# IEGPS-CSIC at SemEval-2025 Task 11: BERT-based approach for Multi-label Emotion Detection in English and Russian texts

**Albina Sarymsakova[1], Patricia Martín-Rodilla[1]**
[1]Instituto de Estudios Gallegos Padre Sarmiento (IEGPS-CSIC), Santiago de Compostela, Spain
{albina.s, p.m.rodilla}@iegps.csic.es

## Abstract

This paper presents an original approach for SemEval 2025 Task 11. Our study investigates various strategies to improve Text-Based Multi-label Emotion Detection task. Through experimental endeavors, we explore the benefits of contextualized vector representations by comparing multiple BERT models, including those specifically trained for emotion recognition. Additionally, we examine the impact of hyperparameters adjustments on model performance. For Subtask A, our approach achieved F1 scores of 0.71 on the English dataset and 0.84 on the Russian dataset. Our findings underscore that (1) monolingual BERT models demonstrate superior performance for English, whereas multilingual BERT models perform better for Russian; (2) pretrained emotion detection models prove less effective for this specific task compared to models with reduced vocabulary and embeddings focused on specific languages; (3) exclusive use of BERT-based models, without incorporating additional methods or optimization techniques, demonstrates promising results for multilabel emotion detection.

## 1 Introduction

The emotion mining framework (Liu, 2020) represents a recently established specialized task that allows for evaluation and identification of emotional tone conveyed in textual data. Traditionally, sentiment analysis classified textual content as positive, negative, or neutral (Rosenthal et al., 2017). However, emotion mining techniques enable the assignment of a broader range of emotions to texts (Troiano et al., 2023; Greschner and Klinger, 2024), such as those defined by the Ekman (Ekman and Friesen, 1978; Ekman, 1992), which identifies six basic emotions (sadness, joy, disgust, surprise, anger, and fear) and the absence of emotion. With rise of Large Language Models (LLM), this emotion classification has been also widely applied in domains like marketing (Wemmer et al., 2024), health (Yang et al., 2023), and media (Zhang et al., 2023), helping researchers and organizations understand emotional aspects, make informed decisions, and improve services.

To date, emotion analysis has shown promising results, especially in English, due to abundant resources (Maks and Vossen, 2011; Valitutti et al., 2004). SemEval 2025 Task 11 (Muhammad et al., 2025b) seeks to bridge the gap in text-based emotion identification across 32 languages, focusing on emotion perception—how most people interpret a speaker's emotions from text (Mohammad, 2022). This is challenging as perceived emotions may differ from actual emotions due to cultural and individual factors (Woensel and Nevil, 2019; Wakefield, 2021). It is also important to highlight that Subtask A accounts for the complexity of expressed emotions in text by allowing multi-label emotion classification. This approach not only enhances the accuracy of emotion extraction but also acknowledges that some emotions can emerge from the interaction of two or more emotions, as described by Muhammad et al. (2025a).

Based on these outcomes, our study seeks to advance the field of text-based emotion detection by providing the following contributions:

1. Comprehensive evaluation and fine-tuning of fifteen pretrained monolingual and multilingual BERT-based models for emotion classification on English and Russian datasets, including those specifically trained for this task.

2. Analysis of the performance of BERT-based models without additional techniques or methods for emotion detection, aimed at evaluating their ability to handle this task independently.

3. Release of the best-performing multi-label emotion detection models for English and Russian.



*Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 532–538
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

This paper is structured as follows: Section 2 reviews state-of-the-art models for text-based emotion identification. Section 3 discusses dataset insights and our approach to architecture selection. Section 4 presents the experimental results. Section 5 provides a qualitative analysis and error categorization of the best-performing models. Finally, Section 6 concludes the paper.

