Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2023"

June 14–16, 2023

Knowledge Transfer Between Tasks and Languages in the Multi-task Encoder-agnostic Transformer-based Models

Dmitry Karpov
MIPT
Dolgoprudny, Russia
dmitrii.a.karpov@phystech.edu

Vasily Konovalov
MIPT
Dolgoprudny, Russia
vasily.konovalov@phystech.edu

Аннотация

We explore the knowledge transfer in the simple multi-task encoder-agnostic transformer-based models on five dialog tasks: emotion classification, sentiment classification, toxicity classification, intent classification, and topic classification. We show that these models' accuracy differs from the analogous single-task models by ∼0.9%. These results hold for the multiple transformer backbones. At the same time, these models have the same backbone for all tasks, which allows them to have about 0.1% more parameters than any analogous single-task model and to support multiple tasks simultaneously. We also found that if we decrease the dataset size to a certain extent, multi-task models outperform single-task ones, especially on the smallest datasets. We also show that while training multilingual models on the Russian data, adding the English data from the same task to the training sample can improve model performance for the multi-task and single-task settings. The improvement can reach 4-5% if the Russian data are scarce enough. We have integrated these models to the DeepPavlov library and to the DREAM dialogue platform.

Keywords: multi-task, transformer, neural, dialog, emotion, sentiment, toxic, knowledge transfer, cross-lingual knowledge transfer, multi-task knowledge transfer, conversational tasks

DOI: 10.28995/2075-7182-2023-22-200-214

Перенос знаний между задачами и языками в многозадачных энкодер-агностичных моделях типа Трансформер

Дмитрий Карпов
Василий Коновалов
Московский физико-технический институт / Долгопрудный, Россия
dmitrii.a.karpov@phystech.edu vasily.konovalov@phystech.edu

Аннотация

В статье изучается перенос знаний в многозадачных энкодер-агностичных моделях типа Трансформер для пяти диалоговых задач – классификации эмоций, сентимента, токсичности, интентов и тематической классификации. В статье показано, что эти модели демонстрируют точность, отличающуюся от аналогичных однозадачных моделей примерно на 0.9%. Эти результаты верны для разных типов трансформеров. В то же время эти многозадачные модели имеют примерно на 0.1% больше параметров, чем любая из аналогичных однозадачных моделей. В статье также показывается, что начиная с определенного достаточно маленького размера набора данных, многозадачные модели начинают превосходить однозадачные модели, особенно на тех задачах, для которых меньше всего данных. Помимо этого, при обучении многоязычных моделей на русскоязычных данных, добавление англоязычных данных в обучающую выборку дополнительно улучшает результат многоязычных моделей в однозадачном и многозадачном режиме. Улучшение может достигать 4-5%, если русскоязычных данных достаточно мало. Эти модели также были интегрированы в библиотеку DeepPavlov и диалоговую платформу DREAM.

Ключевые слова: многозадачные, трансформер, нейросетевые, диалог, эмоции, тональность, токсичность, перенос знаний, межязыковой перенос знаний, многозадачный перенос знаний, разговорные задачи

1 Introduction

Transformer-based models, such as BERT, are widely used for text classification. The original article (Devlin et al., 2019) proposed the use of a separate BERT model for each task in multi-task benchmarks. Therefore, if several classification tasks need to be solved in parallel, several prediction models should be employed, which increases the demand for computational resources. One of the ways of tackling this problem is training one single model that can yield results for these tasks simultaneously.

Multi-task learning (MTL) is one of the transfer-learning techniques. It allows training one single model simultaneously for multiple related tasks so that the knowledge acquired in one task enhances another task's performance.

The real-world conditions require making a trade-off between the quality of neural models and their use of computational resources. Responding to this tradeoff necessitates the use of encoder-agnostic multitask transformer-based models, which allows quick replacement of the transformer backbone for different circumstances. The `transformers` (Wolf et al., 2020) library allows using different transformer-based models including distilled ones to save computational resources and speed up the inference time (Kolesnikova et al., 2022).

Our contributions are as follows:

1. We show how multi-task knowledge transfer occurs in the simple encoder-agnostic transformer-based models during training for multiple dialogue-related tasks.
2. We explore the effects of multi-lingual knowledge transfer in these models.
3. We implement these models in `DeepPavlov` framework.[1]

2 Related Work

Researchers have been conducting experiments with multi-task learning (MTL) for a long time (Caruana, 1997). Since the rise of neural networks, researchers have proposed a wide range of approaches to MTL, including cross-lingual word embeddings (Konovalov and Tumunbayarova, 2018). However, these methods did not develop further, as nowadays NLP is based on transformer-based models. Nevertheless, as transformer architectures come out quite often, this review mostly focuses on agnostic architectures, which work with all kinds of transformers, rather than transformer-specific architectures.

In some kinds (Karpov and Burtsev, 2021) of multi-task encoder-agnostic transformer-based architectures, every sample needs to be labeled or pseudo-labeled for all considered tasks. Even though this approach is successfully used in some dialogue systems (Kuratov et al., 2021), it lacks flexibility.

One of the most frequently used encoder-agnostic transformer-based architectures is (Liu et al., 2019). However, this architecture increases computational demands due to the specific stochastic attention layers for text classification.

The work (Asa Cooper Stickland, 2019) proposed different encoder-agnostic ways to work with BERT output in a multi-task setting.[2] One such way is to supplement the model with an extra BERT layer for each task. However, that way increases the number of required parameters for GLUE (Wang et al., 2018) by 67%, which is computationally heavy. Other encoder-agnostic approaches proposed in the same work worked no better than the plain use of *bert-base-uncased* output in the linear classifier in our experiments on GLUE.

Additionally, utilizing self-attention with a task-embedded module from the paper (Maziarka and Danel, 2021) instead of plain self-attention in the low-rank transformation did not yield improvements over the plain dense task-specific layers on top of BERT in our experiments. The task-embedded architecture presented in the same article is still not encoder-agnostic.

Another work (Huang et al., 2021) suggested a novel way to extract additional features from the BERT output – using lightweight convolutional ghost modules. Despite this approach being encoder-agnostic, utilizing attention with a ghost module in the low-rank transformation did not yield improvements over the plain dense task-specific layers above BERT in our experiments. This also holds for (Ali et al., 2021) architecture from computer vision.

