RuSentiTweet: a sentiment analysis dataset of general domain tweets in Russian

Sergey Smetanin

Department of Business Informatics, Graduate School of Business, National Research University Higher School of Economics, Russia

ABSTRACT

The Russian language is still not as well-resourced as English, especially in the field of sentiment analysis of Twitter content. Though several sentiment analysis datasets of tweets in Russia exist, they all are either automatically annotated or manually annotated by one annotator. Thus, there is no inter-annotator agreement, or annotation may be focused on a specific domain. In this article, we present RuSentiTweet, a new sentiment analysis dataset of general domain tweets in Russian. RuSentiTweet is currently the largest in its class for Russian, with 13,392 tweets manually annotated with moderate inter-rater agreement into five classes: Positive, Neutral, Negative, Speech Act, and Skip. As a source of data, we used Twitter Stream Grab, a historical collection of tweets obtained from the general Twitter API stream, which provides a 1% sample of the public tweets. Additionally, we released a RuBERT-based sentiment classification model that achieved $F_1 = 0.6594$ on the test subset.

INTRODUCTION

Recently, Twitter has been established as a major research platform, utilized in more than ten thousand research articles over the past ten years. Sentiment analysis has proven to be one of the major research areas (*Antonakaki, Fragopoulou & Ioannidis, 2021*). As expected, there is an interest in the sentiment analysis of the Russian-speaking segment of Twitter, not only for training machine learning (ML) models (*Kotelnikova, 2020*; *Araslanov, Komotskiy & Agbozo, 2020*; *Kanev et al., 2022*), but also for applied research—such as studying migration issues (*Borodkina & Sibirev, 2019*), measuring reactions to different events (*Kirilenko & Stepchenkova, 2017*; *Kausar, Soosaimanickam & Nasar, 2021*), and monitoring public sentiment (*Chizhik, 2016*; *Smetanin, 2017*). However, despite the fact that there are several datasets of tweets in Russian (*Smetanin, 2020a*), they are either annotated automatically (*e.g.*, RuTweetCorp by *Rubtsova (2013)*) or annotated only by one annotator; thus, there is no inter-annotator agreement (*e.g.*, Twitter Sentiment for 15 European Languages by *Mozeticar & Smailović (2016)*), or focused on a specific domain (*e.g.*, SentiRuEval-2015 by *Loukachevitch et al. (2015)*). Thus, this research community lacks a general domain sentiment dataset of tweets in Russian that is annotated manually with reported inter-rater agreement score.

Corresponding author
Sergey Smetanin,
sismetanin@gmail.com

In this article, we present RuSentiTweet, a new sentiment analysis dataset of 13,392 general domain tweets in Russian. RuSentiTweet was annotated manually using RuSentiment guidelines (*Rogers et al., 2018*) into five classes (*Positive*, *Neutral*, *Negative*, *Speech Act*, and *Skip*) with moderate inter-rater agreement. The practical and academic contribution of this study is threefold. Firstly, we reviewed existing public sentiment dataset of tweets in Russian. Secondly, we filled the data gap and introduced RuSentiTweet, the only dataset of general domain tweets with manual annotation for the Russian language. Lastly, we trained several ML models to provide further research with a strong baseline.

The rest of the article is organized as follows. In "Related Work", we review related research, identify existing public sentiment datasets of tweets in Russian, and confirm the importance of a new dataset of general domain tweets in Russian. In "Sentiment Dataset", we describe the creation of RuSentiTweet. In "Sentiment Classification Baseline", we document the training of several ML models to provide the research community with public baselines. In "Conclusion", we present conclusions from this study.

RELATED WORK

As of 2022, Russian was the eighth most widely-spoken language worldwide, with a total number of 258.2 million speakers (*Szmigiera, 2022*). Yet as reported in the preliminary results of the All-Russian Census 2020 (*Rosstat, 2022*), only about 147 million people permanently live in Russia. In addition to Russia, where Russian is the official language, it is also widely spoken in a number of other countries that were part of the USSR. According to various sources (*Arefiev, 2013*; *Lopatin & Ulukhanov, 2017*), there are from 52 to 94 million native speakers of the Russian language in these countries. A large number of Russian speakers also live in other countries such as those in Europe, the USA, Canada, Israel, and others (*Lopatin & Ulukhanov, 2017*). Given the significant Russian-speaking population and the ever-growing level of Internet penetration, texts published by Russian-speaking users on social networks are attracting more and more attention from researchers. As a result, every year new works appear both in the classical analysis of the sentiment of Russian-language content (*e.g.*, *Araslanov, Komotskiy & Agbozo, 2020*; *Kanev et al., 2022*; *Kausar, Soosaimanickam & Nasar, 2021*) and in related areas, such as the identification of emotions (*e.g.*, *Babii, Kazyulina & Malafeev, 2020*; *Kazyulina, Babii & Malafeev, 2020*; *Babii, Kazyulina & Malafeev, 2021*), toxicity and hate speech detection (*e.g.*, *Zueva, Kabirova & Kalaidin, 2020*; *Pronoza et al., 2021*; *Smetanin & Komarov, 2021b*), and inappropriate language identification (*e.g.*, *Babakov et al., 2021*; *Babakov, Logacheva & Panchenko, 2022*).

However, the Russian language is not as well-resourced as the English language (*Besacier et al., 2014*), especially in the field of sentiment analysis (*Smetanin & Komarov, 2021a*), so the data options for researchers are quite limited. In our previous study (*Smetanin, 2020a*), we identified 14 publicly available sentiment analysis datasets of Russian texts. In said study, we considered only those datasets that can be accessed *via* instructions from their original papers or official websites. Following this strategy, we omitted several existing datasets—such as ROMIP datasets (*Chetviorkin, Braslavskiy &*

Table 1 Sentiment analysis datasets of Russian language texts. More detailed description of each dataset can be found in *Smetanin (2020a)*, *Smetanin & Komarov (2021a)*, *Kotelnikov (2021)*, as well as in original papers (if published). For datasets that contain several subsets from different data sources, we indicated only those subsets that are made from tweets.

Dataset	Data source	Domain	Annotation	Classes	Size	Link
Twitter Sentiment for 15 European Languages (*Mozeticar & Smailović, 2016*)	Twitter	General	Manual	3	107,773	Project page
SemEval-2016 Task 5 Russian (*Pontiki et al., 2016*)	Twitter	Restaurants	Manual	3	405	Project page
SentiRuEval-2016 (*Lukashevich & Rubtsova, 2016*)	Twitter	Telecom and banks	Manual	3	23,595	Project page
SentiRuEval-2015 (*Loukachevitch et al., 2015*)	Twitter	Telecom and banks	Manual	4	16,318	Project page
RuTweetCorp (*Rubtsova, 2013*)	Twitter	General	Automatic	3	334,836	Project page
Kaggle Russian_twitter_sentiment	Twitter	n/a	n/a	2	226,832	Kaggle page

Loukachevich, 2013; *Chetvirokin & Loukachevitch, 2013*)—because we were unable to obtain access to them. Among these 14 datasets, only six datasets were constructed based on Twitter content, so we selected them for further detailed analysis. Additionally, we analysed the most recent review of sentiment analysis datasets of Russian texts by *Kotelnikov (2021)* but did not find any new Twitter datasets for consideration.

