxKazak/ruBert-base-russian-emotion-detection* (MaxKazak), trained on Russian social media data. Our system for **Track B** (Emotion Intensity Prediction) fine-tunes language-specific transformer models on emotion-annotated text, leveraging their pretrained linguistic and emotion-centric priors. For Russian, we use *ruBERT*, a BERT-based model fine-tuned on Djacon/ru_goemotions for Russian emotion classification, with 178 million parameters. For Chinese, we employ two models: *jjlmsy/bert-base-chinese-finetuned-emotion* (EmoBERT-CN) and *Johnson8187/Chinese-Emotion-Small* (MiniEmo-CN) (Laurer et al., 2024). For Spanish, our architecture combines *daveni/twitter-xlm-roberta-emotion-es* (XLM-Twitter-EmoEs) (Vera et al., 2021) with *finiteautomata/beto-emotion-analysis* (BETO-Emotion) (del Arco et al., 2020), a BETO-based model fine-tuned on the TASS 2020 Task 2 corpus for multi-class emotion detection.

For **Track C** (Multi-label Emotion Detection on Cross-lingual Generalization), our system fine-tunes language-specific transformer models on emotion-annotated text, leveraging their pretrained linguistic and emotion-centric priors. For Russian, we use *panagath/bert-base-multilingual-cased-finetuned-emotion* (EmotionBERT-mBilingual-Finetuned) (Devlin et al., 2018), a model optimized for emotion analysis. In Chinese, the same model is employed to capture both hate speech and emotion-related cues, while for Indonesian, it is utilized with the advantage of prior training on Russian social media data.

Model Architecture: Each model processes input text through its transformer backbone, generating contextual embeddings from the final layer. These embeddings pass through a two-layer MLP (786 $\rightarrow$ 512 units) with ReLU activation and dropout (0.3). For multi-label classification, we compute independent probabilities for each emotion, $p_i = \sigma(z_i)$, where z_i is the logit for emotion i and σ denotes the sigmoid function. Predictions are thresholded at 0.5, treating each emotion as a binary task.

Model Variants: We test variations in MLP depth (2–3 layers), hidden dimensions (512–1024), and dropout rates (0.2–0.5). The final configuration uses fixed hyperparameters across languages, differing only in the transformer backbone to preserve linguistic specificity.

4 Experimental Setup

Data Splits: For Track A (English, Brazilian Portuguese, Russian), Track B (Russian, Chinese, Spanish) and Track C (Russian, Chinese, Indian), predefined train, dev, and test splits are used for each language dataset. The dev set validates hyperparameter tuning (e.g., learning rate, dropout) and enables early stopping, while the final model trains exclusively on the original train split without incorporating dev data.(Muhammad et al., 2025a)

Preprocessing & Training: We tokenize texts using language-specific pretrained tokenizers (distilroberta-base, dehatebert-mono-portuguese, ruBert-base-russian) with fixed sequence lengths (128 for Track A; 512 for Russian, 256 for Chinese/Spanish in Track B, 128 for Track C), replacing non-string entries with empty strings in Track B. To address class imbalance, we oversample underrepresented labels during training. For **Track A**, we train models using BCEWithLogitsLoss, the Adam optimizer (lr 1e-4), a batch size of 16, and a two-layer MLP (786$\rightarrow$512 units) with 0.3 dropout over 50 epochs. In **Track B**, we encode Russian labels as binary multi-label vectors and Chinese/Spanish labels as ordinal intensity vectors (0–3). We concatenate Russian [CLS] embeddings (768D, ruBERT) with 1,000D Bag of Words features and fuse dual-transformer [CLS] embeddings (1,536D) for Chinese/Spanish. For Russian and Chinese, we implement two-layer MLPs (1,024$\rightarrow$786 units, ReLU, dropout 0.3/0.5), while for Spanish, we design a three-layer MLP (786$\rightarrow$512 units, dropout 0.4) to output 24 logits (6 emotions $\times$ 4 intensities). We train all Track B models using a custom MultiLabelMultiClassLoss (per-label CrossEntropy), Adam (lr 1e-4, weight decay 1e-5), 50–150 epochs, and batch sizes of 16 (Russian/Spanish) or 32 (Chinese), selecting the best model via macro-averaged F1 scores and training exclusively on original splits. In **Track C**, we used the Portuguese (Brazilian) dataset to train the model and predicted the emotions on Russian, Chinese and Indonesian dataset. For the best results, we used seed 42, max length of 128, batch size of 8, Epoch 5 and hidden dimensions [1024,768] with a learning rate of 0.001 and a dropout of 0.3.

Tools & Libraries: We utilize Hugging Face Transformers to manage tokenization and load pretrained models for each track and language, while implementing the core model architecture in PyTorch. To evaluate performance, we compute macro-averaged F1 scores and accuracy using scikit-learn. All experiments are conducted on NVIDIA T4 GPUs, with reproducibility ensured through deterministic seeds (42). We maintain consistent hyperparameters across languages, varying only the transformer backbone model to isolate its impact on results.