[1] `https://github.com/deeppavlov/DeepPavlov/`

[2] Projective attention layers, presented in the same article as the superior result, are not encoder-agnostic.

At the same time, the performance of simple encoder-agnostic transformer-based models is still not fully explored. It is especially true for dialogue-specific datasets. Furthermore, the body of work lacks studies on the Russian language multi-task learning in general, and specifically on the dialogue tasks. Multilingual knowledge transfer in multi-task models for such tasks also remains unexplored. Our work is aimed to bridge this gap.

3 MTL Model Description

The MTL architecture we explore allows using different encoder-only Transformer architectures as a backbone. For our experiments, we utilized BERT-based models because they allow effective transfer learning (Chizhikova et al., 2023; Konovalov et al., 2020). However, the same approach can be applied to any Transformer-based model.

1. In the same way as in the original work (Devlin et al., 2019), we return the final hidden states for all tokens and apply the BERT pooling layer to them. Like in this article, we apply the pooler output.
2. Then, for every task, we apply the task-specific linear layer to the pooler output. The task-specific linear layer for every task type looks exactly like the linear fine-tuning layer for the single-task BERT models (see original article or Transformers (Wolf et al., 2020) manual).
3. Then we apply a loss function: mean squared error loss for the regression tasks, categorical cross-entropy loss for single-label tasks or binary cross-entropy loss for multi-label classification task. In this paper, we consider only the single-label classification.

The multi-task model in this setting requires almost no additional parameters and computational overhead, apart from the linear layers, so its simplicity singles it out. Also, the flexibility of this model allows using it with different kinds of backbones, which positively distinguishes it from (Asa Cooper Stickland, 2019).

For the distilBERT-like models, this multi-task model takes only 0.1% more parameters than single-task models. This computational overhead varies around this number, depending on the number of tasks, the number of classes, and the backbone model.

The encoder-agnostic multi-task transformer-based model is integrated into DeepPavlov (Burtsev et al., 2018). This implementation supports all Transformer backbones from the `AutoModel` class from HuggingFace. Our implementation is also successfully used in the Dream dialogue platform (Baymurzina et al., 2021).

4 Datasets

We explored the multi-task models' performance on the Russian and English datasets for five tasks, i.e. emotion classification, toxicity classification, sentiment classification, intent classification, and topic classification. For Russian and English data, the indexes of the same classes used by the models were also the same. We chose these tasks as they are pivotal for dialog systems (Kuratov et al., 2020). The datasets contain naturally occurring data, which are useful for dialogue systems development (Konovalov et al., 2016b; Konovalov et al., 2016a). Therefore, we consider these tasks to be conversational tasks.

4.1 Emotion Classification

We used the `go_emotions` dataset (Demszky et al., 2020) for emotion classification in English. This dataset consists of short comments from Reddit, such as *LOL. Super cute!* or *Yikes. I admire your patience*. We used Ekman-grouped emotions, grouping them into seven types, i.e *anger*, *fear*, *disgust*, *joy*, *surprise*, *sadness*, and *neutral*. After such grouping, we selected only single-label examples. There were 39,555 training examples of that kind. The train/test/validation split of this dataset was approximately 80/10/10.

For the same task in Russian, we used the `CEDR` dataset (Sboev et al., 2021). The dataset contains examples from different social sources: blogs, microblogs, and news. This dataset has five classes – *anger*, *fear*, *joy*, *surprise*, and *sadness* – but the samples from this dataset can belong to more than one single class or (unlike `go_emotions`) belong to no class. For example, the text *Надо утопать на встречу.* belongs to no class.

From this dataset, we selected only examples that belong to one single class or that have no class, labeling no-class examples as *neutral*. The class nomenclature of this dataset was almost the same as for the English dataset, except for the *disgust* class. Nonetheless, as *disgust* examples comprised less than 1.5% of the English training samples, it didn't impact knowledge transfer much.

The work (Sboev et al., 2021) provided only the train-test split of the `CEDR` dataset, which is 80/20. We singled out 12.5% of the training examples from `CEDR` as the validation set. The resulting dataset has 6,557 training samples.

4.2 Sentiment Classification

We used the `DynaSent(r1)` dataset (Potts et al., 2020) for sentiment classification in English. It contains naturally occurring sentences. i.e. *Need a cheap spatula?* We used only examples from the first round of the collection, to match the Russian data by difficulty. This single-label dataset with 80,488 training samples has three classes – positive, negative, and neutral. The dataset has 3,600 validation samples and the same number of test samples.

For the same task in Russian, we used the `RuReviews` dataset (Smetanin and Komarov, 2019). This three-class dataset consists of 90,000 product reviews from the "Women's Clothes and Accessories" category of a large Russian e-commerce website. As the considered product reviews already contain grades from the user, the authors of this dataset classified sentiment according to the grades. For example, the phrase *размер очень мал* was considered to be negative. We have chosen this dataset because it is open source and it has a relatively large size, even though it is domain-specific. As the train/validation/test split of this dataset was not provided, we used the same split as in the `DynaSent(r1)` dataset. After that, the training set had 82,610 training samples.

4.3 Toxicity Classification

For English toxic classification, we used the `Wiki Talk` dataset (Dixon et al., 2018). This Wikipedia comment dataset has two classes: toxic and not toxic. Unsurprisingly, the dataset contains vulgar slang. However, about 90% of examples from this dataset are not toxic, i.e. *Hi! so umm i guess yer incharge here hehehe. so wassup?*. This dataset has 127,656 training samples, 93,342 validation samples, and 31,915 testing samples. For Russian toxic classification, we used the `RuToxic` dataset (Dementieva et al., 2021). This two-class dataset was collected from Dvach, a Russian anonymous imageboard. This dataset originally has 163,187 samples. Among them, most of the samples are not toxic, e.g. *ещё бы. какой красавец.*. But obviously, some samples are toxic, e.g. *дворника тоже надо уничтожить!*. As the authors didn't provide the original split in their repository, we split this dataset in the same proportions as in the `Wiki Talk` dataset. After that, the training set had 93,342 training samples.