As can be seen from Table 1, RuTweetCorp (*Rubtsova, 2013*) is the largest sentiment analysis dataset of general domain tweets in Russian, but it was automatically annotated based on the strategy proposed by *Read (2005)*: each tweet was assigned with the sentiment class based on the emoticons it contains. As a consequence, even a simple rule-based approach based on the presence of the '(' character can achieve $F_1 = 97.39\%$ in the binary (*Positive* and *Negative* classes) classification task (*Smetanin & Komarov, 2021a*). SemEval-2016 Task 5 Russian (*Pontiki et al., 2016*), SentiRuEval-2016 (*Lukashevich & Rubtsova, 2016*) and SentiRuEval-2015 (*Loukachevitch et al., 2015*) are manually annotated and widely used datasets, but they are all tied to a specific domain such as restaurants, automobiles, telecommunication companies, or banks. Twitter Sentiment for 15 European Languages (*Mozeticar & Smailović, 2016*) is a sentiment analysis dataset with manual annotation, but only one annotator was engaged for Russian-language tweets; thus, there is no inter-annotator agreement. The Kaggle dataset did not report data collection and annotation procedure. Thus, there is a lack of general domain sentiment dataset of tweets in Russian that is annotated manually with reported inter-rater agreement score.

SENTIMENT DATASET

Data collection

For a data source of tweets in Russian, we decided to use the Twitter Stream Grab (https://archive.org/details/twitterstream), a publicly available historical collection of JSON grabbed from the general Twitter "Spritzer" API stream. According to Twitter, this API provides a 1% sample of the complete public tweets and is not tied to a specific topic, so we considered it as a good source of general domain tweets. Additionally, several studies (*Wang, Callan & Zheng, 2015*; *Leetaru, 2019*) performed independent validation of the representativeness of this stream. Since the Twitter Stream Grab consists of tweets in
different languages, our first step was to remove tweets written in non-Russian languages. Each tweet from this data source already contained information about the language of the text automatically detected[1] by Twitter, so the language filtering procedure was fairly straightforward.

We downloaded the Twitter Stream Grab for 12 months from January 2020 to December 2020[2]. The main motivation for choosing an entire year as the interval was to cover all months of the year to minimize the effect of seasonality. Previous research has shown that there are daily (*Larsen et al., 2015*; *Prata et al., 2016*), weekly (*Ten Thij, Bhulai & Kampstra, 2014*; *Dzogang, Lightman & Cristianini, 2017b*), and seasonal (*Dzogang et al., 2017a*) patterns of sentiment or emotion expression on Twitter. Also, it has been found (*Baylis et al., 2018*; *Baylis, 2020*) that expressed sentiment correlates with weather, which also tends to depend on the season. After excluding retweets and filtering by language, we obtained ~4.5M tweets in Russian. Since manual labelling of such a volume of tweets is costly and extremely time-consuming, we randomly selected 15,000 tweets for further annotation (tweets evenly distributed over the selected months).

Data annotation

Guidelines

As per recommendations outlined in our previous study (*Smetanin, 2020a*), we decided to use RuSentiment (*Rogers et al., 2018*) annotation guidelines (https://github.com/text-machine-lab/rusentiment/tree/master/Guidelines). To the best of our knowledge, this is the only set of publicly available sentiment annotation guidelines designed for the Russian language. The guidelines are described in detail in the original RuSentiment paper, so this section provides only key summary.

The annotation guidelines cover both implicit and explicit forms of expressions for external attitude (evaluation) and the internal emotional state (mood). The guidelines cover five sentiment classes.

- *Negative* represents both explicit and implicit negative sentiment or attitude towards something.
- *Neutral* represents texts that simply describe some situation in a neutral, matter-of-fact way and have no clear positive or negative sentiment. This class also includes commercial information, factual questions, objective descriptions, and summaries.
- *Positive* represents both explicit and implicit positive sentiment or attitude towards something.
- *Speech Act* represents texts that perform the functions of various speech acts—such as greeting someone, congratulating someone, and expressing gratitude for something. Although these texts also represent a positive sentiment, they are treated as a separate subcategory because they can also be performed under social pressure or out of a feeling of obligation (*Rogers et al., 2018*).
- *Skip* represents noisy and unclear sentiment or attitude towards something—such as when the original meaning is impossible to ascertain without additional context, the

---

[1] Assessing the quality of a given algorithm lies outside the scope of this study. Initial research in this direction has already been done in other studies; for example, *Pavliy and Lewis (2016)* compared the quality of Twitter's language detection algorithm and Google's Compact Language Detector on Ukrainian and Russian tweets. The authors found that Twitter's algorithm correctly detects 92% of texts in Russian and has higher accuracy than Google's Compact Language Detector.

[2] At the time of this writing, all months for 2021 were not available.
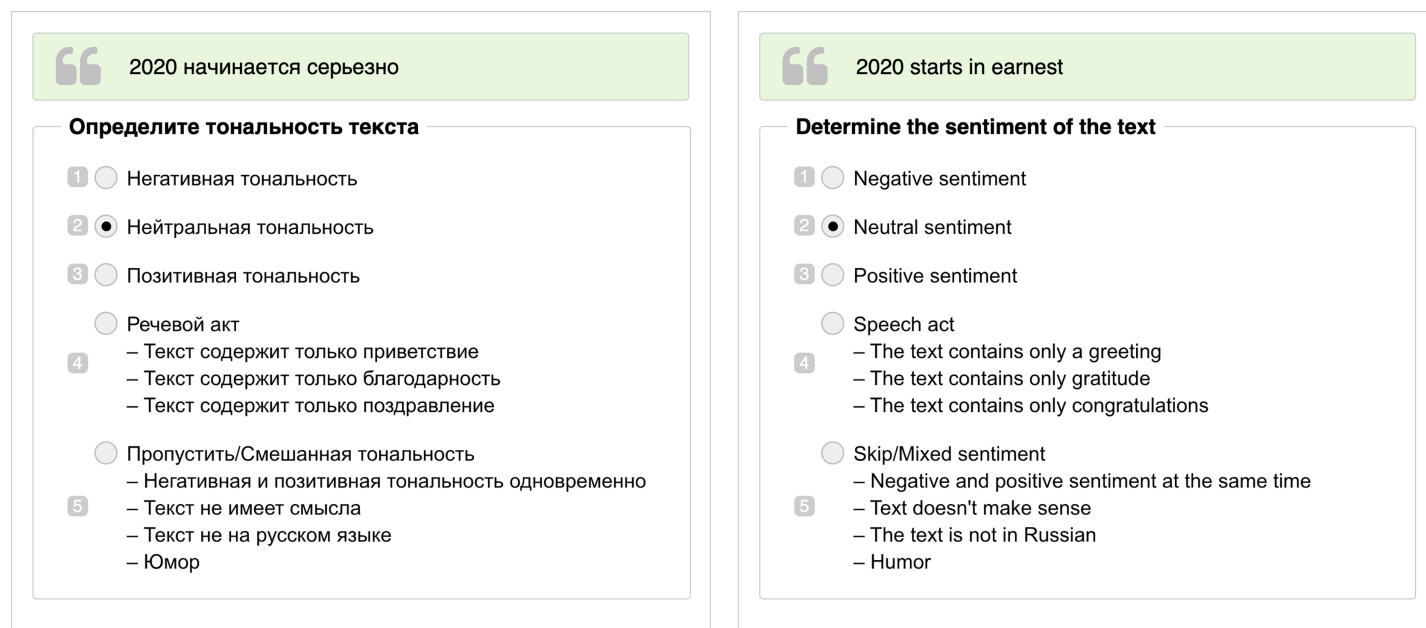
Figure 1 **An example of user interface for annotators in Yandex.Toloka in Russian (on the left) and its translation in English (on the right).** The green block with quotation marks contains the text of the tweet. Under the block with the text, there are numbered sentiment classes, where 1 is *Negative*, 2 is *Neutral*, 3 is *Positive*, 4 is *Speech Act*, and 5 is *Skip*. Numbers are used as hotkeys during annotation.

sentiment of the texts as a whole is not entirely clear, the text is not in Russian, or the text contains jokes.