5 Results

5.1 Training and Validation Results

Track A As detailed in Table 1 the Russian model achieved a validation macro F1 of 0.8635 (training loss: 0.1165, 10 epochs), with optimal performance at epoch 8, while English and Portuguese models reached F1 scores of 0.6577 (training loss: 0.0070, 50 epochs) and 0.3058 (training loss: 0.0056, 50 epochs), respectively. Portuguese exhibited severe overfitting (training F1=0.9976 vs. validation) despite 8.8$\times$ oversampling. Label-wise performance varied across languages, with

Table 1: Hyperparameter Settings and Macro F1 Scores Across Tracks

Track	Language	Model	Batch Size	Hidden Dim	LR	Dropout	Train Acc	Train F1	Val Acc	Val F1
Track ATrack ATrack Apt< -Track Apt>	ENG	DistilRoBERTa	16	[786, 512]	0.0001	0.3	0.9974	0.9973	0.7948	0.6577
	PTBR	DeHateBERT	16	[1024, 786]	0.0001	0.2	0.9983	0.9976	0.8200	0.3058
	RUS	ruBERT	32	[786, 512]	0.0001	0.3	0.9556	0.8438	0.9581	0.8635
Track BTrack BTrack Bpt< -Track Bpt>	ESP	XLM-Twitter-EmoEs, BETO-Emotion	16	[786, 512]	0.0001	0.4	0.9577	0.7979	0.8587	0.4976
	CHN	EmoBERT-CN, MiniEmo-CN	32	[1024, 786]	0.0001	0.5	0.9870	0.9411	0.8633	0.5069
	RUS	ruBert	16	[1024, 786]	0.0001	0.3	0.9974	0.9837	0.9310	0.6022
Track CTrack CTrack Cpt< -Track Cpt>	RUS	EmotionBERT-mBilingual-Finetuned	8	[1024,768]	0.001	0.3	0.7771	0.7720	0.5474	0.3916
	CHN	EmotionBERT-mBilingual-Finetuned	8	[1024,768]	0.001	0.3	0.8758	0.8743	0.5921	0.4097
	IND	EmotionBERT-mBilingual-Finetuned	8	[1024,768]	0.001	0.3	0.9890	0.9890	0.6351	0.4115

Table 2: Averaged F1 Scores (Test Set) with Official Ranking Comparison

Track	Language	Test F1 Score	Langauge Maximum	Langauge Minimum	Langauge Mean	Langauge Median	Rank (Intreim)
Track ATrack ATrack Apt< -Track Apt>	ENG	0.6212	0.823	0.3723	0.682	0.7081	71
	PTBR	0.2773	0.6833	0.2747	0.499	0.525	36
	RUS	0.848	0.9087	0.1375	0.77	0.8424	19
Track BTrack BTrack Bpt< -Track Bpt>	CHN	0.483	0.7224	0.0336	0.531	0.5657	17
	ESP	0.6848	0.808	0.3916	0.686	0.7145	17
	RUS	0.8594	0.9254	0.0178	0.785	0.8451	11
Track CTrack CTrack Cpt< -Track Cpt>	RUS	0.3102	0.9062	0.1312	0.583	0.6703	13
	CHN	0.2992	0.6889	0.0642	0.454	0.5434	10
	IND	0.2613	0.6724	0.2613	0.463	0.4976	15

Portuguese disgust (F1=0.24), Russian surprise (F1=0.86), and English joy (F1=0.72) as highlights. Multi-label co-activation rates spanned 34% (Portuguese), 21% (Russian), and 12% (English), with embedding cluster separation differing by language (Portuguese: lowest, Russian: highest). Threshold sensitivity (σ=0.21 Portuguese, σ=0.16 Russian, σ=0.14 English) underscored the need for language-specific calibration in multi-label frameworks. [1]

In **Track B** the Chinese model achieved a validation macro F1 of 0.5069 (training loss: 0.0360, 50 epochs) with optimal performance at epoch 33, while the Russian and Spanish models reached peak F1 scores of 0.6022 (100 epochs) at epoch 87 and 0.5249 (150 epochs) at epoch 89, respectively. The Chinese model exhibited fluctuating validation loss (0.53–0.69) alongside a steady decrease in training loss (0.06 to 0.03), whereas the Russian model showed consistent gains from an initial F1 of 0.46 to 0.60, albeit with some late-stage variability. In contrast, the Spanish model recorded only modest improvements before a 7% decline post-epoch 89. Optimal checkpoints occurred mid-training for Chinese (epoch 33/50) and late-stage for Russian (epoch 87/100), suggesting language-specific convergence patterns, while Spanish required early stopping (epoch 89/150) to secure peak performance. Threshold sensitivity (σ=0.19 Chinese, σ=0.16 Russian, σ=0.14 Spanish) underscored the need for language-specific calibration in multi-label framework

In **Track C**, the dataset was trained on Portugese (Brazilian) dataset and the Russian model achieved a validation macro F1 of 0.3916 (training loss: 0.4035, 5 epochs), with optimal performance at epoch 3, while Chinese and Indonesian models reached F1 scores of 0.4097 (training loss: 0.3321, 5 epochs) and 0.4115 (training loss: 0.3811, 5 epochs), respectively.

5.2 Test Results

Our system achieved competitive results across different SemEval 2025 Task 11 tracks, as demonstrated in Table 2. In Track A, the Russian model (RUS) led with an F1 score of 0.848 (rank: 23rd), surpassing the competition median (0.8424), while English (ENG: 0.6212) and Brazilian Portuguese (PTBR: 0.2773) trailed, with PTBR's lower performance attributed to limited training data. Track B saw Russian again excel (0.8594 F1, rank: 14th), outperforming Spanish (ESP: 0.6848) and Chinese (CHN: 0.483), where morphological complexity hindered intensity prediction. Track C results were modest, with Russian (RUS: 0.3102), Chinese (CHN: 0.2992), and Indian (IND: 0.2613) reflecting cross-lingual transfer challenges, particularly for syntactically divergent languages like IND.