## 2 Background

Several antecedent works have addressed the problem of emotion annotation and identification in text-based sources. A foundational contribution within the SemEval framework is by Mohammad et al. (2018), who introduced multi-label emotion annotation across 174,356 tweets in English, Arabic, and Spanish, covering twelve emotion categories (e.g., "anger" included related emotions like annoyance and rage). For English tweets, the best results were achieved using an ensemble of pretrained models, feature extraction, and traditional classifiers (e.g., SVM, logistic regression), reaching a Pearson's $r$ of 79.9. Building on this, Chatterjee et al. (2019) improved performance through ensemble methods combined with transfer learning. Since then, particularly in SemEval 2020, 2023, and 2024, BERT models (Devlin, 2018) and transfer learning have become the dominant approaches for emotion detection, consistently achieving high F1 scores (see Table 1).

| Description | F1 macro |
|---|---|
| Chatterjee et al. (2019) | 0.7731 |
| Sharma et al. (2020) | 0.33 |
| Vallecillo Rodrguez et al. (2023) | 0.8245 |
| Wang et al. (2024) | 0.3223 |

Table 1: F1 scores of SemEval BERT-based state-of-the-art models for text-only emotion identification

According to these results, BERT-based approaches are not only regarded as the most effective and preferred method for emotion detection in text within the SemEval NLP community, but their effectiveness has also been confirmed in recent studies by Imran (2024) and Aslan (2024). However, it is important to note that most of the reported results primarily involve English datasets for training and evaluation of the systems (Bujnowski et al., 2024; Šmíd et al., 2024).

Table 2 presents selected BERT-based models specifically trained for emotion detection in English and Russian, chosen based on their reported performance. These models were trained on large-scale emotion datasets using only BERT-based architectures, without additional optimization techniques, as reported by the authors in Table 1. Most were trained and evaluated on the GoEmotions dataset (Demszky et al., 2020), except Model 4, which uses the CEDR corpus for Russian (Sboev et al., 2021). The selection ensures consistency in emotion categorization, as all models overlap in their identification of the same emotional categories.

| # | Description | Languge | F1 macro |
|---|---|---|---|
| 1 | Sam Lowe (2024) | Eng | 0.54 |
| 2 | Pérez et al. (2021) | Eng | 0.45 |
| 3 | Seara (2021a) | Rus | 0.36 |
| 4 | Seara (2021b) | Rus | 0.74 |

Table 2: F1 scores of BERT fine-tuned models for text-based emotion classification

It is important to note that selecting and matching fine-tuned BERT models for emotion analysis in English and Russian is highly challenging due to the limited availability of resources. Consequently, we had to rely on two Russian-targeted models since, as mentioned earlier, the field of text-based emotion identification in Russian remains underexplored.

Building on these insights and in line with the contributions outlined in Section 1, we aim to propose a system for emotion identification in both English and Russian texts. The choice of these languages is motivated by two following reasons: (1) to introduce the Russian language into the SemEval competition board, as there is a notable lack of studies addressing this language; (2) to explore how a system with a similar architecture performs on two linguistically unrelated languages.

## 3 Datasets and Model

The dataset by Muhammad et al. (2025a) contains English and Russian short text snippets. We provide a statistical overview for each dataset in Table 3. It is worth noting that while the Russian dataset follows Ekman's traditional six-emotion classification (Ekman and Friesen, 1978; Ekman, 1992), the English dataset includes only five emotions due to an insufficient representation of the Disgust class (Muhammad et al., 2025a). The En-

| English | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **Class** | **Samples** | **%** | | **Samples** | **%** | | **Samples** | **%** |
| | Anger | 333 | 12.03 | | 16 | 13.79 | | 322 | 11.64 |
| | Fear | 1611 | 58.20 | | 63 | 54.31 | | 1544 | 55.80 |
| **Train** | Joy | 674 | 24.35 | **Dev** | 31 | 26.72 | **Test** | 670 | 24.21 |
| | Sadness | 878 | 31.72 | | 35 | 30.17 | | 881 | 31.84 |
| | Surprise | 839 | 30.31 | | 31 | 26.72 | | 799 | 28.88 |
| | **Total** | 2768 | 100 | | 116 | 100 | | 2767 | 100 |
| **Russian** | | | | | | | | | |
| | Anger | 543 | 20.27 | | 47 | 23.62 | | 226 | 22.60 |
| | Disgust | 273 | 10.19 | | 26 | 13.07 | | 122 | 12.20 |
| | Fear | 328 | 12.24 | | 21 | 10.55 | | 108 | 10.80 |
| **Train** | Joy | 555 | 20.72 | **Dev** | 34 | 17.09 | **Test** | 192 | 19.20 |
| | Sadness | 421 | 15.71 | | 39 | 19.60 | | 141 | 14.10 |
| | Surprise | 355 | 13.25 | | 26 | 13.07 | | 122 | 12.20 |
| | **Total** | 2679 | 100 | | 199 | 100 | | 1000 | 100 |