4.4 Intent Classification and Topic Classification

We used `MASSIVE` dataset (FitzGerald et al., 2022) for the intent classification for the Russian and English languages. The MASSIVE dataset for the English language contains the spoken utterances, which aim for the voice assistant, e.g. *play rock playlist*. All examples in this dataset were labeled and adapted simultaneously for 51 languages, including Russian.[3] This dataset has 11,514 train samples, 2,033 validation samples, and 2,974 test samples. Every sample belongs to one of 60 intent classes. This dataset is widely used for the conversational topic classification (Karpov and Burtsev, 2023).

We used the same dataset in the same way for topic classification as well, as this dataset is labeled by intent and by topic. Every sample from this dataset belongs to one of the 18 topic classes.

5 Experimental Setup

For all the experiments described in our work, the optimizer was AdamW (Kingma and Ba, 2015) with betas (0.9, 0.99), and the initial learning rate was 2e-5. We used average accuracies for all tasks as an early stop metric. The training had validation patience 3, and the learning rate was dropped by two times if the early stopping metric did not improve for two epochs.

[3]For example, the Russian dataset contains sample *расскажи новости russia today* instead of *stell me b. b. c. news*.

Table 1: Accuracy / F1-macro on the English data for the encoder-agnostic transformer-based model. English cased models trained on English data. Mode S stands for single-task, and mode M stands for multi-task.

Model	Mode	Average	Emotions 39.4k	Sentiment 80.5k	Toxic 127.6k	Intents 11.5k	Topics 11.5k	Batches seen
distilbert	S	**82.9/78.4**	**70.3/63.1**	74.7/74.3	**91.5/81.2**	**87.4/82.7**	**91.0/90.6**	11390
distilbert	M	82.1/77.2	67.7/60.7	**75.2/75.0**	90.6/79.8	86.3/80.4	90.8/90.1	14000
bert	S	**83.9/79.7**	**71.2/64.2**	76.1/75.8	**93.2/83.5**	**87.9/84.2**	**91.3/90.7**	9470
bert	M	83.0/78.4	69.0/63.1	**76.5/76.4**	91.4/80.8	87.1/81.2	91.2/90.6	11760

The training was usually completed in less than 10-15 epochs and never exceeded 25 epochs, even though the maximum number of epochs was set to 100.

We set the batch size to 160. We have also tried batch size 32, and the metrics for batch size 160 were just insignificantly better. However, the paper (Godbole et al., 2023) claims that this difference can be eliminated by better fine-tuning. Finally, we settled with batch size 160 because the computations with batch size 160 were performed several times faster.

In the preliminary multi-task experiments, apart from plain sampling (a sampling mode where the example sampling probability is proportional to the task size), we also tried annealed sampling (Asa Cooper Stickland, 2019) and uniform sampling (the same sampling probability for all tasks). We performed such experiments for Russian and English distilbert-like models, for Russian and English tasks. The results for these sampling modes did not bring out a noticeable improvement, thus we used only plain sampling.

We averaged all the experiment results by three runs.

6 Results and Analysis

We conducted experiments in mono-lingual mode with different transformer-based backbones to compare single-task and multi-task approaches. For the English-language tasks, we conducted the experiments for the backbones *bert-base-cased* (Devlin et al., 2019) (*bert*) and *distilbert-base-cased* (Sanh et al., 2019) (*distilbert*).

For the Russian-language tasks, we made experiments for the backbones *DeepPavlov/rubert-base-cased-conversational* (Kuratov and Arkhipov, 2019) (*rubert*) and *DeepPavlov/distilrubert-base-cased-conversational* (Kolesnikova et al., 2022) (*distilrubert*).

As distilled BERTs take 40% less memory than BERTs and are 60% faster, these experiments cover a variety of different model uses for different computational budgets and quality demands.

6.1 Single-task VS Multi-task: Backbones From Different Languages

In the first stage of the experiments, we compared the performance of our multi-task models to analogous single-task models with the same hyperparameters.

We present the results of the first stage of experiments in Tables 1-2. For every experiment, we provide accuracy / macro-averaged F1.

Overall, the performance of multi-task encoder-agnostic transformer-based models closely matches the performance of the analogous single-task models. This effect holds for the Russian language as well as for the English language.

While *distilbert* shows slightly worse metrics than *bert*, *distilrubert* even excels *rubert* on all but the largest tasks.

In the next experiments, we put the main focus on the distilbert-like models to speed up the computations.

Table 2: Accuracy / F1-macro on the Russian data for the encoder-agnostic transformer-based model. Russian cased models trained on Russian data. Mode S stands for single-task, and mode M stands for multi-task. RU means that models were trained and evaluated on Russian data, EN means that models were trained and evaluated on English data.

Model	Mode	Average	Emotions 6.5k	Sentiment 82.6k	Toxic 93.3k	Intents 11.5k	Topics 11.5k	Batches seen
distilrubert	S	**86.9/84.1**	**82.2/76.1**	**77.9/78.2**	**97.1/95.4**	**86.7/81.6**	90.4/89.5	8472
distilrubert	M	86.3/82.6	81.0/74.6	77.7/77.7	96.9/95.0	85.2/75.9	**90.7/89.9**	8540
rubert	S	**86.5/83.4**	**80.9/75.3**	**78.0/78.2**	**97.2/95.6**	**86.2/79.1**	90.0/89.0	7999
rubert	M	86.2/82.6	80.5/73.8	77.6/77.6	96.8/95.0	85.3/76.9	**90.5/89.8**	8113

Table 3: Accuracy / F1-macro on the Russian data for the encoder-agnostic transformer-based model. Multilingual cased models, batch size 160, plain sampling. Mode S stands for single-task, and mode M stands for multi-task. In the 'Training data' column, RU stands for the Russian language, 'RU+EN' means that Russian and English data are merged by task, and 'RU $\oplus$ EN' means that Russian and English tasks are treated as separate tasks.