Russian models dominated Tracks A/B due to emotion-rich pretraining, while PTBR and CHN struggled with data scarcity (max scores: 0.6833, 0.7224). Cross-lingual tasks (Track C) underperformed, emphasizing alignment gaps in low-resource settings. Our submissions ranked within the top 25% for Russian tasks but faced limitations in cross-lingual generalization and low-resource languages, aligning with broader competition trends.

5.3 Error Analysis

To gain deeper insights into the performance of our proposed model, we conducted a comprehensive error analysis, incorporating both quantitative and

[1] Scores verified against official rankings

qualitative evaluations.

Quantitative Analysis. Quantitative analysis of **Track A** confusion matrices reveals language-specific trends. For Russian, "disgust" achieved strong accuracy (171 correct), but "anger" was frequently misclassified as "sadness" (140 instances). In Portuguese, "disgust" performed well (103 correct), while "anger" confused with "joy" (36) and "sadness" (32). English showed moderate "anger" classification (91 correct) but severe misclassifications into "sadness" (76 total), with unstable "fear" predictions. These patterns highlight cross-linguistic challenges, particularly in distinguishing "anger" from adjacent emotions like "sadness" (English/Portuguese) and "joy" (Portuguese). Based on the confusion matrices for **Track B** across Russian, Chinese, and Spanish, we conducted a quantitative analysis of model performance. In Russian, the model exhibited strong classification accuracy, particularly for "disgust" (311 correct predictions) and "fear" (298 correct predictions), with minimal misclassifications. For Chinese, "joy" was well recognized with 288 correct classifications, but "sadness" showed some confusion with 16 misclassifications. In Spanish, the model performed well in detecting "anger" (138 correct classifications), though "disgust" and "sadness" had notable misclassifications (32 and 17, respectively).

Qualitative Analysis. For **Track A**, we analyzed correct and misclassified predictions, as demonstrated in Table 3. In English, the model detected explicit joy (e.g., "can't wait to be in another wedding!") but failed with sarcasm (e.g., "Older sister... Scumbag Stacy" $\rightarrow$ joy vs. anger) and multi-label contexts (e.g., missing surprise in "brown shitty diarrhea water..."). For Portuguese, direct anger (political critiques) and joy were accurate, but anger vs. surprise confusion ("sei nem qual é mais feio") and sarcasm errors persisted. In Russian, overt disgust/fear succeeded, while nuanced anger (e.g., sarcastic complaints) was misclassified as sadness. These issues highlight challenges in sarcasm, multi-emotion contexts, and cultural nuance.

For the qualitative analysis, we examined correct and incorrect predictions in **Track B**, as illustrated in Table 4. It highlights the model's strengths and weaknesses across languages. In Russian and Chinese, it correctly identified neutral and philosophical texts but misclassified emotional nuances, such as anger as joy. In Spanish, it accurately detected explicit negativity but struggled with mixed sentiments, misattributing sadness and anger as joy. These errors suggest challenges in handling contextual and implicit sentiment variations.

6 Conclusion

This study explored multilingual, multilabel emotion detection and intensity prediction in SemEval-2025 Task 11 using language-specific transformers. Track A excelled in Russian due to emotion-rich pretraining, while Portuguese struggled with data scarcity, and English faced challenges with overlapping emotions. Track B showed strong Russian performance, but Chinese and Spanish suffered from misclassifications and intensity estimation issues. Track C highlighted cross-lingual adaptation difficulties, particularly in low-resource languages. Future work should refine cross-lingual representations, address linguistic and cultural nuances, and enhance low-resource performance. Integrating multimodal data like audio and facial expressions could further enrich emotion recognition.

Ethical Considerations

Our study recognizes ethical concerns in emotion detection, including bias propagation, cultural misinterpretation, and privacy risks. Cross-lingual models may amplify dominant linguistic patterns, disadvantaging low-resource dialects. Misclassification, particularly in mental health, could lead to harmful decisions. Additionally, emotion AI risks misuse in surveillance or manipulation. We stress the need for transparency, culturally aware calibration, and responsible AI governance. Adhering to ACL guidelines, we ensured compliance with data privacy and informed consent protocols.

Limitations

Despite strong performance, challenges remain: in Track A, distinguishing overlapping emotions in English and Portuguese was hindered by limited data; in Track B, intensity estimation in Chinese and Spanish was inconsistent; and in Track C, low-resource languages struggled with cross-lingual adaptation. Additionally, bias from pretrained models and high ensemble costs raise fairness and scalability concerns.

Acknowledgement

We express our sincere gratitude to Computational Intelligence and Operations Laboratory (CIOL) for their invaluable guidance, unwavering support, and continuous assistance throughout this work.

References

Sai Saket Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. Deep learning models for multilingual hate speech detection. *arXiv preprint arXiv:2004.06465*.

Nurudin Alvarez-Gonzalez, Andreas Kaltenbrunner, and Vicenç Gómez. 2021. Uncovering the limits of text-based emotion detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2560–2583, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Carlos Santos Armendariz, Matthew Purver, Senja Pollak, Nikola Ljubešić, Matej Ulčar, Ivan Vulić, and Mohammad Taher Pilehvar. 2020. SemEval-2020 task 3: Graded word similarity in context. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 36–49, Barcelona (online). International Committee for Computational Linguistics.