Text with irony was annotated with the dominant sentiment, commonly negative. Hashtags were treated as information units similar to basic words or phrases. Emoticons were not treated as the only sentiment labels but were analysed in combination with the whole text to identify dominant sentiment.

Crowdsourcing platform

The annotation was performed *via* Yandex.Toloka (https://toloka.ai/), a Russian crowd-sourcing platform with a high share of Russian speaking workers. Yandex.Toloka is widely used in the studies on Russian-language content, such as for annotation of semantic change (*Rodina & Kutuzov, 2020*), question answering (*Korablinov & Braslavski, 2020*), and toxic comments (*Smetanin, 2020b*). A depiction of the Yandex.Toloka user interface can be found in Fig. 1. We required annotators to pass training before starting annotation. During the annotation of the dataset, their work was continuously evaluated through honeypots. As training samples and control pairs, we selected texts from RuSentiment, which was annotated using the same guidelines. The threshold was 60% correctly annotated samples for training and 80% samples for honeypots. We selected only Russian speaking annotators who passed an internal exam (https://toloka.ai/ru/docs/guide/concepts/filters.html) on language knowledge.

Aggregation

Following the RuSentiment aggregation strategy, a tweet was deemed to belong to a class if at least two out of the three annotators attributed it to that class. In case all three annotators disagreed, the tweet was removed from the dataset as extremely noisy and unclear (see examples in Table A1). Out of the initially selected 15,000 tweets, 1,608 tweets received all three different annotations, so we excluded these tweets from the final dataset. Thus, the final dataset consists of 13,392 tweets with the following class distribution: 3,298 (24.62%) *Negative* tweets, 5,341 (39.88%) *Neutral* tweets, 2,414 (18.02%) *Positive* tweets, 1,843 (13.76%) *Skip* tweets, and 496 (3.70%) *Speech Act* tweets. We split our dataset into training (80%) and test subset (20%) using stratified random sampling by class labels.

Inter-annotator agreement

For measuring the inter-annotator agreement, we calculated the Krippendorff's α coefficient (*Krippendorff, 1980*) because it applies to any number of annotators and categories, as well as to missing or incomplete data (*Krippendorff, 2004*). For most inter-annotator agreement indices, including Krippendorff's α, it is commonly suggested that a cutoff threshold value of 0.8 is a marker of good reliability, with a range of 0.667 to 0.8 allowing for tentative conclusions and values below 0.667 indicating poor agreement (*Beckler et al., 2018*). However, in the systematic review of crowd-sourced annotation in social computing, *Salminen et al. (2018)* reported that agreement scores in social computing studies are not high, averaging at around 0.60 for both Kappa and Alpha metrics, which is lower than typical threshold values. The authors highlighted that the nature of annotation in social computing tends to be more subjective rather than objective, and the more subjective the task, the worse the agreement, regardless of the metric used. Though it is important to report inter-rater agreement scores, there are suggestions that the results can be misleading in social computing (*Hillaire, 2021*). In fact, low agreement in this case does not necessarily mean the opinions of annotators are incorrect; it may simply indicate that they have different opinions (*Salminen et al., 2018*; *Hillaire et al., 2021*). Sentiment annotation, by nature, is a subjective task because the annotator must subjectively (with some guidelines) identify sentiment and emotions expressed by the author and not just objectively analyse narrated events or situations: we can expect annotators to have different subjective understanding of emotion expressed in a particular text. Thus, considering that in the field of social computing science the mean score is 0.60 (*Salminen et al., 2018*), we followed the same approach as *Hillaire (2021)* and adopted the less conservative interpretation of inter-rater agreement by *Landis & Koch (1977)*, which suggests the following interpretations.

- Scores from 0.0 to 0.2 indicate a slight agreement.
- Scores from 0.21 to 0.40 indicate a fair agreement.
- Scores from 0.41 to 0.60 indicate a moderate agreement.
- Scores from 0.61 to 0.80 indicate a substantial agreement.
- Scores from 0.81 to 1.0 indicate almost perfect or perfect agreement.

Table 2 Distance between classes for interval Krippendorff's α, where 0 means that classes are the same, 1 means that classes are close to each other, and 2 means that classes a far away from each other.

Class	Negative	Neutral	Positive	Speech	Skip
Negative	0	1	2	2	1
Neutral	1	0	1	1	1
Positive	2	1	0	0	1
Speech	2	1	0	0	1
Skip	1	1	1	1	0

Note:
Positive and *Speech* classes have zero distance between them; they both represent positive sentiment as per RuSentiment guidelines.

We calculated Krippendorff's α with binary distance (*e.g.*, all classes have similar distance between each other) using the NLTK library (*Bird, Klein & Loper, 2009*) and obtained the score of 0.5048 for binary distance, which can be interpreted as a moderate agreement between annotators. We considered this level of agreement as satisfactory for our case, since other five-class sentiment datasets also reported this or even lower level of agreement, such as Blog Track at TREC 2008 ($\alpha = 0.4219$, five classes) (*Bermingham & Smeaton, 2009*), LINIS Crowd ($\alpha = 0.541$, five classes) (*Koltsova, Alexeeva & Kolcov, 2016*), RuSentiment (Fleiss' kappa of 0.58, five classes) (*Rogers et al., 2018*), sentiment@USNavy ($\alpha = 0.592$, four classes) (*Fiok et al., 2021*), and NaijaSenti (Fleiss kappa of (0.434, 0.555), five classes) (*Muhammad et al., 2022*). Additionally, we calculated Krippendorff's α with interval distance that takes into account distance between classes: for example, *Neutral* and *Positive* classes are closer to each other than *Negative* and *Positive* classes. The distance matrix is presented in Table 2. The Krippendorff's α coefficient for interval distance was 0.5601, which can also be interpreted as slightly higher but still moderate agreement.