Model	Training data	Mode	Average	Emotions	Sentiment	Toxic	Intents	Topics	Batches seen
distilbert-mult	RU	S	**84.7/81.0**	77.4/69.1	**77.7/77.9**	**96.7/94.8**	**83.5/76.6**	88.1/86.9	10058
distilbert-mult	RU	M	84.3/80.2	**78.1/70.5**	76.8/76.7	96.5/94.4	81.9/72.3	**88.2/87.1**	9821
distilbert-mult	RU+EN	S	**85.2/81.8**	**78.9**/70.2	**77.4/77.3**	**96.8/94.9**	**84.7/79.1**	**88.4/87.4**	31843
distilbert-mult	RU+EN	M	84.5/81.1	77.9/**70.7**	76.6/76.7	96.5/94.5	82.9/76.5	**88.4**/87.2	17790
distilbert-mult	RU $\oplus$ EN	M	84.4/80.6	77.6/70.0	76.8/77.1	96.5/94.5	82.4/73.9	88.3/87.2	23688
bert-mult	RU	S	84.7/80.2	76.6/64.2	**77.8/78.2**	**96.9/95.1**	**83.9**/76.3	88.4/87.0	10884
bert-mult	RU	M	**84.8/81.4**	**78.4/71.4**	76.3/76.3	96.8/94.8	83.7/**76.6**	**89.0/87.8**	12810
bert-mult	RU+EN	S	**85.6/82.3**	78.9/70.1	**77.6/77.8**	**96.9/94.9**	**85.0/80.4**	**89.4/88.5**	23752
bert-mult	RU+EN	M	85.2/**82.3**	**79.2/72.7**	76.4/76.6	96.7/94.8	84.3/79.3	**89.4**/88.3	20755
bert-mult	RU $\oplus$ EN	M	85.0/81.6	78.3/71.4	77.1/77.0	96.7/94.7	84.0/76.7	89.1/88.0	22701

6.2 Multilingual Multi-task Backbones: Cross-lingual Training Impact

In the next stage of experiments, we have put the focus on multilingual knowledge transfer. To investigate this transfer, we utilized only multilingual backbones. In particular, we used *distilbert-base-multilingual-cased* and *bert-base-multilingual-cased*. In Table 3, we label them as *distilbert-mult* and *bert-mult*, respectively. Our main goals were:

- To compare the performance of the multi-task and single-task models with the multilingual backbones for the Russian language.
- To check how the performance of single-task models and the performance of multi-task models varies if we add the English data to them, and the data are merged by task (for every task, the model is trained on English+Russian training data and validated on Russian data).
- To check whether treating English-language tasks as separate tasks yields any improvements if we perform the validation on the Russian data.

As we see, the results of all settings are pretty similar: using Russian+English data puts us on the plateau, while improvements compared to using only Russian data are only moderate.

In the same setting, we also explored whether utilizing English-language tasks as separate tasks is more beneficial than merging Russian and English data by task. This approach did not prove to be any better and even brought out a small deterioration.

The impact of adding English data in case of having limited Russian data required additional investigation. We have researched this impact in the next series of experiments. In real-world conditions, we usually have a huge body of datasets for English data, but not nearly as much for Russian data. This

gives additional practical value to that experiments.

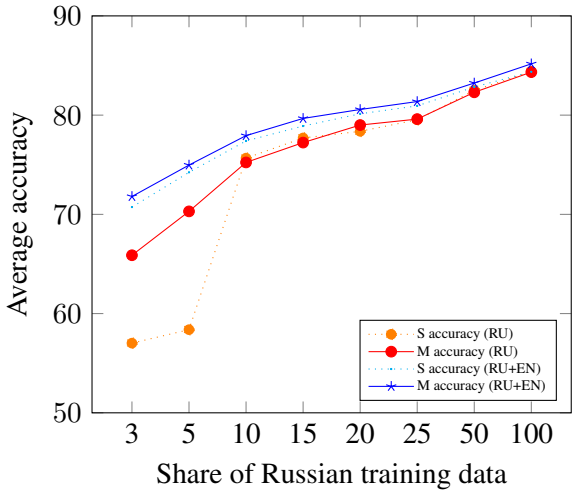
6.3 Impact of Adding the English Data

In this experiment series, we explored multi-task and single-task settings with Russian and English data merged by task. We studied how much the performance of *distilbert-base-multilingual-cased* (multi-task or single-task) improves when it is trained on some part of Russian train data if we add English training data to it and validate on the English validation data.

Specifically, we performed experiments for the following data shares: 0%, 3%, 5%, 15 %, 20%, 25%, 50%, and 100%. For 0%, we added to the table the model trained on English train data and validated on Russian validation data, and the model which is trained on English train data and validated on English validation data (but tested still on Russian test data). We restarted the experiments with three random seeds. For every series of experiments, we randomly shuffled the datasets and then selected all subsets at once, while the larger subsets contained all examples from the smaller subsets (like, 10% subset contains all examples from 5% and also from 3%)

We present the averaged results in Table 9, in Appendix. We averaged the results by three runs. For training on the 3-5% of the Russian data without the English data, we averaged the results by five runs due to the high variability of results. Additionally, we plot the results below, in Figure 1. The task-wise results for the experiments with data shares are also shown in Appendix, in Table 9.

We also note that in all the experiments from Table 9 where 100% share of the English data was used, we performed the experiments also with validation on the Russian data instead of the English data. That change did not impact the scores in any meaningful way (see Table 10).



RU share	S RU	M RU	S RU+EN	M RU+EN
3	57.0	**65.9**	**71.8**	70.7
5	58.4	**70.3**	**75.0**	74.2
10	**75.7**	75.2	**77.9**	77.4
15	**77.7**	77.2	**79.7**	78.9
20	78.4	**79.0**	**80.6**	80.1
25	79.5	**79.6**	**81.4**	80.9
50	**82.5**	82.3	**83.2**	82.8
100	**84.4**	84.3	**85.2**	84.4

Figure 1: Average accuracy on the Russian data for *distilbert-base-multilingual-cased*, batch size 160, plain sampling. 'S' stands for single-task mode, 'M' stands for multi-task mode, 'RU share' means the share of Russian training data, 'RU' means training only on the given percentage of Russian data, 'RU+EN' means training on the given percentage of Russian data with added full size English data. See Table 9 for more details.

For the Russian-only data, starting with a small enough percentage of the training data, the single-task metrics drop and become much lower than the multitask metrics. We do not see this effect for the Russian+English data, as in this case, even with a low share of Russian data, even single-task models still learn a much higher amount of knowledge from the English data.