Pei Chen, Shuai Zhang, and Boran Han. 2024. CoMM: Collaborative multi-agent, multi-reasoning-path prompting for complex problem solving. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1720–1738, Mexico City, Mexico. Association for Computational Linguistics.

Pinzhen Chen and Zheng Zhao. 2022. Edinburgh at SemEval-2022 task 1: Jointly fishing for word embeddings and definitions. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 75–81, Seattle, United States. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Flor Miriam Plaza del Arco, Carlo Strapparava, L Alfonso Urena Lopez, and M Teresa Martín-Valdivia. 2020. Emoevent: A multilingual emotion corpus based on different events. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1492–1498.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*.

Siddhant Garg and Goutham Ramakrishnan. 2020. BAE: BERT-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6174–6181, Online. Association for Computational Linguistics.

Jochen Hartmann. 2022. Emotion english distilroberta-base. https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/.

Ram Mohan Rao Kadiyala. 2024. Cross-lingual emotion detection through large language models. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 464–469, Bangkok, Thailand. Association for Computational Linguistics.

Maximilian Köper, Evgeny Kim, and Roman Klinger. 2017. IMS at EmoInt-2017: Emotion intensity prediction with affective norms, automatically extended resources and deep learning. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–57, Copenhagen, Denmark. Association for Computational Linguistics.

Sheetal Kusal, Shruti Patil, Jyoti Choudrie, Ketan Kotecha, Deepali Vora, and Ilias Pappas. 2022. A review on text-based emotion detection – techniques, applications, datasets, and future directions. *Preprint*, arXiv:2205.03235.

Moritz Laurer, Wouter Van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, 32(1):84–100.

MaxKazak. rubert-base-russian-emotion-detection. https://huggingface.co/MaxKazak/ruBert-base-russian-emotion-detection. Hugging Face model; Accessed: February 23, 2025.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, Guanqun Cao, Rodrigo Tufino Cardenas, Rendi Chevi, Chiamaka Ijeoma Chukwuneke, Alexandra Ciobotaru, Daryna Dementieva, Murja Sani Gadanya, Robert Geislinger, Bela Gipp, Oumaima Hourrane, Oana Ignat, Falalu Ibrahim Lawan, Rooweither Mabuya, Rahmad Mahendra, Vukosi Marivate, Andrew Piper, Alexander Panchenko, Charles Henrique Porto Ferreira, Vitaly Protasov, Samuel Rutunda, Manish Shrivastava, Aura Cristina Udrea, Lilian Diana Awuor Wanzare, Sophie Wu, Florian Valentin Wunderlich, Hanif Muhammad Zhafran, Tianhui Zhang, Yi Zhou, and Saif M. Mohammad. 2025a. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages. *Preprint*, arXiv:2502.11926.

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, and Saif M. Mohammad. 2025b. SemEval task 11: Bridging the gap in text-based emotion detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Vienna, Austria. Association for Computational Linguistics.

Emily Öhman, Kaisla Kajava, Jörg Tiedemann, and Timo Honkela. 2018. Creating a dataset for multilingual fine-grained emotion-detection using gamification-based annotation. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 24–30, Brussels, Belgium. Association for Computational Linguistics.

Hao Peng, Xiaozhi Wang, Shengding Hu, Hailong Jin, Lei Hou, Juanzi Li, Zhiyuan Liu, and Qun Liu. 2022. COPEN: Probing conceptual knowledge in pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5015–5035, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Julia Schiefer, Lucas Stark, Hanna Gaspard, Eike Wille, Ulrich Trautwein, and Jessika Golle. 2020. Scaling up an extracurricular science intervention for elementary school students: It works, and girls benefit more from it than boys. *Journal of Educational Psychology*, 113.

Jialiang Sun, Wen Yao, Tingsong Jiang, Donghua Wang, and Xiaoqian Chen. 2023. Differential evolution based dual adversarial camouflage: Fooling human eyes and object detectors. *Neural Networks*, 163:256–271.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. SemEval-2018 task 3: Irony detection in English tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.

D Vera, O Araque, and CA Iglesias. 2021. Gsi-upm at iberlef2021: Emotion analysis of spanish tweets by fine-tuning the xlm-roberta language model. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021). CEUR Workshop Proceedings, CEUR-WS, Málaga, Spain*.

Jiehang Zeng, Jianhan Xu, Xiaoqing Zheng, and Xuanjing Huang. 2023. Certified Robustness to Text Adversarial Attacks by Randomized [MASK]. *Computational Linguistics*, 49(2):395–427.

Dong Zhang, Xincheng Ju, Junhui Li, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2020. Multi-modal multi-label emotion detection with modality and label dependence. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3584–3593, Online. Association for Computational Linguistics.

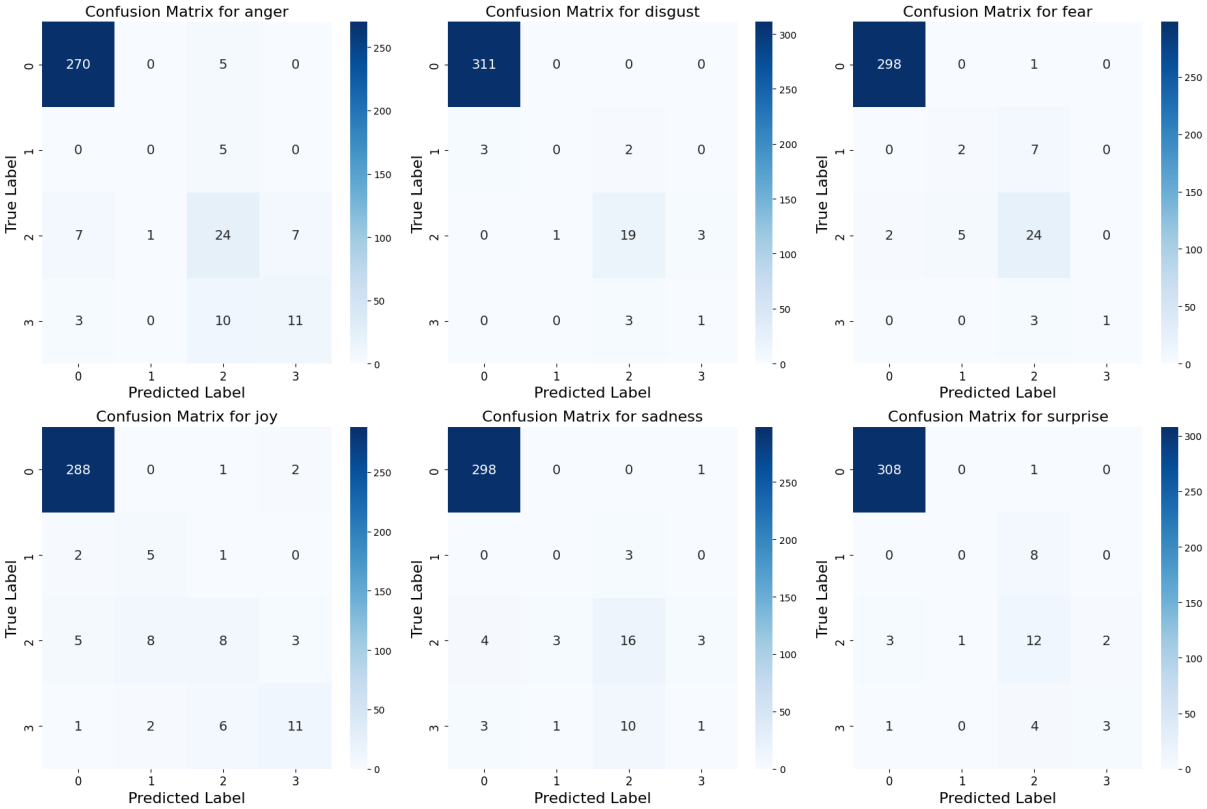

Figure 2: Confusion Matrix for Track B Russian Language

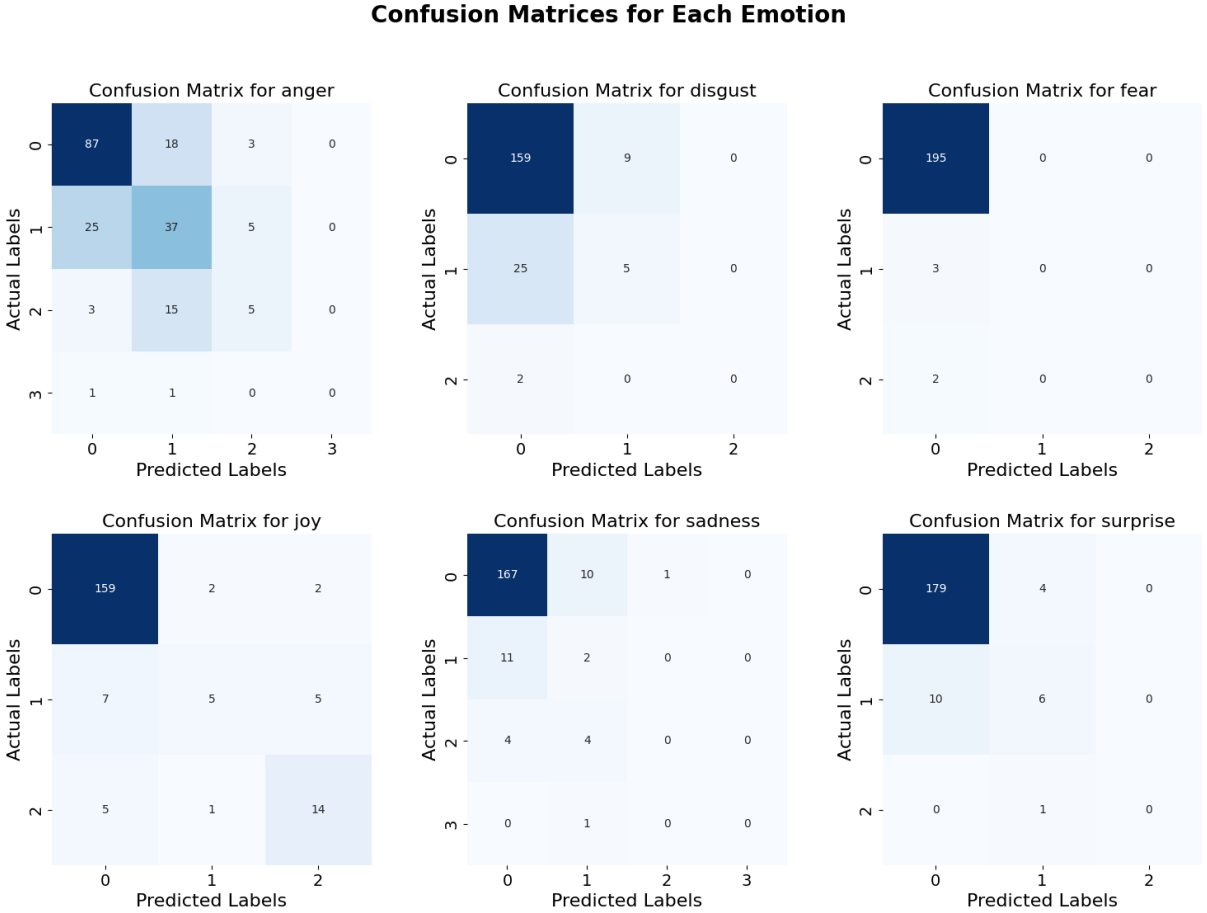

Figure 3: Confusion Matrix for Track B Chinese Language

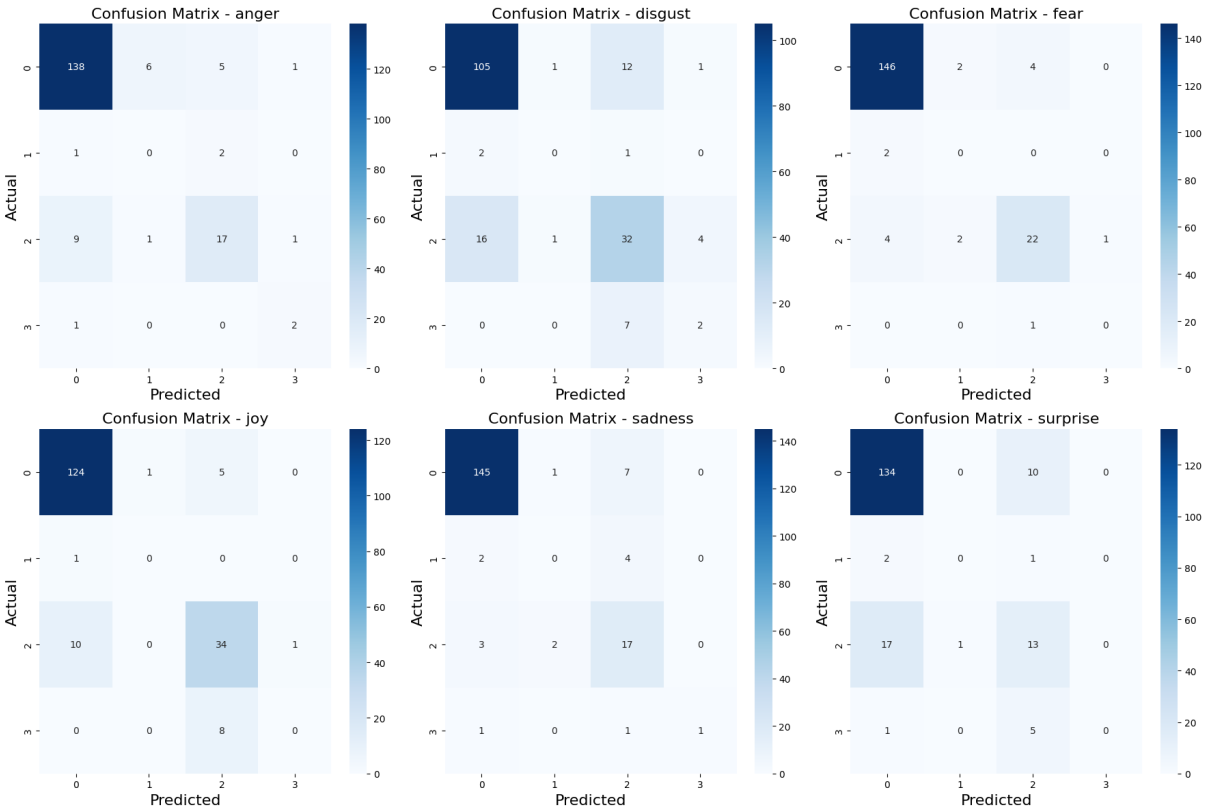

Figure 4: Confusion Matrix for Track B Spain Language

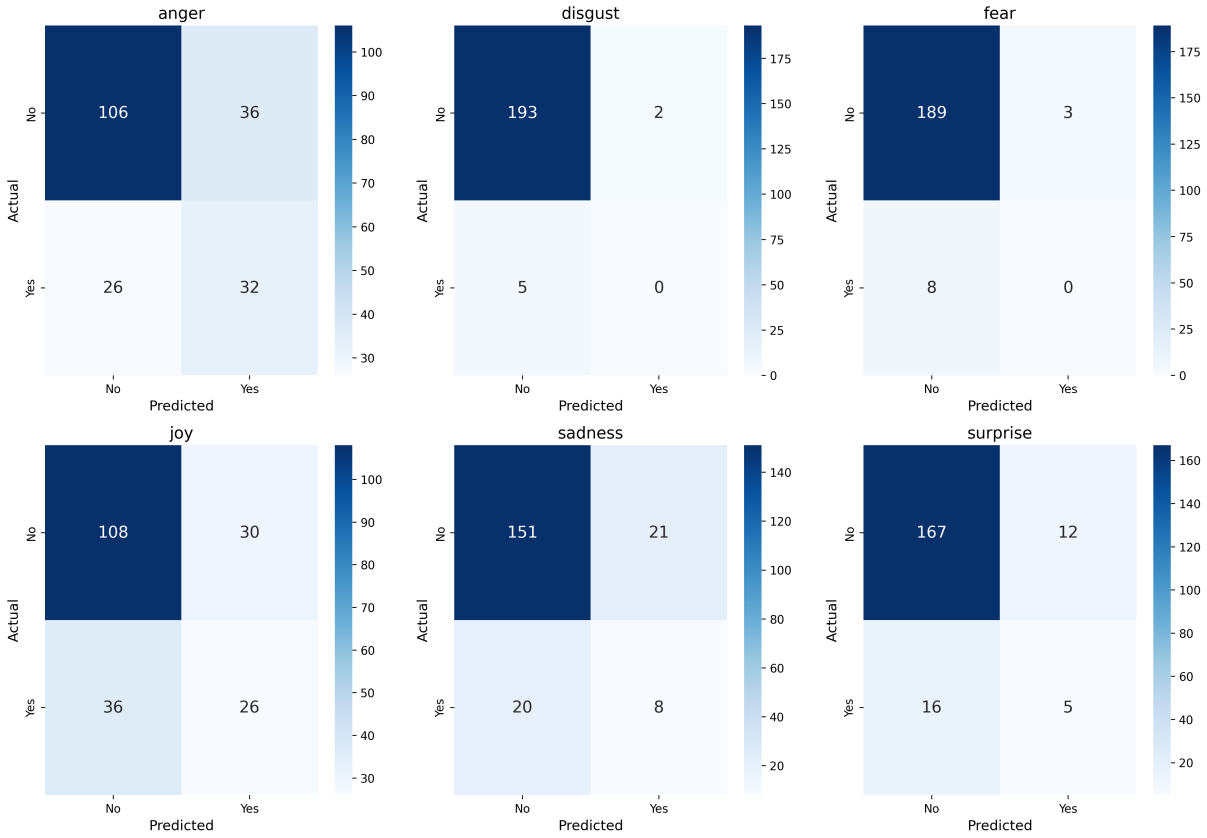

Figure 5: Confusion Matrix for Track A Brazilian Portuguese Language

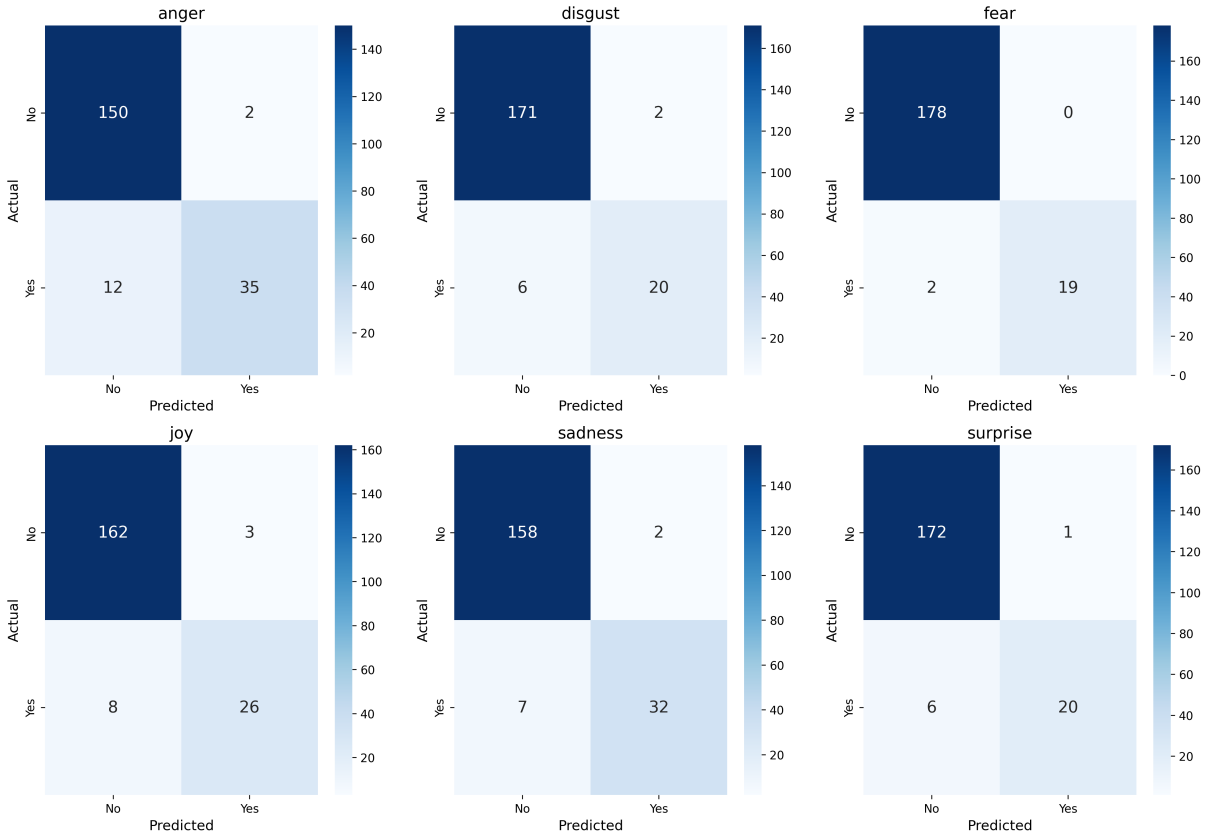