Table 3: Dataset sizes and emotion distribution for English and Russian in Subtask A

glish dataset exhibits a notable imbalance, with Fear dominating the training set at 58.2%, and other emotions, like Anger (12.03%), being more underrepresented. This trend continues in the development and test sets, where Fear makes up 54.31% and 55.80%. In contrast, the Russian dataset is more balanced, with Joy (20.72%, 17.09%, 19.20%) and Anger (20.27%, 23.62%, 22.60%) being the most common across the subsets. Other emotions like Disgust (10.19%, 13.07%, 12.20%) and Fear (12.24%, 10.55%, 10.80%) also have a reasonable presence.

Overall, while there is some variation in class sizes, the Russian dataset provides a more equitable distribution of emotions, which may benefit the model while predicting across different categories. In contrast, the English dataset is slightly more imbalanced, with Fear overrepresented, which may bias model toward this class.

### 3.1 Architecture selection

To address the multi-label classification of emotions in English and Russian textual data, we rely on sequence classification module from BERT pretrained models exclusively to predict probabilities across multiple emotion categories. The architecture of our proposed method is illustrated in Figure 1.

The BERT models we used have already demonstrated strong performance in emotion classification tasks (Creanga and Dinu, 2024). To adapt them for our specific task of multi-label emotion

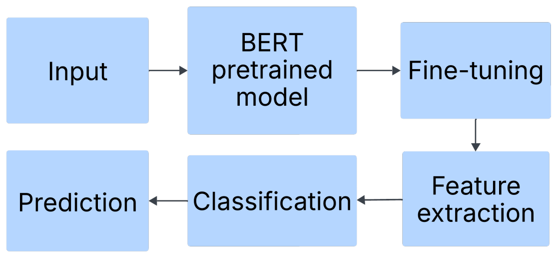


Figure 1: Model architecture

detection, we fine-tuned the models. We used a standard fine-tuning method with pretrained transformer models, based on the *transformers* library[1] from HuggingFace (Wolf et al., 2020). For each model, we experimented with different batch sizes (8, 16, 32) and learning rates ($1e^{-5}$, $2e^{-5}$, $3e^{-5}$, $4e^{-5}$, $5e^{-5}$), selecting the best checkpoint based on its F1 score on the development set after fifty epochs. The final hyperparameters set included a learning rate of $2e^{-5}$, a batch size of 16, and forty epochs.

We used pretrained multilingual and monolingual BERT-based models as the first step in our pipeline to preprocess input. This process involves reading text data from datasets, tokenizing it with a pretrained tokenizer, truncating and padding sequences, converting text into tensors, and finally creating batches for training and evaluation. The tokenized input is then passed through the pretrained transformer models. The models employs a self-

[1] https://github.com/huggingface/transformers

| English | | | |
|---|---|---|---|
| Model | Macro P | Macro R | Macro F1 |
| FacebookAI/roberta-base | 0.6441 | 0.8267 | **0.7170** |
| microsoft/deberta-v3-base | 0.4209 | 0.9334 | 0.5732 |
| google-bert/bert-large-cased | 0.6451 | 0.7129 | 0.6684 |
| google-bert/bert-base-multilingual-cased | 0.5457 | 0.6948 | 0.6073 |
| google-bert/bert-base-cased | 0.6109 | 0.7158 | 0.6492 |
| SamLowe/roberta-base-go_emotions | 0.6011 | 0.7652 | 0.6683 |
| FacebookAI/xlm-roberta-large | 0.6133 | 0.7665 | 0.6743 |
| finiteautomata/bertweet-base-emotion-analysis | 0.5983 | 0.7734 | 0.6672 |
| distilbert/distilbert-base-cased | 0.5945 | 0.7070 | 0.6363 |
| **Russian** | | | |
| DeepPavlov/xlm-roberta-large-en-ru | 0.7995 | 0.8522 | **0.8253** |
| FacebookAI/xlm-roberta-large | 0.7771 | 0.8729 | 0.8205 |
| google-bert/bert-base-multilingual-case | 0.8182 | 0.8019 | 0.806 |
| microsoft/deberta-v3-base | 0.2448 | 0.9679 | 0.3878 |
| seara/rubert-base-cased-russian-emotion-detection-ru-go-emotions | 0.7763 | 0.8176 | 0.7943 |
| r1char9/rubert-tiny2-ru-go-emotions | 0.498 | 0.8936 | 0.634 |
| DeepPavlov/distilrubert-base-cased-conversational | 0.7995 | 0.8422 | 0.8208 |
| DeepPavlov/rubert-base-cased | 0.7742 | 0.8548 | 0.8103 |
| DeepPavlov/rubert-base-cased-sentence | 0.7045 | 0.8269 | 0.7574 |

Table 4: The results on multilabel emotion detections task with BERT-based models and evaluated on English and Russian development set

attention mechanism across multiple layers to generate contextualized embeddings for each token. These embeddings capture semantic relationships between tokens in the context of the entire input sequence, allowing the model to understand the emotion conveyed in it.

For multi-label classification (5 for Russian or 6 for English emotion labels), the model produces logits that represent the prediction for each emotion class. The logits are passed through a sigmoid activation function to obtain probabilities ($p$) for each emotion, using the following formula:

$$p(x) = \frac{1}{1 + e^{-x}}$$

where $x$ is the logit for a given emotion. Since each emotion is independent, a sigmoid function is applied to each logit to obtain a probability between 0 and 1 for each emotion. The predictions are then used for evaluation or testing.

## 4 Experiment results

Using the architecture detailed above, we experimented with 15 different models for both English and Russian datasets. As shown in Table 4, the best-performing model for English was roBERTa-base, while for Russian, it was xlm-roBERTa-large trained with a reduced vocabulary and embeddings specifically tailored to Russian and English. Emotion-specific models like GoEmotions consistently underperform vanilla BERT-base, offering less precision on emotion classification task.

These development dataset evaluation results were subsequently tested on the final test dataset described in Section 3. The final F1 Macro scores achieved were *0.71* for English and *0.8418* for Russian, demonstrating the effectiveness of our chosen models and approach across both languages.

## 5 Discussion

The scores illustrated in Figure 2 represent the accuracy of our model in correctly predicted emotions. In English, Fear has the highest accuracy (82%), followed by Joy and Sadness, while Surprise and Anger perform slightly lower. The Russian dataset shows higher overall accuracy, with Joy at 91% and Fear 88%, while Disgust is the less accurate (77%).

These results could be attributed to both the model's performance and the dataset used for fine-tuning. Nonetheless, as discussed in Section 3, Fear is the most overrepresented emotion in the English dataset, which aligns with our model strong performance predicting it, as shows the Figure 2.