Explanatory analysis

The average text length is 59.36 characters for all text, 67.52 for *Negative*, 59.29 for *Neutral*, 57.85 for *Positive*, 42.71 for *Speech*, and 51.41 for *Skip*. As can be seen from Fig. 2, the frequency of occurrence of texts from a pair of characters in the dataset is extremely low, but with an increase in the number of characters, rapid growth begins. The frequency peak is reached when the text length is from 20 to 40 characters, and then the frequency gradually begins to decrease. Interestingly, for some classes, there is a moderate Pearson's correlation between the length of the text and the proportion of texts with this class relative to all texts. The *Negative* class has a moderate positive correlation ($\rho = 0.68$, $p < 0.01$) with the length of text, whereas *Speech* ($\rho = -0.52$, $p < 0.01$) and *Skip* ($\rho = -0.62$, $p < 0.01$) classes have moderate negative correlation. At the same time, *Neutral* ($\rho = -0.03$, $p = 0.70$) and *Positive* ($\rho = -0.04$, $p = 0.62$) classes do not have statistically significant correlation. The most common unigrams, bigrams, and emojis can be found in Table 3.
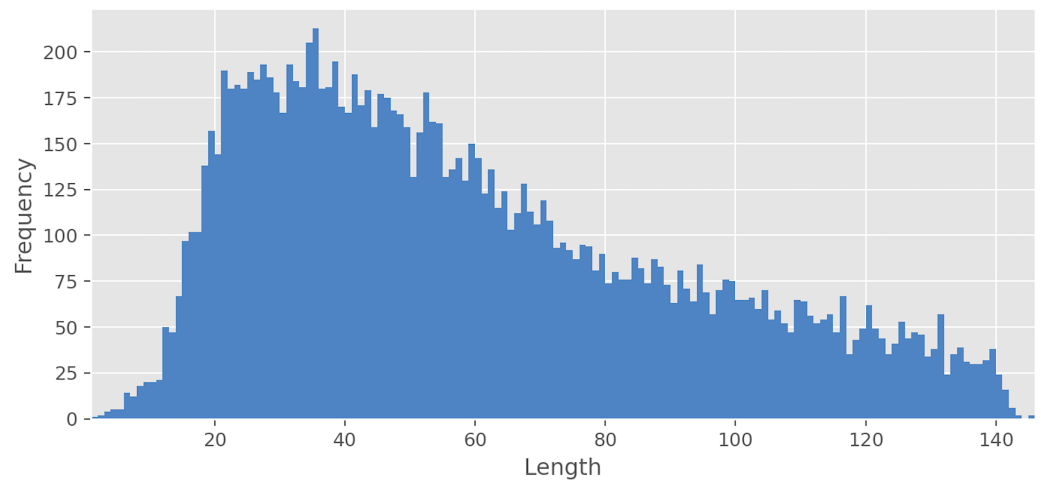
Figure 2 Texts length distribution.

Table 3 Most common unigrams, bigrams, and emojis without stop words, punctuation, and numbers. Stop words were removed using NLTK (*Bird, Klein & Loper, 2009*). Most unigrams and bigrams can have several English translations depending on the context. The table provides only one translation option.

Unigram			Bigram			Emoji	
Item		**Count**	**Item**		**Count**	**Item**	**Count**
Russian	**English**		**Russian**	**English**			
это	it	1,117	доброе утро	good morning	39		443
просто	simply	355	спокойной ночи	good night	26		313
спасибо	thanks	306	спасибо большое	thanks a lot	24		246
хочу	want	249	самом деле	actually	23		240
ещё	yet	223	это просто	it's simple	23		120
почему	why	209	опубликовано фото	published photo	18		119
очень	very	205	сих пор	so far	17		118
всё	all	204	руб г	rub g	16		113
блять	fuck	184	днем рождения	birthday	15		104
вообще	generally	174	все ещё	still	13		100

As mentioned in "Related Work", one of the key limitations of RuTweetCorp (*Rubtsova, 2013*)—the biggest automatically annotated dataset of tweets in Russian—is that *Positive* and *Negative* tweets in it can be easily separated with F_1 = 97.39% by a simple rule-based approach based on the presence of the '(' character. We decided to check that this limitation does not apply to RuSentiTweet. We applied this simple rule-based approach to *Positive* and *Negative* tweets from RuSentiTweet and got F_1 = 0.3450 (*i.e.*, approximately the same result as in the case of a random classification), thereby confirming that RuTweetCorp's limitation does not apply to RuSentiTweet.

Table 4 Five-class sentiment classification on RuSentiTweet.

Model	Precision	Recall	F_1^{macro}	$F_1^{weighted}$
RuBERT	0.6793	0.6449	0.6594	0.6675
MNB	0.5867	0.5021	0.5216	0.5189

SENTIMENT CLASSIFICATION BASELINE

Model selection

As was demonstrated in our recent study (*Smetanin & Komarov, 2021a*), sentiment analysis of the Russian language text based on the language models tends to outperform rule-based and basic ML-based approaches in terms of classification quality. This statement was also supported by other studies (*Golubev & Loukachevitch, 2020*; *Kotelnikova, 2020*; *Konstantinov, Moshkin & Yarushkina, 2021*). Based on the mentioned papers, we decided to fine-tune RuBERT (*Kuratov & Arkhipov, 2019*), a version of BERT (*Devlin et al., 2019*) trained on the Russian part of Wikipedia and Russian news. Over the past few years, this model has been actively used in sentiment analysis studies on the Russian language and constantly demonstrated strong or even new state-of-the-art (SOTA) results (*Golubev & Loukachevitch, 2020*; *Kotelnikova, 2020*; *Konstantinov, Moshkin & Yarushkina, 2021*; *Smetanin & Komarov, 2021a*). For comparison, we also decided to train more a classical ML classifier for sentiment analysis task: Multinomial Naive Bayes (MNB). We used the MNB implementation (https://github.com/sismetanin/sentiment-analysis-of-tweets-in-russian) from our previous paper (*Smetanin & Komarov, 2019*).

Results

During the training stage for RuBERT, we relied on the approach used in *Smetanin & Komarov (2021a)*. Fine-tuning was performed using the Transformers library (*Wolf et al., 2020*) on 1 Tesla V100 SXM2 32GB GPU with the following parameters: four train epochs, 128 max sequence length, 32 batch size, and a learning rate of 5e−5. Since our goal was to provide a baseline classification model and not the most efficient one, we did not search for the most efficient training parameters. We repeated each experiment 3 times and reported mean values of the measurements. For MNB, we used the same parameters as in our previous paper (*Smetanin & Komarov, 2019*): combination of unigrams and bigrams, TF-IDF vectorizer, and an alpha of 0.01.

According to the results presented in Table 4, RuBERT outperformed MNB, as expected, and showed the best classification scores. The classification results obtained on RuSentiTweet are slightly lower but still comparable with the results obtained in other studies on RuSentiment (see Table 5): RuBERT achieved $F_1^{weighted} = 0.7263$ on RuSentiment (*Kuratov & Arkhipov, 2019*), whereas on our dataset this model showed $F_1^{weighted} = 0.6675$. The difference in the results could be caused by the size of the dataset because RuSentiment is more than two times bigger. The classification metrics of five-class sentiment analysis approaches on other datasets in other languages can be found in

Table 5 Five-class sentiment classification studies.