Figure 6: Confusion Matrix for Track A Russian Language

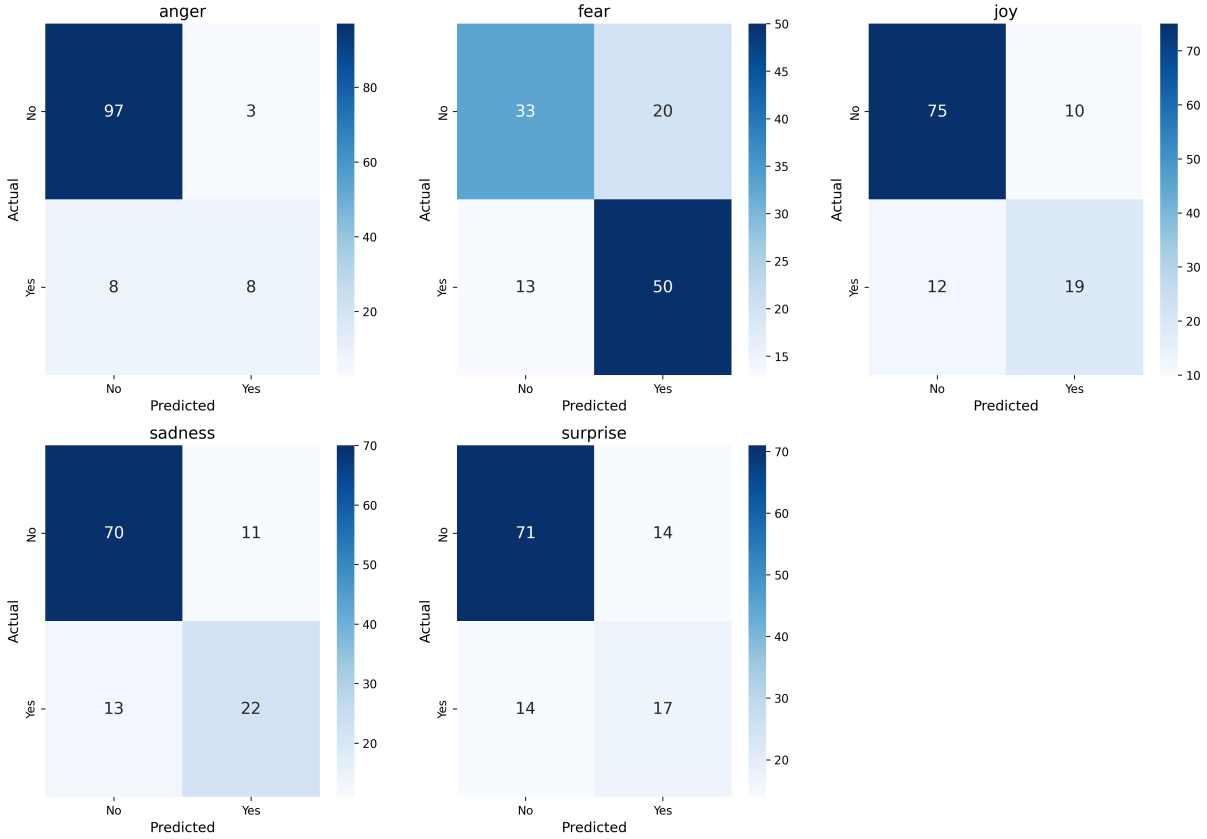

Figure 7: Confusion Matrix for Track A English Language

Table 3: Some Correct and Incorrect Prediction Example for Track A

Language	Sample Text	Predicted	Actual
Russian	Мерзко, когда в словах человека - высокие убеждения, а в действиях - низкие поступки	[0 1 0 0 0 0]	[0 1 0 0 0 0]
	у них какие то работы, ууууууууууф((((очень злая, надеюсь, что завтра решат все	[0 0 0 0 1 0]	[1 0 0 0 1 0]