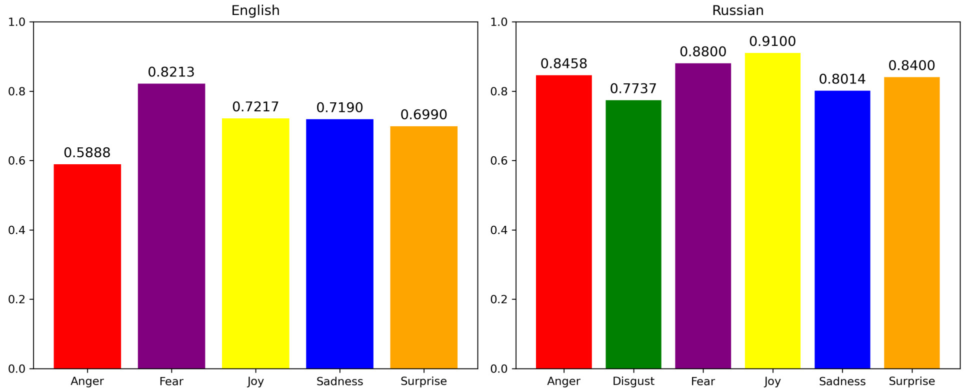


Figure 2: Accuracy of correctly labeled emotions for English and Russian datasets

On the contrary, Anger, being underrepresented in the English dataset, exhibits the lowest accuracy. This indicates that the model is moderately sensitive to class imbalance—performance on minority emotions drops, especially in the English dataset where Fear dominates. The Russian dataset is more balanced, allowing the model to maintain stable performance across different classes. As future research, we aim to improve this dataset by incorporating a more diverse range of text snippets, which, we hope, will benefit the model.

Finally, we analyzed the accuracy of our model in performing multilabel emotion classification based on textual data. As shown in Table 5, the model demonstrated higher accuracy in correctly predicted Russian text snippets compared to English. This suggests that multilabel emotion prediction is more reliable for the Russian dataset. These findings will be also considered in future research to improve multilabel accuracy further.

| Language | Total | Correct | % |
|---|---|---|---|
| English | 2767 | 1190 | 56.99 |
| Russian | 1000 | 771 | 77.10 |

Table 5: Multilabeling accuracy data of our model

## 6 Conclusions

To the best of our knowledge, this study presents several novel contributions to the field of emotion detection in English and Russian, which have not been explored in previous works:

First, the evaluation and fine-tuning of various BERT-based models for emotion detection reveal that monolingual BERT models achieve superior performance for English, while multilingual BERT models perform better for Russian. Notably, a BERT model with a reduced vocabulary and embeddings specifically tailored to Russian and English demonstrated the highest accuracy for the Russian dataset. Compared to the baseline by Muhammad et al. (2025b), which used only fine-tuned multilingual RoBERTa on each language's training data, our approach—fine-tuning the classifier layer with both multilingual and monolingual BERT models—achieved better results. We ranked 44th in Track A for English and 25th for Russian.

Second, the architecture of our model follows state-of-the-art principles used in emotion detection tools. While prior studies suggest that incorporating additional methods and techniques alongside BERT-based models enhances emotion detection performance, our experiments reveal that BERT models alone achieve sufficient accuracy.

In summary, we have contributed to the area of emotion detection in English and Russian by reutilizing the available BERT-based models, refining them for this specific task, which has provided positive results (71% for English and 84% for Russian) showing the high accuracy in multilingual emotion labeling and outperforming results, reported in previous works.

As for limitations, we have not yet extensively explored LLMs due to resource constraints, though we recognize their potential for improving contextual understanding in multilabel emotion classification. While this work focused on transformer-based models, we acknowledge that exploring alternative classifier architectures (e.g., Conv1D + MLP) and traditional methods (e.g., SVM, XGB) could enhance accuracy. We also aim to apply regularization and data augmentation techniques to reduce overfitting and improve generalization in future work.

## Acknowledgments

The authors are extremely grateful to Olga Zamaraeva, PhD in Computational Linguistics, for her technical and theoretical support. This study is funded by CSIC Momentum project, financed by European Union's Recovery and Resilience Facility-Next Generation, in the framework of the General Invitation of the Spanish Government's public business entity Red.es to participate in talent attraction and retention programmes within Investment 4 of Component 19 of the Recovery, Transformation and Resilience Plan.

## References

Zülfikar Aslan. 2024. Emotion detection via bert-based deep learning approaches in natural language processing. *The International Journal of Energy and Engineering Sciences*, 9(2):103–114.

Paweł Bujnowski, Bartłomiej Kuzma, Bartłomiej Paziewski, Jacek Rutkowski, Joanna Marhula, Zuzanna Bordzicka, and Piotr Andruszkiewicz. 2024. Samsemo: New dataset for multilingual and multimodal emotion recognition. In *Proc. Interspeech 2024*, pages 2925–2929.

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. SemEval-2019 task 3: EmoContext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Claudiu Creanga and Liviu P. Dinu. 2024. ISDS-NLP at SemEval-2024 task 10: Transformer based neural networks for emotion recognition in conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 649–654, Mexico City, Mexico. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. In *58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Paul Ekman and Wallace V Friesen. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*.

Lynn Greschner and Roman Klinger. 2024. Fearful falcons and angry llamas: Emotion category annotations of arguments by humans and llms. *arXiv preprint arXiv:2412.15993*.

Mia Mohammad Imran. 2024. Emotion classification in software engineering texts: A comparative analysis of pre-trained transformers language models. In *Proceedings of the Third ACM/IEEE International Workshop on NL-based Software Engineering*, pages 73–80.

Bing Liu. 2020. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge university press.

Isa Maks and Piek Vossen. 2011. A verb lexicon model for deep sentiment analysis and opinion mining applications. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pages 10–18.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Saif M Mohammad. 2022. Best practices in the creation and use of emotion lexicons. *arXiv preprint arXiv:2210.07206*.

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.

Juan Manuel Pérez, Juan Carlos Giudici, and Franco Luque. 2021. pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks. *Preprint*, arXiv:2106.09462.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

Sam Lowe. 2024. roberta-base-go$_emotions(revision 58b6c5b)$.

Alexander Sboev, Aleksandr Naumov, and Roman Rybka. 2021. Data-driven model for emotion detection in russian texts. *Procedia Computer Science*, 190:637–642.

Seara. 2021a. Rubert-base russian emotion detection.

Seara. 2021b. Rubert-base russian emotion detection trained on cedr dataset.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. SemEval-2020 task 8: Memotion analysis- the visuo-lingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773, Barcelona (online). International Committee for Computational Linguistics.

Jakub Šmíd, Pavel Přibáň, and Pavel Král. 2024. Uwb at wassa-2024 shared task 2: Cross-lingual emotion detection. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 483–489.

Enrica Troiano, Laura Oberländer, and Roman Klinger. 2023. Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. *Computational Linguistics*, 49(1):1–72.

Alessandro Valitutti, Carlo Strapparava, and Oliviero Stock. 2004. Developing affective lexical resources. *PsychNology J.*, 2(1):61–83.

María Estrella Vallecillo Rodrguez, Flor Miriam Plaza Del Arco, L. Alfonso Ureña López, and M. Teresa Martín Valdivia. 2023. SINAI at SemEval-2023 task 10: Leveraging emotions, sentiments, and irony knowledge for explainable detection of online sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 986–994, Toronto, Canada. Association for Computational Linguistics.

Jane Wakefield. 2021. Ai emotion-detection software tested on uyghurs. *BBC News*, 26.

Fanfan Wang, Heqing Ma, Rui Xia, Jianfei Yu, and Erik Cambria. 2024. SemEval-2024 task 3: Multimodal emotion cause analysis in conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2039–2050, Mexico City, Mexico. Association for Computational Linguistics.

Eileen Wemmer, Sofie Labat, and Roman Klinger. 2024. Emoprogress: Cumulated emotion progression analysis in dreams and customer service dialogues. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5660–5677.

Lieve Van Woensel and Nissy Nevil. 2019. What if your emotions were tracked to spy on you? *URL https://www. europarl. europa. eu/RegData/etudes/ATAG/2019/634415/EPRS ATA (2019)*, 634415.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. *arXiv preprint arXiv:2304.03347*.

Yuchen Zhang, Xing Su, Jia Wu, Jian Yang, Hao Fan, and Xiaochuan Zheng. 2023. Emoknow: Emotion- and knowledge-oriented model for covid-19 fake news detection. In *International Conference on Advanced Data Mining and Applications*, pages 352–367. Springer.