Study	Dataset	Model	Classification metrics				
			Accuracy	Precision	Recall	F_1^{macro}	$F_1^{weighted}$
Muhammad et al. (2022)	NaijaSenti	XLM-R-base+LAFT	n/a	n/a	n/a	n/a	0.795
Muhammad et al. (2022)	NaijaSenti	M-BERT+LAFT	n/a	n/a	n/a	n/a	0.7700
Fiok et al. (2021)	sentiment@USNavy	BART large + CNN	n/a	n/a	n/a	0.596	n/a
Smetanin & Komarov (2021a)	RuSentiment	M-BERT-Base	n/a	0.6722	0.6907	0.6794	0.7244
Smetanin & Komarov (2021a)	RuSentiment	RuBERT	n/a	0.7089	0.7362	0.7203	0.7571
Smetanin & Komarov (2021a)	RuSentiment	M-USE-CNN	n/a	0.6571	0.6708	0.6627	0.7105
Smetanin & Komarov (2021a)	RuSentiment	M-USE-Trans	n/a	0.6821	0.6982	0.6860	0.7342
Jamadi Khiabani, Basiri & Rastegari (2020)	TripAdvisor	Dempster–Shafer-based model	0.79	0.5	0.47	0.49	n/a
Jamadi Khiabani, Basiri & Rastegari (2020)	CitySearch	Dempster–Shafer-based model	0.79	0.48	0.48	0.48	n/a
Kuratov & Arkhipov (2019)	RuSentiment	Multilingual BERT	n/a	n/a	n/a	n/a	0.7082
Kuratov & Arkhipov (2019)	RuSentiment	RuBERT	n/a	n/a	n/a	n/a	0.7263
Baymurzina, Kuznetsov & Burtsev (2019)	RuSentiment	SWCNN + fastText Twitter	n/a	n/a	n/a	n/a	0.7850
Baymurzina, Kuznetsov & Burtsev (2019)	RuSentiment	BiGRU + ELMo Wiki	n/a	n/a	n/a	n/a	0.6947
Tripto & Ali (2018)	YouTube	LSTM	0.5424	n/a	n/a	0.5320	n/a
Li et al. (2018)	Twitter	Logistic Regression	0.6899	0.6053	0.6899	0.6354	n/a
Ahmadi et al. (2017)	SST-5	RNTN	0.41	n/a	n/a	0.32	n/a
Buntoro, Adji & Purnamasari (2016)	Twitter	Naïve Bayes	0.7177	0.716	0.718	n/a	n/a
Aly & Atiya (2013)	LABR	SVM	0.503	n/a	n/a	n/a	0.491
Chetvirokin & Loukachevitch (2013)	ROMIP-2012 (Movies)	n/a	0.407	n/a	n/a	0.377	n/a
Blinov, Kotelnikov & Pestov (2013)	ROMIP-2012 (Books)	SVM	0.481	0.339	0.496	0.402	n/a
Chetvirokin & Loukachevitch (2013)	ROMIP-2012 (Cameras)	n/a	0.480	n/a	n/a	0.336	n/a
Pak & Paroubek (2012)	ROMIP-2011 (Movies)	SVM	0.599	n/a	n/a	0.286	n/a
Pak & Paroubek (2012)	ROMIP-2011 (Books)	SVM	0.622	n/a	n/a	0.291	n/a
Pak & Paroubek (2012)	ROMIP-2011 (Cameras)	SVM	0.626	n/a	n/a	0.342	n/a

Note:

We selected only those studies, which consideredfive sentiment classes and reported at least one of the following classification measures: Precision, Recall, macro *F1*, weighted *F1*. Among all datasets, only ROMIP (*Chetviorkin, Braslavskiy & Loukachevich, 2013*; *Chetvirokin & Loukachevitch, 2013*) and RuSentiment (*Rogers et al., 2018*) datasets are in Russian.

Table 5. Although direct comparison for different datasets and languages may not be entirely correct, we can see that at least the magnitude of order of our approach corresponds with the average score for five-class classification.

We made our RuBERT-based model publicly available (https://huggingface.co/sismetanin/rubert-rusentitweet) to the research community.

Error analysis

Considering that RuBERT clearly outperformed MNB, we performed error analysis only for RuBERT. As can be seen from confusion matrix for RuBERT (see Fig. 3), the *Skip* class was one of the most scarcely classified classes since it initially consisted of barely
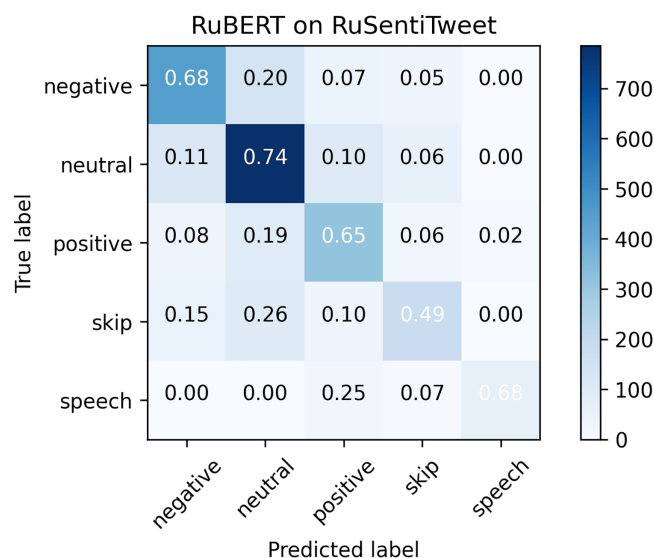
Figure 3 Confusion matrix for RuSentiTweet.
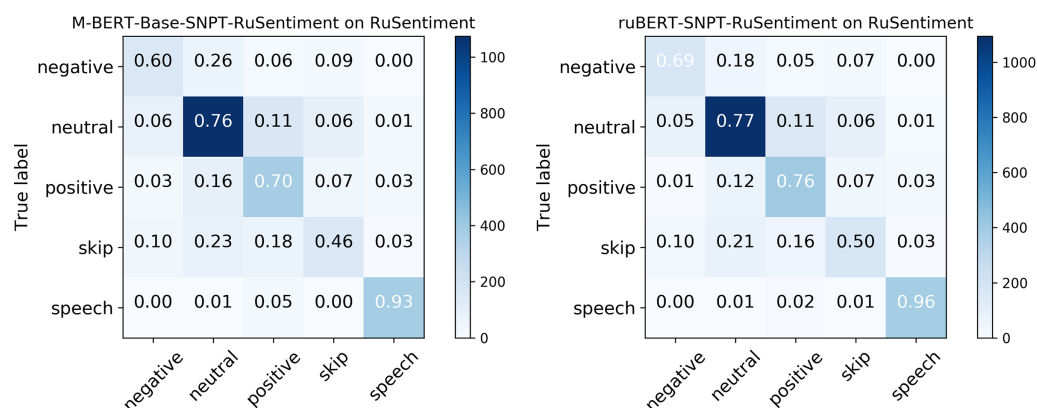
Figure 4 Confusion matrix for RuSentiment were created using molders from *Smetanin & Komarov (2021a)*.

interpretable and noisy tweets. The *Speech Acts* class was clearly distinguished from *Negative* and *Neutral* classes because it consists of a well-defined group of speech constructs, but it was commonly misclassified as *Positive* because it also represents positive sentiment. Predictably, the *Neutral* class was commonly misclassified as *Positive* or *Negative* class because neutral sentiment is logically located between positive and negative sentiment. As was highlighted by *Barnes, Øvrelid & Velldal (2019)*, the issue of neutral sentiment misclassification tends to be a general challenge of non-binary sentiment classification. In general, misclassification errors of our model were quite similar to RuSentiment misclassification errors reported in our previous study (*Smetanin & Komarov, 2021a*) (see Fig. 4), most likely because the same annotation guidelines and models were used. The most noticeable difference was in the recall for the *Speech* class. For RuSentiment, it was much better separated from other classes, with recall in the interval

from 0.88 to 0.96 (*Smetanin & Komarov, 2021a*). We suppose that the reason of such a difference is in the number of texts in this class: RuSentiment contains 3,467 texts of the *Speech*, whereas RuSentiTweet contains only 480 such texts. The examples of misclassified tweets can be found in Table A2.