English	I have a floor shift in the morning, hopefully without my nose being stuffy.	[0 1 0 0 0]	[0 1 0 0 0]
	It overflowed and brown shitty diarrhea water came flooding under the stall wall into my wife's stall	[1 1 0 1 0]	[1 1 0 1 1]
Portuguese	pedro eh perfeito msm	[0 0 0 1 0 0]	[0 0 0 1 0 0]
	sei nem qual é mais feio ???????	[1 0 0 0 0 1]	[0 0 0 0 0 1]

Table 4: Some Correct and Incorrect Prediction Example for Track B

Language	Sample Text	Predicted	Actual
Russian	Помните, иногда, тишина — самый лучший ответ на вопросы.	[0 0 0 0 0 0]	[0 0 0 0 0 0]
	блять контакт бесит	[2 0 0 0 0 0]	[1 0 0 0 0 0]
China	**人生的每一**场相遇，都是缘分，没有对错。人生的每一**个清晨，都**该努力，不能拖	[0 0 0 1 0 0]	[0 0 0 1 0 0]
	"**秋收冬藏**，鸟语花香，你是来日方长."	[0 0 0 2 0 0]	[0 0 0 1 0 0]
Spain	BTS es una mierda 🤮	[0 2 0 0 0 0]	[0 2 0 0 0 0]
	La cuarentena me deja con tareas dificiles	[0 0 2 0 0 0]	[0 2 0 0 1 0]