7 Discussion

Multi-task encoder-agnostic transformer-based models almost match the single-task models by metrics on the dialogue tasks. The gap in average accuracy between the multi-task and single-task monolingual models is about 0.8-0.9% for the English language and about 0.3-0.6% for the Russian language. For the

multilingual models, the gap remains within the same limit, except for the *bert-base-multilingual-cased* trained only on Russian data, for which there is no gap.

We also show that if we train the multilingual model and have Russian and English data for the same tasks with the same classes, combining that into one task is slightly better than treating Russian and English tasks as separate tasks. We can explain it by the fact that while training multilingual models on merged data (see Table 3), the knowledge is transferred by the backbone and by the class-specific linear layers. At the same time, while training multilingual models on separate data, only knowledge transfer by the backbone takes place.

For the small-scale data, we can see that if we train the multilingual distilbert on small shares of Russian training data (2-5%), multi-task models outperform single-task models in the average accuracy. The Talbe 9 shows that this accuracy advantage increases while the dataset size decreases. For intent and topic datasets, this advantage disappears at 1,151 training samples. For the emotion dataset, surprisingly, this advantage holds with any dataset partition, possibly due to the effect of knowledge transfer from the sentiment task.

For experiments with adding English data, multi-task models showed no clear pattern of advantage over single-task ones. This fact also supports the hypothesis of the knowledge transfer dependency of the dataset size. If we added 100% of English training data, dataset sizes became too large for reaching the advantage from the multi-task knowledge transfer.

However, adding the English training data to the Russian training data improves the metrics on the Russian test set. The lower the size of the Russian training data we have, the more substantial the accuracy increase from adding the English data to the training sample. This accuracy gain can reach several percent if we have a limited amount of Russian training data (3-10% RU share in Table 9). This conclusion holds for multi-task and single-task models. The language of validation data (English or Russian) did not matter in our experiments.

The reason for metric improvement for the multilingual models by adding the English data is that while being pretrained on certain languages (in our cases - English and Russian), the models learn to represent the language-independent features of the examples. Therefore, while receiving Russian and English examples for the same tasks, the models fine-tune to the larger number of language-independent features and generalize more broadly, which helps to improve the results.

Our work did not cover the knowledge transfer to languages other than Russian. Also, we did not consider conditions under which multilingual models, with knowledge transferred from English data, excel analogous Russian-only models. We leave that for future work.

8 Conclusion

We explore the knowledge transfer in the simple multi-task encoder-agnostic transformer-based models on five dialog tasks: emotion classification, sentiment classification, toxicity classification, intent classification, and topic classification. We show that these models' accuracy differs from the analogous single-task models by $\sim$0.9%. These results hold for the multiple transformer backbones. At the same time, these models have the same backbone for all tasks, which allows them to have about 0.1% more parameters than any analogous single-task model and to support multiple tasks simultaneously. We also found that if we decrease the dataset size to a certain extent, multi-task models outperform single-task ones, especially on the smallest datasets. We also show that while training multilingual models on the Russian data, adding the English data from the same task to the training sample can improve model performance for the multi-task and single-task settings - up to 4-5% if the Russian data are scarce enough. We also have integrated these models into the DeepPavlov framework and into the DREAM library.

References

Alaaeldin Ali, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, and Herve Jegou. 2021. Xcit: Cross-covariance image transformers. // M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, *Advances in Neural Information Processing Systems*, volume 34, P 20014–20027. Curran Associates, Inc.

Iain Murray Asa Cooper Stickland. 2019. Bert and pals: Projected attention layers for efficient adaptation in multi-task learning. // *Proceedings of the 36th International Conference on Machine Learning*, volume 97, P 5986:5995.

Dilyara Baymurzina, Denis Kuznetsov, Dmitry Evseev, Dmitry Karpov, Alsu Sagirova, Anton Peganov, Fedor Ignatov, Elena Ermakova, Daniil Cherniavskii, Sergey Kumeyko, et al. 2021. Dream technical report for the alexa prize 4. *4th Proc. Alexa Prize*.

Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yuri Kuratov, Denis Kuznetsov, and Vasily Konovalov. 2018. Deeppavlov: An open source library for conversational ai. // *NIPS*.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

Anastasia Chizhikova, Vasily Konovalov, and Mikhail Burtsev. 2023. Multilingual case-insensitive named entity recognition. // Boris Kryzhanovsky, Witali Dunin-Barkowski, Vladimir Redko, and Yury Tiumentsev, *Advances in Neural Computation, Machine Learning, and Cognitive Research VI*, P 448–454, Cham. Springer International Publishing.

Daryna Dementieva, Daniil Moskovskiy, Varvara Logacheva, David Dale, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Russian toxicity dataset from 2ch.hk. dataset retrieved from `https://github.com/s-nlp/rudetoxifier`. *CoRR*, abs/2105.09052.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan S. Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *CoRR*, abs/2005.00547.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. // *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, P 4171:4186.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification(paper for wiki talk dataset, cleaned version of the dataset retrieved from `https://huggingface.co/datasets/OxAISH-AL-LLM/wiki_toxic`). // *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, P 67–73, New York, NY, USA. Association for Computing Machinery.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2022. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages.

Varun Godbole, George E. Dahl, Justin Gilmer, Christopher J. Shallue, and Zachary Nado†. 2023. Tuning neural networks (google research github). `https://github.com/google-research/tuning_playbook`.

Zhiqi Huang, Lu Hou, Lifeng Shang, Xin Jiang, Xiao Chen, and Qun Liu. 2021. Ghostbert: Generate more features with cheap operations for bert. // *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, P 6512–6523, 01.

Dmitry Karpov and Michail Burtsev. 2021. Data pseudo-labeling while adapting bert for multitask approaches. // *Proceedings of the International Conference "Dialogue 2021"*.

Dmitry Karpov and Mikhail Burtsev. 2023. Monolingual and cross-lingual knowledge transfer for topic classification.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. // Yoshua Bengio and Yann LeCun, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Alina Kolesnikova, Yuri Kuratov, Vasily Konovalov, and Mikhail Burtsev. 2022. Knowledge distillation of russian language models with reduction of vocabulary.

VP Konovalov and ZB Tumunbayarova. 2018. Learning word embeddings for low resource languages: the case of buryat. // *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, P 331–341.

