Information Processing and Management 58 (2021) 102484

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Deep transfer learning baselines for sentiment analysis in Russian

Sergey Smetanin *, Mikhail Komarov

National Research University Higher School of Economics, 101000 Moscow, Russia

ARTICLE INFO

MSC:
00-01
99-00

Keywords:
Sentiment analysis
Transfer learning
Russian texts

ABSTRACT

Recently, transfer learning from pre-trained language models has proven to be effective in a variety of natural language processing tasks, including sentiment analysis. This paper aims at identifying deep transfer learning baselines for sentiment analysis in Russian. Firstly, we identified the most used publicly available sentiment analysis datasets in Russian and recent language models which officially support the Russian language. Secondly, we fine-tuned Multilingual Bidirectional Encoder Representations from Transformers (BERT), RuBERT, and two versions of the Multilingual Universal Sentence Encoder and obtained strong, or even new, state-of-the-art results on seven sentiment datasets in Russian: SentRuEval-2016, SentiRuEval-2015, RuTweetCorp, RuSentiment, LINIS Crowd, and Kaggle Russian News Dataset, and RuReviews. Lastly, we made fine-tuned models publicly available for the research community.

1. Introduction

With the rapid growth of user-generated content, sentiment analysis is becoming ever more urgent research direction in natural language processing, which is broadly used in academic and industry segments. Scholars and companies commonly apply sentiment analysis to content from social media, user reviews, and news in order to extract user opinions about different topics. For instance, sentiment analysis can be used for predicting the stock market (e.g. Carosia, Coelho, & Silva, 2020), computing the Subjective Well-Being Index (e.g. Iacus, Porro, Salini, & Siletti, 2020), predicting election results (e.g. Sharma, Datta, & Pabreja, 2020), and measuring reactions to particular events or news (e.g. Georgiadou, Angelopoulos, & Drake, 2020).

The quality of sentiment analysis outcomes relates directly to the classification quality of the sentiment approach. Transfer learning to a variety of natural language processing tasks has come a long way, ranging from the usage of context-independent word vectors from unsupervised models (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Pennington, Socher, & Manning, 2014) to the current direct use of pre-trained transformer blocks (Devlin, Chang, Lee, & Toutanova, 2019; Yang et al., 2020) with an additional output layer stacked for the task-specific fine-tuning. Recent studies showed that transfer learning from pre-trained language models have proven to be effective in the sentiment classification task, confidently achieving strong results Liu, Shi, Ji, and Jia (2019). However, a minimal amount of research (Baymurzina, Kuznetsov, & Burtsev, 2019; Kuratov & Arkhipov, 2019) was focused on transfer learning from pre-trained language models in the sentiment analysis of Russian texts. In our recent survey Smetanin (2020), we comprehensively reviewed the applications of sentiment analysis of the Russian language content and identified current challenges and future research directions. We identified *transfer learning of language models for sentiment analysis of Russian-language texts* as one of the most relevant research opportunities, which can increase the quality of the applications of sentiment analysis for Russian-language texts. To obtain profound insights into the classification quality of language models on Russian texts and to provide scholars with strong classification baselines, it is required to conduct a synthesis of extant language models and transfer learning experiments. Thus, as the logical continuation of our survey, this paper addresses the following research question:

* Corresponding author.
E-mail address: sismetanin@gmail.com (S. Smetanin).

https://doi.org/10.1016/j.ipm.2020.102484
Received 15 April 2020; Received in revised form 29 November 2020; Accepted 23 December 2020
Available online 27 January 2021

Table 1
Sentiment analysis datasets.

Dataset	Description	Classes	Data Type	Domain	Academic Interest	Access Link
SentiRuEval-2016 Pontiki et al. (2016)	An aspect-based sentiment dataset of users' reviews from Twitter.	3	Tweets	Telecommunications companies and banks	Yes	Project page
SentiRuEval-2015 Loukachevitch et al. (2015)	An aspect-based sentiment dataset of users' reviews from Twitter.	3		Subset 1: cards and restaurants, Subset 2: telecommunications companies and banks	Yes	Project page
RuTweetCorp Rubtsova (2013)	A sentiment dataset of tweets with minor manual data filtering.	3		General	Yes	Project page
Twitter Sentiment for 15 European Languages Mozetič, Grčar, and Smailović (2016)	A sentiment dataset of tweets in 15 European languages.	3		General	Yes	Project page
Kaggle Russian_twitter_sentiment	A sentiment datasets of tweets.	2		N/S	No	Kaggle page
RuSentiment Rogers et al. (2018)	A sentiment dataset of posts from Vkontakte.	5	VKontakte posts	General	Yes	Project page
LINIS Crowd Koltsova, Alexeeva, and Kolcov (2016)	A sentiment dataset blog posts from social media sites.	5	Blog posts	Social and politics	Yes	Project page
Kaggle Russian News Dataset	A sentiment dataset of news articles.	3	News	General	Yes	Kaggle page
Kaggle Sentiment Analysis Dataset	A sentiment dataset of news articles.	3		N/S	No	Kaggle page
Kaggle IS161AIDAY	N/S	3	N/S	N/S	No	Kaggle page
RuReviews Smetanin and Komarov (2019)	A sentiment dataset of online reviews.	3	Reviews	Woman clothes and accessories	Yes	Project page

RQ: What classification quality in sentiment analysis of Russian language texts can be achieved by fine-tuning of the language models?

To get the answer to the formulated research question, this paper addresses the consistent set of research steps. To begin with, it is necessary to identify the most common public sentiment analysis datasets, which are commonly utilised by scholars. In addition to the list of the most common datasets, it is necessary to identify current state-of-the-art (SOTA) sentiment analysis approaches for each of these datasets. Once the most common datasets and current SOTA approaches are found, it is necessary to identify language models with the support of the Russian language and conduct transfer learning experiments. Based on the results of the experiments, we can compare the classification quality of the language models with existing SOTA approaches.

The rest of the article is organised as follows. In Section 2, we provided an overview of the related work, including a list of the most common sentiment datasets in Russian. In Section 3, we described the adoption of language models for the text classification task. In Section 4, we conducted transfer learning experiments for different datasets. In Section 5, we discussed obtained outcomes and performed error analysis. In conclusion, we presented the classification quality of fine-tuned models.

2. Literature review

2.1. Sentiment analysis datasets in Russian

Even though Russian is one of the most common languages in the World Wide Web,[1] generally, it is not as well-resourced as the English language, especially in the field of sentiment analysis. Although many studies aim at sentiment classification, only a few of them make their datasets publicly available for the research community. Based on the previously identified list of 14 public sentiment datasets of Russian-language texts Smetanin (2020), we selected only those datasets to which general methods of sentiment analysis (i.e. not aspect-based analysis) can be applied (see Table 1). Following this logic, we excluded Russian Hotel Reviews Dataset (Rybakov & Malafeev, 2018), RuSentRel (Rusnachenko & Loukachevitch, 2018), SemEval-2016 Task 5: Russian Pontiki et al. (2016), and SentuRuEval-2015 Subset 1 (Loukachevitch et al., 2015) from the further analysis. Formally,

[1] https://www.internetworldstats.com/stats7.htm

Table 2
Statistics of selected datasets.

Dataset	Classes	Average lengths	Max lengths	Train samples	Test samples	Overall samples	Access link
SentiRuEval-2016 Lukashevich and Rubtsova (2016)	3	87.0928	172	18,035	5,560	23,595	Project page
SentiRuEval-2015 Subtask 2 Loukachevitch et al. (2015)	3	81.4986	172	8,580	7,738	16,318	Project page
RuTweetCorp Rubtsova (2013)	3	89.1725	189	n/s	n/s	334,836	Project page
RuSentiment Rogers et al. (2018)	5	82.0279	800	28,218	2,967	31,185	Project page
LINIS Crowd Koltsova et al. (2016)	5	n/s	n/s	n/s	n/s	n/s	Project page
Kaggle Russian News Dataset Kaggle (2017)	3	3911.8501	381,498	n/s	n/s	8,263	Kaggle page
RuReviews Smetanin and Komarov (2019)	3	130.0693	1007	n/s	n/s	90,000	Project page

SentiRuEval-2016 Lukashevich and Rubtsova (2016) and SentiRuEval-2015 Subtask 2 (Loukachevitch et al., 2015) datasets are entity-oriented, but in most cases, texts contain only one entity, therefore general methods of sentiment analysis can be applied. Next, we categorised datasets based on the type of data and domain. After that, we measured the academic interest in each dataset following the simple heuristic: we considered the dataset to be academically interesting if there is at least one academic article that cites this dataset. The threshold value for heuristics was chosen so low since the total number of datasets for analysis is also not very large. We counted citations by performing a literature search in scientific databases and search engines: *Google Scholar*, *Russian Science Citation Index*, *IEEE Xplore*, *ACM Digital Library*, *ScienceDirect*, *SAGE Journals Online*, *Springer Link*. If the dataset had no academic interest, we excluded it from the further analysis. If several datasets were constructed within the same data source and domain, we selected for further analysis the one that had more citations in the academic literature. For example, both RuTweetCorp (Rubtsova, 2013) and Twitter Sentiment for 15 European Languages (Mozetič et al., 2016) datasets were created from general-domain tweets, but RuTweetCorp had much more citations in literature, so we considered only RuTweetCorp for further analysis.

Thus, we selected seven datasets for conducting transfer learning experiments.

1. **SentiRuEval-2016** Lukashevich and Rubtsova (2016) is a dataset of tweets about telecommunication companies and banks, which was used in the evaluation of Russian sentiment analysis systems in 2016.
2. **SentiRuEval-2015 Subtask 2** Loukachevitch et al. (2015) is a dataset of tweets about telecommunication companies and banks, which was used in the evaluation of Russian sentiment analysis systems in 2015.
3. **RuTweetCorp** Rubtsova (2013) is a dataset of general-domain tweets, which were labelled automatically.
4. **RuSentiment** Rogers et al. (2018) is a dataset of general-domain posts from the largest Russian social network, VKontakte.
5. **LINIS Crowd** Koltsova et al. (2016) is a dataset of social and political blog posts from social media sites.
6. **Kaggle Russian News Dataset** Kaggle (2017) is a public sentiment dataset of news, which was anonymously published at Kaggle.
7. **RuReviews** Smetanin and Komarov (2019) is a dataset of online reviews from the "Women's Clothes and Accessories" product category on the primary e-commerce site in Russia.

These datasets were constructed based on different data sources: reviews (e.g. Smetanin & Komarov, 2019), news (e.g. Kaggle, 2017), tweets (e.g. Loukachevitch et al. (2015), Lukashevich and Rubtsova (2016), Rubtsova (2013)), posts from VKontakte (e.g. Rogers et al., 2018), and blog posts from social media sites (e.g. Koltsova et al., 2016). The statistics for each dataset can be found in Table 2. We primarily considered only those datasets, which can be accessed based on instructions from the corresponding papers or official sites. For instance, following this strategy, we have not listed ROMIP datasets Chetvirokin, Braslavskiy, and Loukachevitch (2012), Chetvirokin and Loukachevitch (2013), because we were unable to get access to the datasets using their official website (the similar issue was also mentioned in Smetanin, 2020). However, even though RuSentiment is no longer available at the official GitHub due to request from VKontakte,[2] we listed RuSentiment for several reasons. Firstly, it is the first sentiment analysis dataset of posts from VKontakte, the largest social network in Russia. Secondly, it is the largest manually annotated sentiment analysis dataset in Russian with a high inter-rater agreement score. Lastly, we hope that the authors of RuSentiment will be able to successfully resolve differences about privacy issues with VKontakte and make the dataset publicly available again.

2.1.1. SentiRuEval-2016

The SentiRuEval-2016 competition Lukashevich and Rubtsova (2016) was dedicated to reputation monitoring of banks and telecommunications companies on Twitter. The authors utilised training and test collections from SentiRueval-2015 Subtask 2 Loukachevitch et al. (2015) as a part of the training collection. Additionally, they collected tweets about eight entities from the bank domain and seven entities from the telecom domain using Twitter Streaming API. Tweets for the test collection were obtained in two parts: during July 2015 and during November 2015. In the annotation phase, each tweet from the test collection was labelled by at least four annotations regarding a particular organisation into three classes: *Positive Class*, *Neutral Class*, and *Negative Class*.

[2] We obtained the copy of the dataset before VKontakte's request to remove it. However, the dataset is still available on Github on forked repositories.

Once the annotation phase was finished, the authors applied the "strong agreement" voting scheme, where the strong agreement for a tweet was considered in case of votes for a particular sentiment class exceeded votes for other classes with a margin of 2. All tweets, which did not satisfy the conditions of strong agreement, were excluded from the final collection.

2.1.2. SentiRuEval-2015 Telecommunications Companies and Banks

The SentiRuEval-2015 competition Loukachevitch et al. (2015) was dedicated to the evaluation of systems for automatic sentiment-analysis of Russian-language texts in relation to a given object or its properties. Participants were offered two subtasks. The first subtask was an object-oriented analysis of reviews of restaurants and cars. The main purpose of this subtask was to identify words and expressions denoting important characteristics of the entity (i.e. aspect terms) and to classify them by sentiment and generalised categories. The second subtask was to analyse the influence of tweets on the reputation of given companies. Such tweets can either express the user's opinion about the company, its products or services, or contain negative or positive facts about these companies. Formally, SentiRuEval-2015 Subtask 2 was entity-oriented, but in most cases, tweets contain only one entity, therefore we applied general methods of sentiment analysis. In our paper, we considered only the SentiRuEval-2015 Subtask 2 dataset.

The SentiRuEval-2015 Subtask 2 dataset consists of two subsets: tweets about telecommunication companies and tweets about banks. Each subset consists of training and test collections, which contains tweets published during different time intervals. The tweets of the test collection were published in 2013, while tweets of the training collection were published in 2014. During the annotation phase, each tweet was annotated by three assessors regarding a particular organisation into three classes: *Positive Class*, *Neutral Class*, and *Negative Class*. In case if tweet did not contain mentions of a particular organisation, it was skipped from the annotation procedure for this organisation. Thus, in most cases, tweets contained only one entity.

2.1.3. RuTweetCorp

RuTweetCorp Rubtsova (2013) is the largest automatically annotated dataset of public posts from the Russian-language segment of Twitter. The dataset was collected automatically, based on the Read (2005) strategy, i.e. each text was associated with the sentiment class based on the emoticons it contains. Following this strategy, the author collected posts with negative and positive emoticons via Twitter API. To obtain neutral posts, the author retrieved messages from microblogging news accounts. All retrieved posts were filtered out according to the following criteria: all tweets containing both positive and negative emotions were deleted; identical posts and retweets were deleted; uninformative tweets, with a length of fewer than 40 characters, were deleted. Thus, the annotation scheme of the final version of the dataset includes three classes: *Negative Class* (tweets with negative emoticons), *Neutral Class* (tweets from the news accounts), and *Positive Class* (tweets with positive emoticons).

As a consequence of the annotation approach, even a simple rule-based approach is able to demonstrate outstanding results. For example, if a model classifies text as positive if it contains the '(' character and as a negative overwise, it achieves $F_1 = 97.39\%$ in the binary classification task. To deal with the automatic sentiment analysis task, the authors recommend removing emoticons and URLs during the preprocessing stage. Moreover, the utilised annotation strategy makes an assumption that users express their opinions directly, i.e. without complicated literature techniques such as irony or sarcasm.

2.1.4. RuSentiment

RuSentiment Rogers et al. (2018) is the largest, manually-annotated dataset of general-domain public posts from VKontakte, the largest social network in Russian with about 100,000,000 active users per month.[3] RuSentiment was constructed based on materials of the previously conducted research on political bias during Euromaidan. To compile RuSentiment, the authors excluded, from the initial collection, texts relevant to the Euromaidain discourse, so the created collection contains general-domain posts. To filter out noise posts from the corpus, the authors defined the following criteria: the length of a post is between 10 and 800 characters, at least 50% of characters are alphabetic, at least 30% of characters are from the Russian Cyrillic alphabet, a post does not include more than four hashtags, and a post has not less than two comments. Additionally, the authors excluded from the posts URLs and non-textual content. The RuSentiment consists of 31,185 general domain posts, which were manually annotated into five classes:

1. *Positive Sentiment Class*, which covers both implicit and explicit sentiment.
2. *Negative Sentiment Class*, which covers both implicit and explicit sentiment.
3. *Neutral Sentiment Class*, which covers texts without any sentiment.
4. *Speech Act Class*, which covers formulaic greetings, thank-you posts and congratulatory posts.
5. *Skip Class*, which covers unclear cases, noisy posts, and texts that were likely not created by the users themselves.

The posts were annotated by six native speakers with a linguistics background. During the annotation phase, the annotators achieved a Fleiss' kappa score of 0.58. In addition to the dataset, the authors published sentiment annotation guidelines,[4] that can be utilised in further studies.

[3] https://vk.com/about

[4] https://github.com/text-machine-lab/rusentiment

2.1.5. LINIS Crowd

LINIS Crowd (Koltsova et al., 2016) is a project which consists of a public sentiment dataset and a sentiment lexicon. Both the sentiment dataset and the sentiment lexicon were constructed from social and political blog posts from Russian social media sites. The LINIS Crowd dataset was constructed based on 1,500,000 posts by the top 2,000 LiveJournal bloggers, which were published between March of 2013 and March of 2014. To select posts regarding political and social topics, the authors applied a topic modelling approach implemented in TopicMiner[5] in combination with a partial manual annotation. For the sentiment annotation phase, the authors selected 87 assessors, as diverse as possible, in terms of socio-geographical characteristics, e.g. education, gender, and region. The texts were manually annotated into five scores, from −2 for *Strong Negative Class* to 2 for *Strong Positive Class*. During the annotation phase, the annotators obtained Krippendorff's alpha of 0.278 for all posts and 0.312 for the posts that received non-zero scores.

The authors published the sentiment analysis dataset without annotators' scores aggregation, which can pose both challenges and opportunities. One the one hand, scholars can apply certain labels aggregation techniques in order to fit their specific needs. One the other hand, the lack of a single version of aggregated labels makes it difficult to compare the quality of different proposed approaches since they commonly apply different aggregation techniques.

2.1.6. Kaggle Russian News Dataset

Kaggle Russian News Dataset (Kaggle, 2017) is a public sentiment dataset of 8263 news articles. The dataset was originally composed of a train and test subsets. However, only the train subset was openly published, and there is no way to get access to the test subset. As a consequence, academic studies (Kirekov & Krajvanova, 2018; Nenashev, 2019; Shalkarbayuli, Kairbekov, & Amangeldi, 2018) analysed only the openly available subset. The primary language of news is Russian, but there are also some Kazakh and English names and titles in the texts. The texts were annotated into three classes: *Negative Class*, *Neutral Class*, and *Positive Class*. However, since the dataset was anonymously published at Kaggle, there is no information regarding the annotation procedure and inter-rater agreement metrics.

2.1.7. RuReviews

RuReviews (Smetanin & Komarov, 2019) is an automatically-annotated dataset of 90,000 online reviews from the "Women's Clothes and Accessories" product category on the primary e-commerce site in Russia, where user-ranked scores were used as class labels. During the construction of the dataset, the authors applied several noise removal techniques. For example, in some cases, different authors posted similar reviews even for different products. To discard duplicate reviews, the authors merged reviews with similar texts and assigned them with the most common class labels. It was also found, that, sometimes, it was challenging to distinguish reviews with the close score values, so the authors transformed the initial five-point scale to the three-point scale. Thus, the annotation scheme of the final version of the dataset includes three classes: *Negative Class* (for reviews with rating of "1" and "2"), *Neutral Class* (for reviews with rating of "3"), and *Positive Class* (for reviews with rating of "4" and "5").

One of the main drawbacks of the automatic sentiment annotation using user-ranked scores, lies in the fact, that the final class label assignments are based on only one annotation provided by the author of the review. As a consequence, in the case of any error made by the author of review during assigning rating, the final class label in the dataset can be incorrect. Moreover, since there are no sentiment annotation guidelines, each review's author, guided by the intuition and subjective perception of sentiment labels, which may vary significantly from author to author. The aggregation of reviews with the same texts and transformation of a five-point scale into a three-point scale reduce this pitfall only at some level.

2.2. Embeddings in Russian-language NLP tasks

When dealing with different Russian language NLP tasks, researches generally applied a variety of different methods for obtaining word and sentence embeddings, ranging from simple Bag-of-Words and Word2Vec models to pre-trained language models as ELMo or BERT. For example, Rodina et al. measured the intensity of diachronic semantic shifts in adjectives in English, Norwegian and Russian across 5 decades using Word2Vec embeddings Rodina et al. (2019). Malykh et al. in the context of the aspect extraction task explored different strategies for sentence embeddings generation, including averaging, self-attention models, positional encodings, Recurrent Neural Networks (RRNs), bidirectional RNNs, and Convolutional Neural Networks (CNNs) Malykh, Alekseev, Tutubalina, Shenbin, and Nikolenko (2019). Alimova et al. explored the automatic drug reactions and medical conditions from user-generated texts in Russian using Linear SVM with different features Alimova, Tutubalina, Alferova, and Gafiyatullina (2017), e.g. word embedding from RusVectores (Kutuzov & Kuzmenko, 2017), disease terms, drug names, and RuSentiLex sentiment lexicons (Loukachevitch & Levchik, 2016). Kutuzov et al. summarised the experience of applying Continuous Bag-of-Words and Continuous Skip-gram neural network models to the task of calculating semantic similarity for Russian (Kutuzov & Andreev, 2015). Alekseev and Nikolenko introduced a novel approach to constructing user profiles for recommender systems based on full-text items such as posts in a social network and likes using Word2Vec for word embeddings (Alekseev & Nikolenko, 2016). Karyaeva et al. investigated several techniques for hypernym extraction from a large collection of dictionary definitions in Russian (Karyaeva, Braslavski, & Kiselev, 2018) based on Word2Vec models from the RusVectores (Kutuzov & Kuzmenko, 2017). Khodak et al. introduced a fully unsupervised method for automated construction of WordNets based upon distributional representations of

[5] https://linis.hse.ru/soft-linis

sentences and word-senses combined with readily available machine translation tools (Khodak, Risteski, Fellbaum, & Arora, 2017). On RUSSE'2018 (Panchenko et al., 2018), that is, a shared task on word sense induction for the Russian language, a significant amount of approaches demonstrated good results ranking in top 2–5 were the methods based on clustering of textual contexts generated using word embeddings models pre-trained on large corpora, such as the Russian National Corpus.[6] Enikeeva and Popov proposed an approach to constructing a semantic lexicon for the Russian language based on distributional word representations, where they considered semantic relation as a linear transformation (Enikeeva & Popov, 2018). Loukachevitch and Parkhomenko investigated the task of extracting multiword expressions for Russian thesaurus RuThes, exploring several embedding-based features for phrases and their components (Loukachevitch & Parkhomenko, 2018). Smetanin and Komarov applied shadow-and-wide CNNs on top of Word2Vec embeddings for predicting rating of online order reviews (Smetanin & Komarov, 2019). Tutubalina et al. presented a new open-access corpus named Russian Drug Reaction Corpus (RUDREC) for the research community of biomedical natural language processing and pharmacovigilance (Tutubalina et al., 2020). As baseline models, they applied the following pre-trained language models: Multilingual Cased $\text{BERT}_{\text{Base}}$ (Devlin et al., 2019), RuBERT (Kuratov & Arkhipov, 2019) and RuDR-BERT (Multilingual Cased BERT pre-trained on the raw part of the RUDREC dataset).

Whereas previously the primary attention of NLP researches was focused on word embeddings, now their attention is shifting to sentence embeddings (Popov, Pugachev, Svyatokum, Svitanko, & Artemova, 2019) in a variety of NLP tasks. This trend can also be observed in applied studies on sentiment analysis for Russian-language content, where the percentage of machine learning-based approaches has significantly exceeded the share of rule-based approaches since 2019 (Smetanin, 2020). Thus, the usage of fine-tuned language models is potentially able to significantly increase sentiment classification quality and therefore improve the accuracy of the sentiment monitoring results (Smetanin, 2020).

2.3. Transfer learning in sentiment analysis

The key concept of transfer learning is to store knowledge gained while solving one task and apply it to a different but related task. In this paper, we decided to focus on the current most promising area in natural language processing, sequential transfer learning where tasks are learned in sequence (Ruder, Peters, Swayamdipta, & Wolf, 2019). Sequential transfer learning includes of two phases. The first one is a pretraining phase during which general representations are learned on a source task or domain. The second one is an adaptation phase in which the previously learned knowledge is applied to a target task or domain.

Pre-trained representations can either be context-free or contextual. Such context-free models as Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and FastText (Joulin, Grave, Bojanowski, & Mikolov, 2017) have proven the ability to properly represent words in a way that captures meaning-related, syntactic and grammar-based relationships. The adaptation of these word embeddings was widely explored in a range of sentiment analysis studies, including research on the Russian language (Rogers et al., 2018; Rusnachenko & Loukachevitch, 2018; Smetanin & Komarov, 2019). Since these models are context-independent, they are commonly unable to capture different meanings of the word. However, learning different vectors for multiple word meanings explicitly is also possible (Bartunov, Kondrashkin, Osokin, & Vetrov, 2016). While context-free models generate a single word embedding representation for each word, contextual models, instead, generate a representation of each word dynamically based on the other words in the sentence (Camacho-Collados & Pilehvar, 2018). Contextualised representations are usually deep, that means, that they commonly a function of all of the internal states of the model. Contextualised representations can, further, be unidirectional or bidirectional. In unidirectional contextualised models each word is only contextualised using the words to its left or right, while in bidirectional models each word is contextualised using both its left and right context. Some models combine representations from separate left-context and right-context models, but only in a "shallow" manner. For instance, Flair (Akbik, Blythe, & Vollgraf, 2018) and ELMo (Peters et al., 2018) incorporate word-level or character-level characteristics as well as contextual information. These models are shallow bidirectional since they use a shallow concatenation of independently trained left-to-right and right-to-left LSTM blocks (Hochreiter & Schmidhuber, 1997).

At the same time, pre-trained representations can either be world-level or sentence-level. Word-level embeddings are intended to map the meaning of words or phrases from the vocabulary to vectors of real numbers. Conceptually, this set of language modelling and feature learning techniques (e.g. Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and FastText (Joulin et al., 2017)) involves a mathematical embedding from multi-dimensional space per word or phrase to a continuous vector space with a much lower number of dimensions. In contrast with word-level representations, another group of models aims at the generation of sentence embeddings, that is, a vector representation of a sentence, paragraph or a whole text. In 2018, Facebook published InferSent (Conneau, Kiela, Schwenk, Barrault, & Bordes, 2017), which provides semantic representations for sentences trained on a natural language inference. ULMFiT (Howard & Ruder, 2018) achieved the new state-of-the-art on six text classification tasks, reducing the error by 18%–24% on the majority of datasets. It introduced a language model and a methodology of effective fine-tuning for an array of natural languages processing tasks. Thus, contextualised representations achieved extraordinary results in the sentiment analysis task. For instance, previously mentioned ULMFiT demonstrated an accuracy score up to 95.4% on the IMDb dataset (Maas et al., 2011), the error rate of 29.98% in fine-grained classification and 2.16% in binary classification on the Yelp Review dataset (Zhang, Zhao, & LeCun, 2015). Authors of (McCann, Bradbury, Xiong, & Socher, 2017) utilised the biattentive classification network with CoVe embeddings, thereby achieving accuracy scores up to 91.8% on the IMDb dataset, 53.7% on SST-5 and 90.3% on SST-2. ELMo demonstrated an impressive accuracy score of 54.7% on SST-5 (Socher et al., 2013). In the case of

[6] https://ruscorpora.ru/old/en/index.html

the Russian language, the ELMo-based approach (Baymurzina et al., 2019) achieved the state-of-the-art results on the RuSentiment dataset.

Currently, pre-trained language models have proven to be effective in capturing language representations by training on a large amount of unlabelled data. The most recent pre-trained language models such as Generative Pre-trained Transformer (OpenAI GPT) (Radford, Narasimhan, Salimans, & Sutskever, 2018), Generative Pre-trained Transformer 2 (OpenAI GPT-2) (Radford et al., 2019), Universal Sentence Encoder (USE) (Cer et al., 2018; Yang et al., 2020), Bidirectional Encoder Representations from Transformers (Devlin et al., 2019), XLNet (Yang et al., 2019) demonstrated state-of-the-art results in various natural language processing tasks. For instance, XLNet achieved a state-of-the-art accuracy score up to 96.21% on the IMDb dataset, while BERT took second place with an accuracy score up to 95.79% (Sun, Qiu, Xu, & Huang, 2019). The same situation can be found on the Yelp Review dataset, where XLNet and BERT took first and second places, consequently. USE achieved a top-5 classification score on the Subjectivity dataset (Pang & Lee, 2005). Taking into consideration recent surveys on contextual embeddings (Liu, Kusner, & Blunsom, 2020; Qiu et al., 2020), we compiled a list of modern language models. Since only BERT and M-USE officially supports the Russian language, we proceeded to examine these two models in more detail.

From its inception in 2018, BERT has become widespread in sentiment analysis applications. Li et al. proposed a new method GBCN, which utilises a gating mechanism, with context-aware aspect embeddings, to enhance BERT representation for aspect-based sentiment analysis (Li et al., 2020). The proposed approach achieved state-of-the-art results of $F_1 = 88.0\%$ and $F_1 = 92.9\%$ on the SentiHood and SemEval-2014 datasets, respectively. BERT representations became an efficient tool for the construction of classification neural network models. For example, in the cases of Deep-Attention with Bidirectional Encoder Representations from Transformers (DA-BERT) (Pei, Wang, Shen, & Ning, 2019), Encoder Representations from Transformers (TD-BERT) (Gao, Feng, Song, & Wu, 2019), and A Lexicalised Domain Ontology and a Regularised Neural Attention model (ALDONAr) (Meškel & Frasincar, 2020), BERT embedding allowed the authors to achieve strong classification quality. Since several of the released pre-trained versions of BERT were multilingual, it also boosted sentiment analysis studies on Non-English language content. For instance, to classify the sentiment of e-commerce product reviews in Chinese, Yang and colleagues proposed a hybrid approach (Yang, Li, Wang, & Sherratt, 2020). The approach was based on sentiment lexicon and combination of Convolutional Neural Network (CNN) and Attention-Based Bidirectional Gated Recurrent Unit (BiGRU), which utilise BERT for obtaining word vectors. For the analysis of the Russian language texts, Kuratov and Arkhipov performed language-specific unsupervised training with multilingual initialisation of the Multilingual BERT version and performed fine-tuning on a range of NLP tasks, including sentiment analysis (Kuratov & Arkhipov, 2019).

As one of the state-of-the-art models, Universal Sentence Encoder (USE) also has got much attention from scholars. For example, Wang et al. proposed a novel approach to sentiment analysis, for stock market prediction, based on a combination of SVM and USE models, which allows the authors to achieve encouraging results (Wang, Xu, & Wang, 2018). In the SymantoResearch model (Basile et al., 2019), one of the top systems submitted to the SemEval-2019 Task 3 (Chatterjee, Narahari, Joshi, & Agrawal, 2019), the authors ensembled fine-tuned BERT and USE models, which allowed them to distinguish sad, happy and angry emotions in conversational texts and separate them from the rest of the emotions more accurately. Within the paper (Ruseti et al., 2020), Ruseti and colleagues constructed a corpus of around 200,000 games reviews and performed a series of classification experiments to build a custom-tailored sentiment analysis model. The transfer-learning from the USE representation achieved the best classification accuracy of 67%, thereby establishing a new state-of-the-art result. In 2019, in addition to the monolingual USE model, a Multilingual USE was introduced (Yang et al., 2020), allowing researchers to analyse content in Non-English languages. However, to the best of our knowledge, the Multilingual USE has not been applied to any natural language processing task on Russian-language content.

Thus, we decided to evaluate the multilingual version of Bidirectional Encoder Representations from Transformers (M-BERT) (Devlin et al., 2019), RuBERT (Kuratov & Arkhipov, 2019) and Multilingual Universal Sentence Encoder (M-USE) (Yang et al., 2020) in the Russian-language sentiment analysis task. The decision was made based on the following factors. We identified M-BERT and M-USE as the only recent language models that officially supports the Russian language. M-BERT has already been widely recognised by scholars dealing with content analysis in Non-English language, so evaluation of this language model in the context of Russian language sentiment analysis became a priority task that needed to be done. In comparison with the M-BERT model, M-USE received slightly less attention from scholars. However, based on the classification metrics reported in the original paper, we assumed that this language model also holds a significant potential of the sentiment analysis of Russian language content. As a transfer learning approach, we decided to utilise fine-tuning, since recent fine-tuning studies reported the best classification results.

3. Language models fine-tuning for text classification

On the top of the pre-trained representations, a simple softmax classifier is usually applied to predict the final probability of class labels c:

$$p(c|h) = \mathrm{softmax}(W h), \tag{1}$$

where W is the task-specific parameter matrix of the added softmax layer. During the training stage, we fine-tuned both the pre-trained model parameters and W by maximising the log-probability of the correct label.

3.1. Bidirectional Encoder Representations from Transformers

In a major advance in 2018, Google's research group introduced Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019). This is the first unsupervised, deeply bidirectional neural network for pre-training natural language processing tasks, implementing bidirectional transformers architecture. It was trained on plain text using two unsupervised tasks: masked word prediction and next sentence prediction. The authors released the $\text{BERT}_{\text{Base}}$ and $\text{BERT}_{\text{Large}}$ models, which differs in the number of Transformer blocks, the hidden size, and the number of self-attention heads. Currently, two multilingual versions of $\text{BERT}_{\text{Base}}$ are available, but only the Cased version is officially recommended. As input data, BERT takes a sequence of no more than 512 tokens and outputs the representation of this sequence. The tokenization is performed by WordPiece tokenizer (Johnson et al., 2017) with the preliminary text normalisation and punctuation splitting. The tokenized sequence has a *[CLS]* token at the beginning of the first sentence and a *[SEP]* token at the end of each sentence. In order to perform text classification tasks, BERT utilises the final hidden state h of the first token *[CLS]* as the representation of the whole sequence of tokens.

We used pre-trained Multilingual BERT-base Cased,[7] which supports 104 languages including Russian, with 12 stacked Transformer blocks, a hidden size of 768, 12 self-attention heads, and 110,000,000 parameters in general. Additionally, we used RuBERT[8] (Kuratov & Arkhipov, 2019), that is, a BERT model trained on the Russian part of Wikipedia and news data, which have the same parameters configuration as Multilingual BERT-base Cased. The authors of RuBERT used this training data to build a vocabulary of Russian subtokens and took the multilingual version of BERT-base as initialisation for RuBERT. The fine-tuning stage was performed on 1 Tesla V100 SXM2 32GB GPU, with the recommended parameters from the paper (Sun et al., 2019) and the official repository[9]: a number of train epochs of 3, a number of warm-up steps of 10%, a max sequence length of 128 or 256, a batch size of 32, and a learning rate of 5e-5.

3.2. Multilingual Universal Sentence Encoder

Released in 2018 by Google, Universal Sentence Encoder (Cer et al., 2018) encodes greater-than-word length English texts into high dimensional embedding vectors that specifically target transfer learning to an array of natural language processing tasks. In 2019, based on previous USE versions, a research group from Google introduced Multilingual Universal Sentence Encoder (Yang et al., 2020) that includes training on multiple tasks (Chidambaram et al., 2019) across 16 languages including Russian. The authors published two versions of the USE, which target different design goals. $\text{USE}_{\text{Trans}}$ was based on Transformer architecture (Vaswani et al., 2017) and designed for high evaluations scores at the cost of greater model complexity and resource consumption. USE_{CNN} was trained with Convolutional Neural Network (CNN) and designed for efficient inference with slightly reduced evaluation scores. As input data, Multilingual $\text{USE}_{\text{Trans}}$ takes a sequence of no more than 100 tokens, while Multilingual USE_{CNN} takes a sequence of no more than 256 tokens. SentencePiece tokenization (Kudo & Richardson, 2018) is used for all supported languages.

We used pre-trained Multilingual $\text{USE}_{\text{Trans}}$[10] which supports 16 languages including Russian, with a Transformer encoder equipped with 6 transformer layers, 8 attention heads, filter size of 2048, hidden size of 512. We also used pre-trained Multilingual USE_{CNN}[11] which supports 16 languages including Russian, with CNN encoder with 2 CNN layers, filter width of [1, 2, 3, 5], and filter size of 256. For both models, the fine-tuning stage was performed on 1 Tesla V100 SXM2 32GB GPU, with the recommended parameters from and the TensorFlow Hub page[12]: a number of train epochs of 100, a batch size of 32, and a learning rate of 3e-4.

4. Experiment

We evaluated transfer learning approaches on seven public sentiment datasets in Russian: SentRuEval-2016, SentiRuEval-2015, RuTweetCorp, RuSentiment, LINIS Crowd, Kaggle Russian News Dataset, and RuReviews. These datasets were obtained through diverse sources and have different subject areas. Consequently, they have differences in sentiment annotation schema and in the text's characteristics. The statistics for each dataset can be found in Table 2. For datasets without preliminary division into training, validation and test subsets, we manually performed division into three subsets: training subset (70%), validation subset (15%), and test subset (15%). In order to compare our models with the existing ones, we use the classification metrics reported in the original paper of the dataset.

4.1. SentiRuEval-2016

In the SentiRuEval-2016 competition, macro-averaged F_1-measure for positive and negative classes was used as the main quality measure. Macro-averaged F_1^{PN}-measure was calculated as the average value between F_1-measure of the positive class and F_1-measure of the negative class ignoring the neutral class. Arkhipenko et al. (2016) submitted one of the winning systems, which achieves the best classification results for macro-averaged and micro-averaged F_1^{PN} on Banks subset and the best macro-averaged

[7] https://storage.googleapis.com/bert_models/2018_11_23/multi_cased_L-12_H-768_A-12.zip
[8] http://files.deeppavlov.ai/deeppavlov_data/bert/rubert_cased_L-12_H-768_A-12_v2.tar.gz
[9] https://github.com/google-research/bert
[10] https://tfhub.dev/google/universal-sentence-encoder-multilingual-large/1
[11] https://tfhub.dev/google/universal-sentence-encoder-multilingual/1
[12] https://www.tensorflow.org/hub/tutorials/text_classification_with_tf_hub

Table 3
Three-class sentiment classification on SentiRuEval-2016 Telecommunication Companies and Banks subsets. P, R and F_1 are reported in macro-averaged evaluation measures. F_1^{PN} is the average value between F_1-measure of the positive class and F_1-measure of the negative class ignoring the neutral class, which was used in the original competition as the main quality measure.

System	Telecommunications Companies					Banks				
	P	R	F_1	macro F_1^{PN}	micro F_1^{PN}	P	R	F_1	macro F_1^{PN}	micro F_1^{PN}
SentiRuEval-2016 SOTA (Arkhipenko et al., 2016; Lukashevich & Rubtsova, 2016)	n/s	n/s	n/s	55.94%	68.13%	n/s	n/s	n/s	55.17%	58.81%
Current SOTA (Golubev & Loukachevitch, 2020)	n/s	n/s	68.42%	66.07%	74.11%	n/s	n/s	**79.51%**	**67.44%**	**70.09%**
M-BERT-Base-FiT-SemEval2016	65.73%	67.00%	66.29%	61.78%	72.45%	62.74%	70.13%	65.31%	58.00%	60.52%
RuBERT-FiT-SemEval2016	70.82%	70.57%	**70.68%**	**66.40%**	**76.71%**	71.05%	74.62%	72.83%	65.89%	68.43%
M-USE-CNN-FiT-SemEval2016	64.30%	63.12%	63.64%	58.97%	71.31%	66.06%	68.28%	66.71%	58.73%	62.41%
M-USE-Trans-FiT-SemEval2016	69.45%	67.44%	68.27%	62.77%	75.00%	73.04%	71.94%	72.40%	65.04%	68.21%

Table 4
Three-class sentiment classification on SentiRuEval-2015 Telecommunications Companies and Banks subsets. P, R and F_1 are reported in macro-averaged evaluation measures. F_1^{PN} is the average value between F_1-measure of the positive class and F_1-measure of the negative class ignoring the neutral class, which was used in the original competition as the main quality measure.

System	Telecommunications Companies					Banks				
	P	R	F_1	macro F_1^{PN}	micro F_1^{PN}	P	R	F_1	macro F_1^{PN}	micro F_1^{PN}
SentiRuEval-2015 SOTA (Adaskina, Panicheva, & Popov, 2015; Loukachevitch et al., 2015)	n/s	n/s	n/s	48.8%	53.6%	n/s	n/s	n/s	36.0%	36.6%
Current SOTA (Golubev & Loukachevitch, 2020)	n/s	n/s	**68.54%**	**63.47%**	**67.51%**	n/s	n/s	**79.51%**	**67.44%**	**70.09%**
M-BERT-Base-FiT-SemEval2016	59.20%	68.91%	60.47%	53.16%	57.03%	65.07%	72.05%	67.65%	56.97%	59.32%
RuBERT-FiT-SemEval2016	62.52%	73.18%	64.39%	57.76%	61.38%	68.10%	75.18%	70.58%	60.95%	63.33%
M-USE-CNN-FiT-SemEval2016	58.93%	66.95%	60.57%	52.37%	57.76%	64.97%	68.29%	66.32%	54.74%	57.61%
M-USE-Trans-FiT-SemEval2016	62.54%	73.61%	64.28%	57.60%	61.18%	68.75%	70.97%	69.62%	59.12%	62.17%

F_1^{PN} on Telecommunications Companies subset. The authors tested different neural network-based approaches, including Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), Long-Short Term Memory neural networks (LSTMs) and Gated Recurrent Units (GRUs). Two-layer GRU on the top of pre-trained Word2Vec demonstrated the best classification scores on both subsets in terms of macro-averaged and micro-averaged F_1^{PN} within their submission. The best classification result of micro-averaged F_1^{PN} on Telecommunications Companies subset was achieved by the team #9 (Lukashevich & Rubtsova, 2016), which utilised Support Vector Machine over unigrams, bigrams, and trigrams with two sentiment vocabularies. The best classification scores from SentiRuEval-2016 competition were surpassed by Golubev and Loukachevitch (Golubev & Loukachevitch, 2020), which used different approaches based on pre-trained Conversational RuBERT models (Kuratov & Arkhipov, 2019).

Within our research, we fine-tuned models for two subsets at once. The results of the fine-tuning Multilingual $\text{BERT}_{\text{Base}}$, RuBERT, and two versions of Multilingual USE can be found in Table 3. RuBERT achieved the best macro-averaged and micro-averaged F_1^{PN} scores on the Telecommunications Companies subset, demonstrating the new state-of-the-art classification score.

4.2. SentiRuEval-2015 subtask 2

In the SentiRuEval-2015 competition, macro-averaged F_1-measure for positive and negative classes was used as the main quality measure. Adaskina et al. (2015) submitted one of the winning systems, which achieves the best classification results for macro-averaged and micro-averaged F_1 on Telecommunications Companies subset and micro-averaged F_1 on Banks subset. The authors trained Support Vector Machine with features comprised word and letter n-grams, syntactic links presented as triples (a head word, a dependent word, type of relation), and techniques involving rule-based fact-extraction. The best classification scores from SentiRuEval-2015 competition were surpassed by Golubev and Loukachevitch (Golubev & Loukachevitch, 2020), which used different approaches based on pre-trained Conversational RuBERT models (Kuratov & Arkhipov, 2019).

Within our research, we fine-tuned models for two subsets at once. The results of the fine-tuning Multilingual $\text{BERT}_{\text{Base}}$, RuBERT, and two versions of Multilingual USE can be found in Table 4. However, our models failed to achieve the new state-of-the-art results on SentiRuEval-2015 Telecommunications Companies and Banks subsets.

4.3. RuSentiment

Within the current state-of-the-art approach (Baymurzina et al., 2019), the authors conducted a series of experiments of pre-training FastText and ElMo embeddings on different sources of the Russian-language texts: articles from Wikipedia, Russian WMT

Table 5
Five-class sentiment classification on RuSentiment. P, R and F_1 are reported in macro-averaged evaluation measures. Weighted F_1 is also reported since it was used in the original paper as the main quality measure.

System	P	R	F_1	weighted F_1
Current SOTA (Baymurzina et al., 2019)	n/s	n/s	n/s	**78.50%**
M-BERT-Base-RuSentiment	67.22%	69.07%	67.94%	72.44%
RuBERT-FiT-RuSentiment	70.89%	73.62%	72.03%	75.71%
M-USE-CNN-FiT-RuSentiment	65.71%	67.08%	66.27%	71.05%
M-USE-Trans-FiT-RuSentiment	68.21%	69.82%	68.60%	73.42%

Table 6
Three-class sentiment classification on Kaggle Russian News Dataset. P, R, and F_1-measure are reported in macro-averaged evaluation measures.

System	P	R	F_1
Current SOTA (Shalkarbayuli et al., 2018)	70.00%	70.00%	70.00%
M-BERT-Base-FiT-KSDRN	72.16%	70.68%	71.36%
RuBERT-FiT-KSDRN	74.33%	73.03%	**73.63**%
M-USE-CNN-FiT-KSDRN	72.68%	70.59%	71.27%
M-USE-Trans-FiT-KSDRN	74.00%	71.62%	72.66%

News, and posts from Twitter. On the top of the obtained embeddings, they trained shallow-and-wide Convolutional Neural Networks (CNN) and Bidirectional Gated Recurrent Units (BiGRU) on RuSentiment dataset. Shallow-and-wide CNN trained on top of Twitter ELMo embeddings achieved the best classification quality of weighted $F_1 = 78.5\%$.

In our study, we performed the pre-processing stage during BERT training since there are some issues in emoji tokenization in BERT. We replaced all emoji with the corresponding short names using the Emoji[13] library. The results of the fine-tuning Multilingual $\text{BERT}_{\text{Base}}$, RuBERT, and two versions of Multilingual USE can be found in Table 5. RuBERT achieved the best score of weighted $F_1 = 75.71\%$ among our models, thereby outperforming basic multilingual BERT score of weighted $F_1 = 72.44\%$. Although it also exceeded the existing BERT-based approach (Kuratov & Arkhipov, 2019), it failed to achieve the new state-of-the-art result on RuSentiment.

4.4. Kaggle Russian News Dataset

The dataset of Russian news from Kaggle (Kaggle, 2017) was originally composed of a train and test subsets. However, only the train subset was openly published, and there is no way to get access to the test subset. As a consequence, academic studies (Kirekov & Krajvanova, 2018; Nenashev, 2019; Shalkarbayuli et al., 2018) analysed only an openly available subset. Within the current state-of-the-art approach (Shalkarbayuli et al., 2018), the authors conducted a series of experiments of machine learning algorithms training. At the preprocessing stage, the authors removed stopwords, punctuations, digits, and dates, then stemmed words and removed named entities. Next, they applied TF-IDF vectorisation and trained a set of machine learning models: Support Vector Machine, Naive Bayes, Decision Tree, Random Forest, Linear Regression. The SVM model achieved the best classification quality of weighted $F_1 = 70\%$.

The results of fine-tuning Multilingual $\text{BERT}_{\text{Base}}$, RuBERT, and two versions of multilingual USE can be found in Table 6. RuBERT achieved $F_1 = 73.63\%$ (3.63 point absolute improvement), demonstrating the new state-of-the-art classification score.

4.5. LINIS Crowd

In LINIS Crowd (Koltsova et al., 2016), the annotations made by assessors were not aggregated. As a consequence, different studies utilised different aggregation strategies. For example, Lagutina (Lagutina, Larionov, Petryakov, Lagutina, & Paramonov, 2018) conducted the research where only 2,160 positive and 2,160 negative posts were randomly selected from the entire dataset. Within our study, we decided to construct an aggregation script based on a relative majority voting system, which can be further used in other studies. To begin with, we concatenated datasets between 2015 and 2016 in one dataset. Next, we removed all posts, which were labelled by less than three annotators. After that, we converted class labels into a three-point scale. Then, we performed aggregation of annotators labels, i.e. each text was associated with the most common class label assigned by annotators. We also removed texts with the unclear sentiment, i.e. texts where the relative majority of the labels was related to different classes. The created dataset consists of 5,866 texts, where 520 texts refer to *Positive Class*, 2,928 texts refer to *Neutral Class*, and 2,418 texts refer to *Negative Class*. The average length of the texts is 1,073 characters, and the maximum length is 22,515.

[13] https://github.com/carpedm20/emoji/

Table 7
Three-class sentiment classification on LINIS Crowd. All metrics are reported in macro-averaged evaluation measures.

System	P	R	F_1
SentiStrength (Thelwall, Buckley, Paltoglou, Cai, & