JellyK at SemEval-2025 Task 11: Russian Multi-label Emotion Detection with Pre-trained BERT-based Language Models

Khoa Anh-Nguyen Le[1,2], Dang Van Thin[1,2],
[1]University of Information Technology-VNUHCM, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
23520742@gm.uit.edu.vn, thindv@uit.edu.vn

Abstract

This paper presents our system developed for track A of the SemEval-2025 Task 11: Multi-label Emotion Detection. The primary goal of this task is to identify the emotions that most people would associate with a given sentence from a speaker, allowing multiple labels per instance. Our system focuses on detecting emotions in Russian-language sentences. To enhance model performance, we perform data pre-processing on special characters, punctuation, and expressions to better understand the relationship between textual features and emotion labels. We fine-tune a pre-trained language model specifically designed for Russian. To identify the best-performing model architecture, we employ a K-Fold Cross-Validation strategy during the model selection phase. Our final system achieved fourth place on the official leaderboard for the Russian sub-task.

1 Introduction

Emotion detection is a subfield of natural language processing (NLP) that involves identifying and classifying emotions expressed in text according to what most people are likely to perceive the speaker to be feeling. Importantly, the task does not aim to determine the actual emotional state of the speaker or the emotions of other entities mentioned in the text. For example, a sentence like "I am really happy now" would commonly be interpreted as expressing happiness due to the presence of emotional cues. This task has broad applications in various domains, making it an essential component in modern AI systems. In customer service, emotion detection enables companies to analyze user feedback, detect dissatisfaction, and improve overall user experience (Guo et al., 2024). In the mental health domain, it can help identify signs of emotional distress such as depression or anxiety from user input, supporting early intervention and automated screening processes (Francese and Attanasio, 2022). Given these applications, SemEval-2025 Task 11 focuses on the detection of multi-label emotion from sentences in several languages, including Russian, to model how the general public would interpret the emotional content of a speaker's statement (Muhammad et al., 2025b). In this paper, we describe our approach for Track A of the task, which includes comprehensive data preprocessing techniques, the use of a pre-trained Russian language model, and a K-Fold Cross-Validation strategy to identify potential model weaknesses and reduce overfitting. We also perform model selection to find the best-fitting architecture for our dataset. Our final system achieved fourth place on the official leaderboard for the Russian sub-task. The implementation of our system is available on Github[1].

2 Related Work

2.1 Models

The main goal of the multi-label classification problem is to find the relevance between classes and corresponding samples. Additionally, in the emotion detection problem, each emotion will correspond to special expressions or characters. There are many methods for this task, such as Decision Tree (Rokach and Maimon, 2005) or Support Vector Machine (SVM) (Evgeniou and Pontil, 2001). But today's language models outperform traditional machine learning algorithms (Liu et al., 2023). We tested many different language models that have been trained in Russian.

Since we want to fine-tune only Russian, we will focus on Encoder-only Models. Because of the support of HuggingFace, we can make predictions directly from the pre-trained model. We tested many different language models and compared them. These models are XLM-RoBERTa

[1]https://github.com/LeNguyenAnhKhoa/Russian-Emotion-Detection

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 2090–2095

Model	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10	Avg.↑
ruRoberta-large	0.848	**0.856**	**0.842**	**0.845**	0.805	**0.851**	**0.859**	**0.860**	**0.875**	**0.869**	**0.841**
ruBert-large	**0.856**	0.847	0.856	0.806	0.775	0.800	0.807	0.829	0.843	0.827	0.825
TwHIN-BERT-large	0.814	0.789	0.775	0.794	0.790	0.788	0.806	0.827	0.849	0.855	0.809
XLM-RoBERTa-large	0.788	0.782	0.795	0.747	**0.838**	0.798	0.806	0.795	0.828	0.849	0.803
XLM-RoBERTa-base	0.798	0.780	0.780	0.786	0.814	0.818	0.849	0.770	0.848	0.835	0.800
rubert-tiny2	0.817	0.787	0.761	0.804	0.810	0.765	0.827	0.732	0.798	0.848	0.795
TwHIN-BERT-base	0.803	0.795	0.766	0.747	0.811	0.798	0.761	0.815	0.824	0.809	0.793
DistilBERT	0.726	0.712	0.715	0.785	0.728	0.727	0.788	0.744	0.807	0.764	0.750

Table 1: F1-score after fine-tuning the models using the first 10 folds as validation set. We **bold** the best value of the folds.

(Conneau et al., 2019), DistilBERT (Sanh et al., 2019), TwHIN-BERT (Zhang et al., 2022), ruBert and ruRoberta (Zmitrovich et al., 2023). These models were trained on a large dataset including Russian.

2.2 Dataset

The provided Russian data are divided into 3 parts for training, validation, and testing (train/val/test, 2679/200/1000). The final ranking will be decided on the test set based on the last submission. Each sample in the datasets will have a sentence with 6 labels corresponding to 6 emotions: *Anger, Disgust, Fear, Joy, Sadness, Surprise* with value **1** for that emotion existing in the sentence and **0** otherwise. In addition, a distinct id is assigned to each sentence. For instance, the saying: *Hooray, I got an iron man* has the corresponding label **Joy**, so the value of Joy is 1 and the remaining emotions are 0 (Muhammad et al., 2025a).

3 System Overview

In this section, we describe the system in detail. We first clean the data and generate correlations between text and sentiments through data pre-processing. Then we use K-FOLD Cross-Validation to evaluate the model and analyze the mispredicted patterns. Furthermore, we use the sigmoid function (Han and Kaliraj, 1995) to make predictions with a single threshold for all labels. Finally, we experiment with different models to select the best models.

3.1 Pre-Processing

Each emotion type is associated with expressions in text or special punctuation and characters to express that emotion type. We encoded these expressions into Russian words that correspond to each emotion type. The Anger label is associated with red-faced expressions of anger, and we encoded these expressions as the word **anger**. In addition, anger emotion often appears in sentences with exclamation marks, and we encode exclamation marks (which are often grouped together and with the number 1) as a single exclamation mark. Next, the label Fear is associated with sentences with panic expressions (expressions with blue heads), we encoded it as **fear**. The emotion Sad is associated with sad facial expressions, crying faces, and a series of consecutive closing parentheses, which we encoded as **sadness**. The emotion of surprise is associated with a flushed face and two "O"s together (usually with a dot or underscore in the middle), which we encoded as the word **surprise**. The embarrassed expression (an underscore between two dots or two dashes) is associated with the three emotions Anger, Disgust, and Surprise, which we encode as the word **embarrassed**. Finally, the label Joy is associated with a smiling face, a heart, and a series of parentheses, which we encode as the word **joy**.

The data also contains swear words and asterisks, which are often associated with the anger label, we replace the asterisks to get the full word. In addition, we remove all remaining special characters such as periods, pound signs, or semicolons and replace them with spaces. Moreover, the dataset still contained spam words in which characters were excessively repeated, such as "xxXxxxx" or "OooOoo". To address this issue, we initially normalized these patterns by merging repeated characters into a single one. However, this approach unintentionally distorted some valid words, such as "running" being transformed into "runing". Therefore, we applied a reverse normalization step to restore such words to their correct forms, based on a predefined vocabulary or context-aware correc-

tion. Finally, there are some Russian letters that look similar to letters in the alphabet but are unique to Russian. For example, the characters "o" and "3" have different writings in Russian. All steps were implemented using the **regex** library.

3.2 K-FOLD Cross Validation

To evaluate the model and the data pre-processing methods, we used all labeled samples (including the validation set) and divided them into k folds. Due to the limited amount of data, we divided them into 30 folds, with 29 folds for the training set and only 1 fold for the validation set. We tested each model 10 times on 10 different validation folds to get an overview of the model. The standard binary cross-entropy loss (BCE) is used to optimize the model.

3.3 Model selection

We evaluated the model on the validation set using K-FOLD Cross-Validation, we computed the average F1-score (Doe and Smith, 2020) over the validation set. The best model is the one with the highest average value. Finally, we fine-tune the best model on the entire labeled dataset (training set and validation set) and use this model to make our final submission. According to Table 1, we choose **ruRoberta-large** as the best model.

4 Experimental Setup

Model	Pre-processing	Batch size	F1-score
ruRoberta-large	Yes	32	**0.841**
ruRoberta-large	No	32	0.831
ruBert-large	Yes	32	0.825
ruBert-large	No	32	0.819
rubert-tiny2	Yes	128	0.795
rubert-tiny2	No	128	0.788
DistilBERT	Yes	128	0.750
DistilBERT	No	128	0.739
XLM-RoBERTa	Yes	128	0.800
XLM-RoBERTa	No	128	0.789
XLM-RoBERTa-large	Yes	128	0.803
XLM-RoBERTa-large	No	128	0.799
TwHIN-BERT-base	Yes	64	0.793
TwHIN-BERT-base	No	64	0.785
TwHIN-BERT-large	Yes	16	0.809
TwHIN-BERT-large	No	16	0.798

Table 2: Performance of different models on the validation set based on pre-processing

We use the AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of 10^{-5}, a weight decay of 0.01, and an epsilon of 0.09. Due to GPU limitations, each model has a different batch size, as detailed in Table 2. Furthermore, we conducted a small search to find the optimal threshold, using **ruRoberta-large** as the model and the F1-score to evaluate it in the validation set. The complete search results are shown in Table 5.

Model	F1-score	Language	Size (GB)
ruRoberta-large	**0.841**	Russian	1.42
ruBert-large	0.825	Russian	1.71
TwHIN-BERT-large	0.809	Multilingual	2.25
XLM-RoBERTa-large	0.803	Multilingual	2.24
XLM-RoBERTa-base	0.800	Multilingual	1.12
rubert-tiny2	0.795	Russian	0.12
TwHIN-BERT-base	0.793	Multilingual	1.12
DistilBERT	0.750	Multilingual	0.54

Table 3: Comparison of models based on F1-score, language, and size.

Rank	System	Score
1	Heimerdinger	0.9008
2	JNLP	0.8912
3	CSIRO-LT	0.8910
4	**Ours**	**0.8890**

Table 4: Track A performance on the test set.

5 Results

5.1 Main Result

Based on Table 3, we see that the fine-tuned model only on Russian gives better results. Additionally, within the same model type, the larger version will give better results. The ruRoberta-large is the best model when it outperforms all other models with the best F1-score. Data pre-processing plays a crucial role in helping the model establish the relationship between words and their corresponding emotions. As shown in Table 2, all models performed better when pre-processing was applied. We also found that the optimal threshold for all labels lies between 0.4 and 0.5. In our final submission, we set **0.43** as the threshold and fine-tuned **ruRoberta-large** throughout the training and development sets. There were a total of 53 final submissions on Track A for Russian (rus), our system ranked fourth with a score of **0.889** on the test set, as shown in Table 4.

5.2 Error Analysis

In this section, we will analyze in detail the cases where our model went wrong. First, the relationship between words and their corresponding emo-

tions is so tight that any sentence containing that word always predicts the corresponding emotion. For example, in the sentence "Yeah, damn it! Yes baby, you gotta be awesome", the phrase "damn it" is associated with the emotion Anger, but this sentence has the emotion Joy. Secondly, our model cannot distinguish between the speaker's and referred person's emotions. For instance, in the text "Turtles are afraid of small spaces", the word "afraid" describes the turtle's sad emotion, not the speaker's, but our model predicts the turtle's feelings. Moreover, happy expressions often appear on the Joy and Surprise labels, confusing our model when predicting these two emotions. The text "Oh my gosh, what's going on tonight? :3 I think it's normal" has the emotion ":3" which usually appears in sentences labeled Joy but appears in sentences labeled Surprise. Finally, in a text, there are two sentences with two different emotions like "I love you. I'm so sad now", our model can only predict the label Joy or the label Sadness but cannot predict both.

6 Conclusion

In this paper, we introduced a good system to classify the speaker's emotions for the SemEval Challenge 2025 Task 11 track A. The key point of our system is to find the relationship between data and labels by pre-processing and testing different models to choose the model that fits the dataset. Our system achieved a top five ranking for Russian in the rankings.

Acknowledgements

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

John Doe and Jane Smith. 2020. A comprehensive study on macro f1-score. *Journal of Machine Learning Metrics*, 10(2):123–145.

Theodoros Evgeniou and Massimiliano Pontil. 2001. *Support Vector Machines: Theory and Applications*, pages 249–257. Springer Berlin Heidelberg, Berlin, Heidelberg.

Rita Francese and Pasquale Attanasio. 2022. Emotion detection for supporting depression screening. *Multimedia Tools and Applications*, 82:12771 – 12795.

Yiting Guo, Yilin Li, De Liu, and Sean Xin Xu. 2024. Measuring service quality based on customer emotion: An explainable ai approach. *Decision Support Systems*, 176:114051.

J. Han and P. K. Kaliraj. 1995. Influence of the sigmoid function parameters on the speed of backpropagation learning. *Neural Networks*, 8(3):351–362.