CONCLUSION

In this article, we present RuSentiTweet, a new general domain sentiment dataset of tweets in Russian with manual annotation. RuSentiTweet includes 13,392 tweets annotated by three annotators with moderate inter-rater agreement into five classes: *Positive*, *Neutral*, *Negative*, *Speech Act*, and *Skip*. Currently, RuSentiTweet is the only dataset of general domain tweets in Russian with manual annotation by more than one annotator and is the largest in its class for Russian. Additionally, we presented a RuBERT-based model trained on RuSentiTweet, which demonstrated $F_1 = 0.6594$ in five-class classification. The code, data, and model were made publicly available to the research community.

Further research might focus on several areas. Firstly, considerably more work must be done to determine the most efficient ML algorithm in terms of classification quality for RuSentiTweet. In particular, it could be interesting to apply explainable sentiment analysis approaches (*e.g.*, *Szczepański et al., 2021*; *Kumar & Raman, 2022*) to allow a deeper understanding of the reasons for misclassification errors on particular texts. Secondly, it would be interesting to measure a subjective well-being index based on historical Russian tweets. Lastly, another possible area of future research would be to perform additional toxicity annotation of negative tweets from RuSentiTweet.

ACKNOWLEDGEMENTS

We thank the authors of the Twitter Stream Grab for collecting public tweets and making them available to the research community. We also thank the authors of RuSentiment for making their annotation guidelines publicly available.

This research was supported in part through computational resources of HPC facilities at HSE University (*Kostenetskiy, Chulkevich & Kozyrev, 2021*).

APPENDIX

Table A1 Examples of tweets with no agreement between annotators.

Tweet		Annotation		
Russian	**English**	**Annotator 1**	**Annotator 2**	**Annotator 3**
ю ноу блин	yu know damn it	Skip	Neutral	Negative
@USER доброе утро всем дэдди сегодня	@USER good morning daddies to everyone today	Speech	Positive	Skip
Путешествуем по Уэльсу. не уважать мужчин	Traveling in Wales. disrespect men	Negative	Neutral	Skip
тот факт что в энимал кроссинге так мало прикольных мышиных жителей	the fact that there are so few funny mouse inhabitants in animal crossing	Negative	Positive	Neutral
Кто в нижнем родился, в верхнем не сгодился.	Who was born in the bottom, did not fit in the top.	Negative	Neutral	Skip

Table A2 Examples of tweets classification. All usernames and URLs were replaced with keywords for anonymity purposes.

Tweet		True class	Predicted class
Russian	**English**		
@USERNAME @USERNAME @USERNAME @USERNAME @USERNAME Помедорус	@USERNAME @USERNAME @USERNAME @USERNAME @USERNAME Pomedorus	Skip	Skip
@USERNAME ты не лохушка ЛОЛ я тебе завидую…. у меня травма из за интернета вот я лохушка	@USERNAME you're not a sucker LOL I envy you…. I'm traumatized because of the internet I'm a sucker	Skip	Negative
@USERNAME Котиков Одриосолу Дождь	@USERNAME Cats Odriosolu Rain	Skip	Neutral
@USERNAME Уж лучше твоя грудь	@USERNAME Your breasts are better	Skip	Positive
Как сережки URL	How do you like the earrings URL	Neutral	Neutral
@USERNAME Реквием по мечте	@USERNAME Requiem for a dream	Neutral	Positive
@USERNAME ПОДОЖДИ НУ МНЕ КАЗАЛОСЬ ДА	@USERNAME WAIT I THINK YES	Neutral	Negative
@USERNAME Спокойной ночи и сладких снов	@USERNAME Good night and sweet dreams	Speech	Speech
@USERNAME как дела моя хорошая?? ( )	@USERNAME how are you my dear?? ( )	Speech	Positive
@USERNAME Это классно что у тебя есть эти люди	@USERNAME It's great that you have these people	Positive	Positive
На самом деле я ловлю уруру с этого облачка.	In fact, I catch ururu from this cloud.	Positive	Neutral
@USERNAME Это классно что у тебя есть эти люди	What kind of morons are you, many have a school/work day tomorrow	Negative	Negative
интересный факт: смысла в клипах тхт больше, чем в твоей жизни	Interesting fact: there is more sense in txt clips than in your life	Negative	Neutral

Funding
There was no external funding received for this study.

Competing Interests
The authors declare that they have no competing interests.

Author Contributions
- Sergey Smetanin conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

Data Availability
The code and dataset are available at GitHub: https://github.com/sismetanin/rusentitweet.

The sentiment classification model is available at HuggingFace: https://huggingface.co/sismetanin/rubert-rusentitweet.

Supplemental Information
Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.1039#supplemental-information.

REFERENCES

Ahmadi Z, Skowron M, Stier A, Kramer S. 2017. An in-depth experimental comparison of RNTNs and CNNs for sentence modeling. In: *International Conference on Discovery Science*. Springer, 144–152.

Aly M, Atiya A. 2013. LABR: a large scale Arabic book reviews dataset. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Short Papers)*. Vol. 2. Association for Computational Linguistics, 494–498.

Antonakaki D, Fragopoulou P, Ioannidis S. 2021. A survey of Twitter research: data model, graph structure, sentiment analysis and attacks. *Expert Systems with Applications* **164**:114006 DOI 10.1016/j.eswa.2020.114006.

Araslanov E, Komotskiy E, Agbozo E. 2020. Assessing the impact of text preprocessing in sentiment analysis of short social network messages in the Russian language. In: *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*. Piscataway: IEEE, 1–4.

Arefiev A. 2013. Demographic changes are not good for the Russian language. *Demoscope Weekly* 571–572.

Babakov N, Logacheva V, Kozlova O, Semenov N, Panchenko A. 2021. Detecting inappropriate messages on sensitive topics that could harm a company's reputation. In: *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*. Kiyv, Ukraine: Association for Computational Linguistics, 26–36.

Babakov N, Logacheva V, Panchenko A. 2022. Beyond plain toxic: detection of inappropriate statements on flammable topics for the Russian language. *Available at https://arxiv.org/abs/2203.02392*.

Babii A, Kazyulina M, Malafeev AY. 2020. Automatic emotion identification in Russian text messages. In: *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue" 2020*. Russian State University for the Humanities, 1002–1010.

Babii A, Kazyulina M, Malafeev A. 2021. FastText-based methods for emotion identification in Russian internet discourse. In: *WebSci '21: 13th ACM Web Science Conference 2021*. 112–119.

Barnes J, Øvrelid L, Velldal E. 2019. Sentiment analysis is not solved! assessing and probing sentiment classification. In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics, 12–23.