Vasily Konovalov, Ron Artstein, Oren Melamud, and Ido Dagan. 2016a. The negochat corpus of human-agent negotiation dialogues. // *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, P 3141–3145, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Vasily Konovalov, Oren Melamud, Ron Artstein, and Ido Dagan. 2016b. Collecting Better Training Data using Biased Agent Policies in Negotiation Dialogues. // *Proceedings of WOCHAT, the Second Workshop on Chatbots and Conversational Agent Technologies*, Los Angeles, September. Zerotype.

Vasily Konovalov, Pavel Gulyaev, Alexey Sorokin, Yury Kuratov, and Mikhail Burtsev. 2020. Exploring the bert cross-lingual transfer for reading comprehension. // *Komp'juternaja Lingvistika i Intellektual'nye Tehnologiithis*, P 445–453.

Yuri Kuratov and Mikhail Y. Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *CoRR*, abs/1905.07213.

Yuri Kuratov, Idris Yusupov, Dilyara Baymurzina, Denis Kuznetsov, Daniil Cherniavskii, Alexander Dmitrievskiy, Elena Ermakova, Fedor Ignatov, Dmitry Karpov, Daniel Kornev, et al. 2020. Dream technical report for the alexa prize 2019. *Alexa Prize Proceedings*.

Y M Kuratov, I F Yusupov, D R Baymurzina, D P Kuznetsov, D V Cherniavskii, A Dmitrievskiy, E S Ermakova, F S Ignatov, D A Karpov, D A Kornev, T A Le, P Y Pugin, and M S Burtsev. 2021. Socialbot dream in alexa prize challenge 2019. *Proceedings of Moscow Institute of Physics and Technology*, 13(3):62–89.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-Task Deep Neural Networks for Natural Language Understanding. // *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, P 4487:4496.

Lukasz Maziarka and Tomasz Danel. 2021. Multitask learning using BERT with task-embedded attention. // *2021 International Joint Conference on Neural Networks (IJCNN)*, P 1–6. IEEE, 7.

Christopher Potts, Zhengxuan Wu, Atticus Geiger, and Douwe Kiela. 2020. Dynasent: A dynamic benchmark for sentiment analysis. *CoRR*, abs/2012.15349.

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

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

Sergey Smetanin and Michail Komarov. 2019. Sentiment analysis of product reviews in russian using convolutional neural networks. // *2019 IEEE 21st Conference on Business Informatics (CBI)*, volume 01, P 482–486, July.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. // *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, P 353:355.

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. // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, P 38–45, Online, October. Association for Computational Linguistics.

9 Appendix

Table 4: Dataset sizes for the emotion classification task, Russian and English data.

data type	train	valid	test	train	valid	test
class	EN			RU		
Total	39555	4946	4968	6557	864	1862
joy	15216	1941	1863	1346	162	341
neutral	12823	1592	1606	2682	361	734
anger	4293	555	572	339	45	121
surprise	3858	459	488	491	77	165
sadness	2326	266	283	1207	158	368
fear	541	72	80	492	61	133
disgust	498	61	76	0	0	0

Table 5: Dataset sizes for the toxicity classification task, Russian and English data

data type	train	valid	test	train	valid	test
class	EN			RU		
Total	127656	31915	63978	93342	23010	46835
not_toxic	114722	28624	57735	75452	18669	37659
toxic	12934	3291	6243	17890	4341	9176

Table 6: Dataset sizes for the sentiment classification task, Russian and English data.

data type	train	valid	test	train	valid	test
class	EN			RU		
Total	80488	3600	3600	82610	3695	3695
positive	21391	1200	1200	27570	1220	1210
neutral	45076	1200	1200	27531	1234	1235
negative	14021	1200	1200	27509	1241	1250

Table 7: Dataset sizes for the topic classification task, Russian and English data.

data type	train	valid	test	train	valid	test
class	EN			RU		
Total	11514	2033	2974	11514	2033	2974
calendar	1688	280	402	1688	280	402
play	1377	260	387	1377	260	387
qa	1183	214	288	1183	214	288
email	953	157	271	953	157	271
iot	769	118	220	769	118	220
general	652	122	189	652	122	189
weather	573	126	156	573	126	156
transport	571	110	124	571	110	124
lists	539	112	142	539	112	142
news	503	82	124	503	82	124
recommendation	433	69	94	433	69	94
datetime	402	73	103	402	73	103
social	391	68	106	391	68	106
alarm	390	64	96	390	64	96
music	332	56	81	332	56	81
audio	290	35	62	290	35	62
takeaway	257	44	57	257	44	57
cooking	211	43	72	211	43	72

Table 8: Dataset sizes for the intent classification task, Russian and English data.

data type	train	valid	test	train	valid	test
class	EN			RU		
Total	11514	2033	2974	11514	2033	2974
calendar_set	810	131	209	810	131	209
play_music	639	123	176	639	123	176
weather_query	573	126	156	573	126	156
calendar_query	566	102	126	566	102	126
general_quirky	555	105	169	555	105	169
qa_factoid	544	90	141	544	90	141
news_query	503	82	124	503	82	124
email_query	418	73	119	418	73	119
email_sendemail	354	63	114	354	63	114
datetime_query	350	64	88	350	64	88
calendar_remove	312	47	67	312	47	67
play_radio	283	46	72	283	46	72
social_post	283	50	81	283	50	81
qa_definition	267	55	57	267	55	57
transport_query	227	36	51	227	36	51
cooking_recipe	207	41	72	207	41	72
lists_query	198	50	51	198	50	51
play_podcasts	193	34	63	193	34	63
recommendation_events	190	26	43	190	26	43
alarm_set	182	31	41	182	31	41
lists_createoradd	177	25	39	177	25	39
recommendation_locations	173	31	31	173	31	31
lists_remove	164	37	52	164	37	52
music_query	154	30	35	154	30	35
iot_hue_lightoff	153	17	43	153	17	43
qa_stock	152	24	26	152	24	26
play_audiobook	150	35	41	150	35	41
qa_currency	142	32	39	142	32	39
takeaway_order	135	20	22	135	20	22
alarm_query	130	19	34	130	19	34
transport_ticket	127	25	35	127	25	35
email_querycontact	127	16	26	127	16	26
iot_hue_lightchange	125	22	36	125	22	36
iot_coffee	124	14	36	124	14	36
takeaway_query	122	24	35	122	24	35
transport_traffic	117	22	15	117	22	15
music_likeness	113	16	36	113	16	36
play_game	112	22	35	112	22	35
audio_volume_mute	110	15	32	110	15	32
audio_volume_up	110	12	13	110	12	13
social_query	108	18	25	108	18	25
transport_taxi	100	27	23	100	27	23
iot_cleaning	93	19	26	93	19	26
alarm_remove	78	14	21	78	14	21
qa_maths	78	13	25	78	13	25
iot_hue_lightdim	76	17	21	76	17	21
iot_hue_lightup	76	12	27	76	12	27
general_joke	72	15	19	72	15	19
recommendation_movies	70	12	20	70	12	20
email_addcontact	54	5	12	54	5	12
datetime_convert	52	9	15	52	9	15
iot_wemo_off	52	5	18	52	5	18
audio_volume_down	52	8	11	52	8	11
music_settings	51	8	6	51	8	6
iot_wemo_on	48	7	10	48	7	10
general_greet	25	2	1	25	2	1
iot_hue_lighton	22	5	3	22	5	3
audio_volume_other	18	0	6	18	0	6
music_dislikeness	14	2	4	14	2	4
cooking_query	4	2	0	4	2	0