Rundong Liu, Wenhan Liang, Weijun Luo, Yuxiang Song, He Zhang, Ruohua Xu, Yunfeng Li, and Ming Liu. 2023. Recent advances in hierarchical multi-label text classification: A survey. *arXiv preprint arXiv:2307.16265*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

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 Zhangand Yi Zhou, and Saif M. Mohammad. 2025a. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages.

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.

Lior Rokach and Oded Maimon. 2005. *Decision Trees*, pages 165–192. Springer US, Boston, MA.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2022. Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations. *arXiv preprint arXiv:2209.07562*.

Dmitry Zmitrovich, Alexander Abramov, Andrey Kalmykov, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Tatiana Shavrina, Sergey Markov, Vladislav Mikhailov, and Alena Fenogenova. 2023. A family of pretrained transformer language models for russian.

A Finding optimal threshold

We conducted a grid-search for an optimal threshold of 0.3 to 0.6 using the **ruRoberta** model during 10 folds.

Threshold	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10
0.30	0.851	0.829	0.830	0.829	0.824	0.843	0.854	0.840	0.917	0.851
0.31	0.851	0.829	0.830	0.829	0.824	0.843	0.854	0.840	0.917	0.851
0.32	0.851	0.829	0.823	0.829	0.821	0.843	0.854	0.840	0.917	0.857
0.33	0.860	0.837	0.823	0.829	0.830	0.843	0.854	0.837	0.917	0.857
0.34	0.860	0.834	0.823	0.829	0.830	**0.859**	0.854	0.829	0.910	0.861
0.35	0.860	0.834	0.823	0.829	0.830	**0.859**	0.854	0.829	0.910	0.861
0.36	0.852	0.838	0.826	0.834	0.832	**0.859**	0.863	0.832	0.914	0.866
0.37	0.852	0.838	0.826	0.834	0.832	0.845	0.863	0.832	0.914	0.859
0.38	0.852	0.843	0.826	0.834	0.832	0.845	0.863	0.832	0.914	0.859
0.39	0.858	0.843	0.826	0.834	0.832	0.845	0.866	0.832	**0.923**	0.867
0.40	0.854	0.843	0.826	0.834	0.832	0.845	**0.872**	0.832	**0.923**	0.870
0.41	0.857	0.838	0.826	0.834	0.832	0.845	0.869	0.832	0.915	0.870
0.42	0.857	0.838	0.826	0.834	0.829	0.845	0.869	0.832	0.915	0.870
0.43	0.857	0.838	0.826	0.834	0.829	0.842	0.869	**0.843**	0.915	0.870
0.44	0.857	0.838	0.826	0.828	0.829	0.842	0.869	**0.843**	0.915	0.870
0.45	0.857	0.838	0.826	**0.840**	0.829	0.842	0.814	0.832	0.915	0.870
0.46	0.857	0.838	0.831	**0.840**	0.829	0.842	0.814	0.815	0.907	0.870
0.47	**0.880**	**0.844**	0.831	**0.840**	0.829	0.842	0.814	0.809	0.907	**0.878**
0.48	0.876	**0.844**	0.838	0.835	0.829	0.842	0.819	0.809	0.907	**0.878**
0.49	0.876	**0.844**	0.828	0.835	0.829	0.842	0.819	0.809	0.907	0.874
0.50	0.879	0.843	0.828	0.835	**0.832**	0.842	0.821	0.809	0.907	0.874
0.51	0.879	0.843	**0.833**	0.835	**0.832**	0.842	0.821	0.809	0.907	0.874
0.52	0.873	0.843	**0.833**	0.828	**0.832**	0.842	0.821	0.809	0.907	0.874
0.53	0.873	0.843	**0.833**	0.828	0.830	0.842	0.821	0.803	0.907	0.874
0.54	0.879	0.843	0.831	0.828	0.830	0.842	0.821	0.803	0.907	0.874
0.55	0.879	0.837	0.831	0.828	0.830	0.839	0.813	0.803	0.889	0.874
0.56	0.879	0.837	0.831	0.819	0.830	0.839	0.813	0.803	0.889	0.859
0.57	0.879	0.837	0.831	0.819	0.830	0.839	0.813	0.803	0.889	0.859
0.58	0.860	0.837	0.831	0.819	0.830	0.839	0.813	0.794	0.889	0.840
0.59	0.860	0.837	0.831	0.819	0.817	0.839	0.813	0.794	0.889	0.840
0.60	0.860	0.837	0.831	0.819	0.817	0.839	0.811	0.794	0.889	0.840

Table 5: Results of a grid-search on 10 different fold validation sets using the ruRoberta model. The highest results are in **bold**.