Baylis P. 2020. Temperature and temperament: evidence from Twitter. *Journal of Public Economics* **184**:104161 DOI 10.1016/j.jpubeco.2020.104161.

Baylis P, Obradovich N, Kryvasheyeu Y, Chen H, Coviello L, Moro E, Cebrian M, Fowler JH. 2018. Weather impacts expressed sentiment. *PLOS ONE* **13(4)**:e0195750 DOI 10.1371/journal.pone.0195750.

Baymurzina D, Kuznetsov D, Burtsev M. 2019. Language model embeddings improve sentiment analysis in Russian. In: *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2019"*. Vol. 18. 53–63.

Beckler DT, Thumser ZC, Schofield JS, Marasco PD. 2018. Reliability in evaluator-based tests: using simulation-constructed models to determine contextually relevant agreement thresholds. *BMC Medical Research Methodology* **18(1)**:1–12 DOI 10.1186/s12874-018-0606-7.

Bermingham A, Smeaton AF. 2009. A study of inter-annotator agreement for opinion retrieval. In: *SIGIR '09: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval.* 784–785.

Besacier L, Barnard E, Karpov A, Schultz T. 2014. Automatic speech recognition for under-resourced languages: a survey. *Speech Communication* **56**:85–100 DOI 10.1016/j.specom.2013.07.008.

Bird S, Klein E, Loper E. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* Newton: O'Reilly Media, Inc.

Blinov P, Kotelnikov E, Pestov O. 2013. Research of lexical approach and machine learning methods for sentiment analysis. In: *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue" 2013.* Vol. 2. Russian State University for the Humanities, 51–61.

Borodkina O, Sibirev V. 2019. Migration issues in Russian Twitter: attitudes to migrants, social problems and online resources. In: El Yacoubi S, Bagnoli F, Pacini G, eds. *Internet Science.* Cham: Springer International Publishing, 32–46.

Buntoro GA, Adji TB, Purnamasari AE. 2016. Sentiment analysis candidates of Indonesian Presiden 2014 with five class attribute. *International Journal of Computer Applications* **136(2)**:23–29 DOI 10.5120/ijca2016908288.

Chetviorkin I, Braslavskiy P, Loukachevich N. 2013. Sentiment analysis track at ROMIP 2011. In: *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue" 2012.* Vol. 2. Russian State University for the Humanities, 1–14.

Chetvirokin I, Loukachevitch N. 2013. Sentiment analysis track at ROMIP 2012. In: *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue" 2013.* 2.

Chizhik A. 2016. Factors for forming social mood on the basis of the analysis of the emotional coloring of posts in the Russian-language Twitter. In: *New Information Technologies in Automated Systems.* Vol. 19. HSE Tikhonov Moscow Institute of Electronics and Mathematics, 61–64.

Devlin J, Chang M-W, Lee K, Toutanova K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Long and Short Papers.* Minneapolis, Minnesota: Association for Computational Linguistics, 4171–4186.

Dzogang F, Goulding J, Lightman S, Cristianini N. 2017a. Seasonal variation in collective mood via Twitter content and medical purchases. In: *International Symposium on Intelligent Data Analysis.* Springer, 63–74.

Dzogang F, Lightman S, Cristianini N. 2017b. Circadian mood variations in Twitter content. *Brain and Neuroscience Advances* **1**:2398212817744501 DOI 10.1177/2398212817744501.

Fiok K, Karwowski W, Gutierrez E, Wilamowski M. 2021. Analysis of sentiment in tweets addressed to a single domain-specific Twitter account: comparison of model performance and explainability of predictions. *Expert Systems with Applications* **186**:115771 DOI 10.1016/j.eswa.2021.115771.

Golubev A, Loukachevitch N. 2020. Improving results on Russian sentiment datasets. In: Filchenkov A, Kauttonen J, Pivovarova L, eds. *Artificial Intelligence and Natural Language.* Cham: Springer International Publishing, 109–121.

Hillaire GE. 2021. *Understanding emotions in online learning: using emotional design and emotional measurement to unpack complex emotions during collaborative learning*. Milton Keynes, UK: Open University.

Hillaire G, Rienties B, Fenton-O'Creevy M, Zdrahal Z, Tempelaar D. 2021. Incorporating student opinion into opinion mining. In: *Open World Learning: Research, Innovation and the Challenges of High-Quality Education*. London: Routledge, 171–185 DOI 10.4324/9781003177098-15.

Jamadi Khiabani P, Basiri ME, Rastegari H. 2020. An improved evidence-based aggregation method for sentiment analysis. *Journal of Information Science* **46(3)**:340–360 DOI 10.1177/0165551519837187.

Kanev AI, Savchenko GA, Grishin IA, Vasiliev DA, Duma EM. 2022. Sentiment analysis of multilingual texts using machine learning methods. In: *2022 Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus)*. Piscataway: IEEE, 326–331.

Kausar MA, Soosaimanickam A, Nasar M. 2021. Public sentiment analysis on Twitter data during COVID-19 outbreak. *International Journal of Advanced Computer Science and Applications* **12(2)**:415–422 DOI 10.14569/issn.2156-5570.

Kazyulina M, Babii A, Malafeev A. 2020. Emotion classification in Russian: feature engineering and analysis. In: *International Conference on Analysis of Images, Social Networks and Texts*. Springer, 135–148.

Kirilenko AP, Stepchenkova SO. 2017. Sochi 2014 Olympics on Twitter: perspectives of hosts and guests. *Tourism Management* **63**:54–65 DOI 10.1016/j.tourman.2017.06.007.

Koltsova OY, Alexeeva S, Kolcov S. 2016. An opinion word lexicon and a training dataset for Russian sentiment analysis of social media. In: *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue" 2016*. Russian State University for the Humanities, 277–287.

Konstantinov A, Moshkin V, Yarushkina N. 2021. Approach to the use of language models BERT and Word2Vec in sentiment analysis of social network texts. In: *Recent Research in Control Engineering and Decision Making*. Cham: Springer International Publishing, 462–473.

Korablinov V, Braslavski P. 2020. RUBQ: a Russian dataset for question answering over wikidata. In: *International Semantic Web Conference*. Cham: Springer, 97–110.

Kostenetskiy P, Chulkevich R, Kozyrev V. 2021. Hpc resources of the higher school of economics. *Journal of Physics: Conference Series* **1740**:12050 IOP Publishing DOI 10.1088/1742-6596/1740/1/012050.

Kotelnikov E. 2021. Current landscape of the Russian sentiment corpora. In: *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue" 2021*. Russian State University for the Humanities, 433–444.

Kotelnikova A. 2020. Comparison of deep learning and rule-based method for the sentiment analysis task. In: *2020 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon)*. Piscataway: IEEE, 1–6.

Krippendorff K. 1980. *Content analysis: an introduction to its methodology*. Thousand Oaks: SAGE.

Krippendorff K. 2004. Reliability in content analysis: some common misconceptions and recommendations. *Human Communication Research* **30(3)**:411–433 DOI 10.1111/j.1468-2958.2004.tb00738.x.

Kumar P, Raman B. 2022. A BERT based dual-channel explainable text emotion recognition system. *Neural Networks* **150**:392–407 DOI 10.1016/j.neunet.2022.03.017.

Kuratov Y, Arkhipov M. 2019. Adaptation of deep bidirectional multilingual transformers for Russian language. In: *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue" 2019*. Vol. 18. Russian State University for the Humanities, 333–340.

Landis JR, Koch GG. 1977. The measurement of observer agreement for categorical data. *Biometrics* **33(1)**:159–174 DOI 10.2307/2529310.

Larsen ME, Boonstra TW, Batterham PJ, O'Dea B, Paris C, Christensen H. 2015. We Feel: mapping emotion on Twitter. *IEEE Journal of Biomedical and Health Informatics* **19(4)**:1246–1252 DOI 10.1109/JBHI.2015.2403839.

Leetaru K. 2019. Is Twitter's spritzer stream really a nearly perfect 1% sample of its firehose? *Forbes* .

Li M, Ch'ng E, Chong AYL, See S. 2018. Multi-class Twitter sentiment classification with emojis. *Industrial Management & Data Systems* **118(9)**:1804–1820 DOI 10.1108/IMDS-12-2017-0582.

Lopatin V, Ulukhanov I. 2017. Russian language. *Languages of the World*. 276–540.

Loukachevitch N, Blinov P, Kotelnikov E, Rubtsova Y, Ivanov V, Tutubalina E. 2015. SentiRuEval: testing object-oriented sentiment analysis systems in Russian. In: *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue" 2015*. Vol. 2. Russian State University for the Humanities, 3–13.

Lukashevich N, Rubtsova YV. 2016. Sentirueval-2016: overcoming time gap and data sparsity in tweet sentiment analysis. In: *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue" 2016*. Russian State University for the Humanities, 416–426.

Mozeticar M, Smailović J. 2016. Multilingual Twitter sentiment classification: the role of human annotators. *PLOS ONE* **11(5)**:e0155036 DOI 10.1371/journal.pone.0155036.

Muhammad SH, Adelani DI, Ahmad IS, Abdulmumin I, Bello BS, Choudhury M, Emezue CC, Aremu A, Abdul S, Brazdil P. 2022. NaijaSenti: a Nigerian Twitter sentiment corpus for multilingual sentiment analysis. *Available at* https://arxiv.org/abs/2201.08277.

Pak A, Paroubek P. 2012. Language independent approach to sentiment analysis (LIMSI participation in ROMIP'11). In: *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue" 2012*. Vol. 11. 37–50.

Pavliy B, Lewis J. 2016. The performance of Twitter's language detection algorithm and Google's Compact Language Detector on language detection in Ukrainian and Russian tweets. *Bulletin of Toyama University of International Studies* **8**:99–106.

Pontiki M, Galanis D, Papageorgiou H, Androutsopoulos I, Manandhar S, AL-Smadi M, Al-Ayyoub M, Zhao Y, Qin B, De Clercq O, Hoste V, Apidianaki M, Tannier X, Loukachevitch N, Kotelnikov E, Bel N, Jiménez-Zafra SM, Eryigit G. 2016. SemEval-2016 task 5: aspect based sentiment analysis. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. San Diego, California: Association for Computational Linguistics, 19–30.

Prata DN, Soares KP, Silva MA, Trevisan DQ, Letouze P. 2016. Social data analysis of Brazilian's mood from Twitter. *International Journal of Social Science and Humanity* **6(3)**:179–183 DOI 10.7763/IJSSH.2016.V6.640.

Pronoza E, Panicheva P, Koltsova O, Rosso P. 2021. Detecting ethnicity-targeted hate speech in Russian social media texts. *Information Processing & Management* **58(6)**:102674 DOI 10.1016/j.ipm.2021.102674.

Read J. 2005. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In: *Proceedings of the ACL Student Research Workshop*. Ann Arbor, Michigan: Association for Computational Linguistics, 43–48.

Rodina J, Kutuzov A. 2020. RuSemShift: a dataset of historical lexical semantic change in Russian. In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain: International Committee on Computational Linguistics, 1037–1047.

Rogers A, Romanov A, Rumshisky A, Volkova S, Gronas M, Gribov A. 2018. RuSentiment: an enriched sentiment analysis dataset for social media in Russian. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 755–763.

Rosstat. 2022. How many people live in Russia: rosstat announced the first results of the census. *Available at* https://www.strana2020.ru/novosti/skolko-lyudey-zhivet-v-rossii-rosstat-ozvuchil-pervye-itogi-perepisi/.

Rubtsova Y. 2013. A method for development and analysis of short text corpus for the review classification task. In: *Trudy XV Vserossiiskoy Naychnoy Konferencii RCDL'2013*. 269–275.

Salminen JO, Al-Merekhi HA, Dey P, Jansen BJ. 2018. Inter-rater agreement for social computing studies. In: *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*. IEEE, 80–87.

Smetanin S. 2017. The program for public mood monitoring through Twitter content in Russia. *Proceedings of the Institute for System Programming of the RAS* **29(4)**:315–324 DOI 10.15514/ISPRAS-2017-29(4)-22.

Smetanin S. 2020a. The applications of sentiment analysis for Russian language texts: current challenges and future perspectives. *IEEE Access* **8**:110693–110719 DOI 10.1109/ACCESS.2020.3002215.

Smetanin S. 2020b. Toxic comments detection in Russian. In: *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue" 2020*. Russian State University for the Humanities.

Smetanin S, Komarov M. 2019. Sentiment analysis of product reviews in Russian using convolutional neural networks. In: *2019 IEEE 21st Conference on Business Informatics (CBI)*. Vol. 1. IEEE, 482–486.

Smetanin S, Komarov M. 2021a. Deep transfer learning baselines for sentiment analysis in Russian. *Information Processing & Management* **58(3)**:102484 DOI 10.1016/j.ipm.2020.102484.

Smetanin S, Komarov M. 2021b. Share of toxic comments among different topics: the case of Russian social networks. In: *2021 IEEE 23rd Conference on Business Informatics (CBI)*. Vol. 2. Piscataway: IEEE, 65–70.

Szczepański M, Pawlicki M, Kozik R, Choraś M. 2021. New explainability method for BERT-based model in fake news detection. *Scientific Reports* **11(1)**:1–13 DOI 10.1038/s41598-021-03100-6.

Szmigiera M. 2022. The most spoken languages worldwide in 2022. *Available at* https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/.

Ten Thij M, Bhulai S, Kampstra P. 2014. Circadian patterns in Twitter. In: *IARIA Data Analytics 2014*. 12–17.

Tripto NI, Ali ME. 2018. Detecting multilabel sentiment and emotions from Bangla YouTube comments. In: *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*. Piscataway: IEEE, 1–6.

Wang Y, Callan J, Zheng B. 2015. Should we use the sample? Analyzing datasets sampled from Twitter's stream API. *ACM Transactions on the Web* **9(3)**:1–23 DOI 10.1145/2746366.

Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Davison J, Shleifer S, von Platen P, Ma C, Jernite Y, Plu J, Xu C, Scao TL, Gugger S, Drame M, Lhoest Q, Rush AM. 2020. Transformers: state-of-the-art natural language processing. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations.* Association for Computational Linguistics, 38–45.

Zueva N, Kabirova M, Kalaidin P. 2020. Reducing unintended identity bias in Russian hate speech detection. In: *Proceedings of the Fourth Workshop on Online Abuse and Harms.* 65–69.