Mode	RU share	EN share	Validated on	Average	Emotions	Sentiment	Toxic	Intents	Topics	Batches seen
S	100%	100%	EN	85.2/82.1	79.0/70.8	77.2/77.4	96.5/94.5	84.5/80.6	88.4/87.4	15946
M	100%	100%	EN	84.4/80.9	77.2/70.5	75.8/75.8	96.4/94.4	83.5/76.3	88.9/87.8	20737
S	50%	100%	EN	83.2/79.5	75.6/65.8	75.6/75.7	96.1/93.9	82.2/76.5	86.8/85.5	16672
M	50%	100%	EN	82.8/78.1	76.2/64.5	74.0/73.4	95.9/93.5	80.9/72.7	87.2/86.1	19336
S	25%	100%	EN	81.4/76.7	73.7/61.4	73.7/73.9	95.5/92.7	78.8/71.9	85.1/83.6	16589
M	25%	100%	EN	80.9/76.4	73.1/63.9	73.7/73.7	95.1/92.2	77.5/68.1	85.3/83.9	16665
S	20%	100%	EN	80.6/76.0	71.8/60.3	74.0/74.0	95.1/92.1	78.0/71.1	83.9/82.4	12951
M	20%	100%	EN	80.1/75.0	71.9/61.2	73.5/73.5	94.9/91.9	76.1/65.5	84.2/82.8	17429
S	15%	100%	EN	79.7/74.7	70.8/57.8	72.6/72.7	94.6/91.3	77.3/70.1	83.1/81.6	13037
M	15%	100%	EN	78.9/73.5	70.0/58.4	71.9/71.5	94.5/91.2	74.7/65.0	83.5/81.8	15599
S	10%	100%	EN	77.9/72.0	68.3/52.1	72.3/72.7	93.9/90.0	73.9/65.8	81.2/79.4	13545
M	10%	100%	EN	77.4/70.9	67.9/51.2	71.7/71.7	93.7/90.1	72.3/61.5	81.6/79.9	14471
S	5%	100%	EN	75.0/67.9	64.1/45.0	70.2/70.4	92.7/87.8	69.9/60.5	77.9/75.8	12567
M	5%	100%	EN	74.2/66.3	63.4/41.2	70.1/70.2	92.3/87.6	67.6/56.6	77.7/75.9	12779
S	3%	100%	EN	71.8/64.6	59.1/38.8	68.4/68.6	91.0/85.6	65.9/57.4	74.6/72.6	12065
M	3%	100%	EN	70.7/64.2	58.5/44.8	67.8/67.7	90.9/85.5	62.4/51.3	74.0/71.6	14896
S	0%	100%	EN	52.4/42.0	48.3/26.6	43.8/43.1	80.0/58.6	37.5/31.5	52.3/50.4	15469
M	0%	100%	EN	51.0/41.5	42.6/23.8	45.4/42.8	78.6/61.6	38.0/30.6	50.0/48.4	14000
S	100%	0%	RU	84.4/80.4	76.5/66.5	77.2/77.3	96.7/94.7	83.5/76.4	88.2/87.0	11199
M	100%	0%	RU	84.3/80.4	77.9/70.4	76.4/76.5	96.5/94.4	82.3/73.3	88.4/87.4	11956
S	50%	0%	RU	82.5/78.0	74.0/63.2	76.4/76.4	96.1/93.8	80.0/71.9	86.1/84.8	5878
M	50%	0%	RU	82.3/78.0	75.0/66.5	74.6/74.7	96.0/93.7	79.5/69.8	86.4/85.2	8090
S	25%	0%	RU	79.5/72.5	67.0/45.0	75.1/75.4	95.4/92.5	76.6/68.1	83.6/81.5	3496
M	25%	0%	RU	79.6/74.3	72.3/62.1	72.7/72.8	95.3/92.6	73.8/61.5	83.7/82.1	5830
S	20%	0%	RU	78.4/70.3	64.3/36.2	74.4/74.7	95.1/92.0	75.3/67.5	82.9/81.0	2796
M	20%	0%	RU	79.0/74.2	71.4/61.4	73.0/73.2	95.0/91.9	73.5/63.8	82.3/80.8	5773
S	15%	0%	RU	77.7/70.1	66.1/44.5	74.0/74.0	94.8/91.5	72.2/61.2	81.6/79.5	1997
M	15%	0%	RU	77.2/71.3	70.7/59.6	71.7/72.0	94.6/91.4	68.6/54.9	80.6/78.7	5320
S	10%	0%	RU	75.7/67.1	64.5/41.1	73.3/73.5	93.9/90.0	67.7/54.7	78.8/76.2	1469
M	10%	0%	RU	75.2/68.2	68.7/55.3	71.5/71.7	94.0/90.2	64.0/48.4	77.8/75.5	2836
S	5%	0%	RU	58.4/47.9	48.3/20.3	71.0/71.1	92.7/87.9	29.9/18.2	50.1/41.8	739
M	5%	0%	RU	70.3/61.6	64.8/48.3	70.1/70.3	92.6/88.0	53.0/35.0	71.2/66.3	2095
S	3%	0%	RU	57.0/45.2	49.1/20.5	69.5/69.6	91.5/85.8	38.9/24.7	36.2/25.6	521
M	3%	0%	RU	65.9/55.1	62.6/41.3	69.0/69.2	91.2/85.6	42.6/24.2	63.9/55.1	1132

Table 9: Impact of small-scale training and adding parts of Russian data to the English data. Accuracy / F1-macro on the Russian data for *distilbert-base-multilingual-cased*, batch size 160, plain sampling. Mode S stands for singletask, mode M stands for multitask, RU share is the share of samples from every train Russian dataset, and EN share is the share of samples from every train English dataset. Averaged by 3-5 runs.

Table 10: Accuracy / f1 macro on the Russian data for the transformer-agnostic *distilbert-base-multilingual-cased*, batch size 160, plain sampling. Mode S stands for singletask, mode M stands for multitask, Impact of small-scale training and adding parts of Russian data to the English data. RU share is the share of samples from every train Russian dataset, and EN share is the share of samples from every train English dataset. Averaged by three runs. Comparison of validation on Russian and English data.

Mode	RU share	EN share	Validate on	Average	Emotions	Sentiment	Toxic	Intents	Topics	Batches seen
S	100%	100%	EN	85.2/82.1	79.0/70.8	77.2/77.4	96.5/94.5	84.5/80.6	88.4/87.4	15946
S	100%	100%	RU	85.3/81.9	79.2/71.4	77.2/77.3	96.7/94.7	84.6/78.2	88.6/87.7	29204
M	100%	100%	EN	84.4/80.9	77.2/70.5	75.8/75.8	96.4/94.4	83.5/76.3	88.9/87.8	20737
M	100%	100%	RU	84.4/80.7	77.6/69.8	76.8/76.9	96.6/94.6	82.4/74.5	88.8/87.8	21726
S	50%	100%	EN	83.2/79.5	75.6/65.8	75.6/75.7	96.1/93.9	82.2/76.5	86.8/85.5	16672
S	50%	100%	RU	83.5/79.6	76.7/67.6	76.1/76.2	96.2/93.9	81.7/74.9	86.7/85.4	17882
M	50%	100%	EN	82.8/78.1	76.2/64.5	74.0/73.4	95.9/93.5	80.9/72.7	87.2/86.1	19336
M	50%	100%	RU	82.7/78.6	75.5/66.3	74.5/74.7	96.0/93.6	80.7/72.8	86.8/85.8	23203
S	25%	100%	EN	81.4/76.7	73.7/61.4	73.7/73.9	95.5/92.7	78.8/71.9	85.1/83.6	16589
S	25%	100%	RU	81.8/77.3	74.5/63.4	74.6/74.9	95.4/92.6	79.1/71.7	85.1/83.7	15304
M	25%	100%	EN	80.9/76.4	73.1/63.9	73.7/73.7	95.1/92.2	77.5/68.1	85.3/83.9	16665
M	25%	100%	RU	81.0/76.6	73.3/63.8	73.5/73.8	95.0/92.2	78.1/69.5	85.1/83.9	19329
S	20%	100%	EN	80.6/76.0	71.8/60.3	74.0/74.0	95.1/92.1	78.0/71.1	83.9/82.4	12951
S	20%	100%	RU	81.0/76.3	73.2/62.3	74.5/74.6	95.2/92.2	77.6/69.6	84.4/83.0	15798
M	20%	100%	EN	80.1/75.0	71.9/61.2	73.5/73.5	94.9/91.9	76.1/65.5	84.2/82.8	17429
M	20%	100%	RU	80.3/75.3	72.3/61.5	73.9/74.1	94.9/92.0	76.1/66.1	84.5/83.1	14847
S	15%	100%	EN	79.7/74.7	70.8/57.8	72.6/72.7	94.6/91.3	77.3/70.1	83.1/81.6	13037
S	15%	100%	RU	80.0/75.4	71.5/60.5	73.7/73.9	94.8/91.6	76.6/69.2	83.3/81.7	18014
M	15%	100%	EN	78.9/73.5	70.0/58.4	71.9/71.5	94.5/91.2	74.7/65.0	83.5/81.8	15599
M	15%	100%	RU	79.0/73.1	69.8/54.1	71.7/71.5	94.5/91.3	75.5/66.6	83.5/82.1	17471
S	10%	100%	EN	77.9/72.0	68.3/52.1	72.3/72.7	93.9/90.0	73.9/65.8	81.2/79.4	13545
S	10%	100%	RU	78.2/72.0	69.4/50.5	72.1/72.4	94.3/90.6	74.4/66.8	81.0/79.5	17812
M	10%	100%	EN	77.4/70.9	67.9/51.2	71.7/71.7	93.7/90.1	72.3/61.5	81.6/79.9	14471
M	10%	100%	RU	77.3/71.2	67.8/54.3	72.0/72.1	93.4/89.7	71.4/59.7	81.7/79.9	13267
S	5%	100%	EN	75.0/67.9	64.1/45.0	70.2/70.4	92.7/87.8	69.9/60.5	77.9/75.8	12567
S	5%	100%	RU	75.2/68.7	64.8/47.9	70.5/70.7	93.0/88.4	69.5/59.9	78.1/76.4	16024
M	5%	100%	EN	74.2/66.3	63.4/41.2	70.1/70.2	92.3/87.6	67.6/56.6	77.7/75.9	12779
M	5%	100%	RU	73.6/66.3	61.3/44.7	70.1/70.1	92.4/87.6	66.1/52.8	78.0/76.1	11618
S	3%	100%	EN	71.8/64.6	59.1/38.8	68.4/68.6	91.0/85.6	65.9/57.4	74.6/72.6	12065
S	3%	100%	RU	72.1/64.8	59.9/40.1	68.7/69.0	91.8/86.5	65.5/55.9	74.6/72.5	12298
M	3%	100%	EN	70.7/64.2	58.5/44.8	67.8/67.7	90.9/85.5	62.4/51.3	74.0/71.6	14896
M	3%	100%	RU	70.7/63.0	58.7/39.5	67.8/67.2	90.6/85.2	62.1/50.8	74.2/72.1	14323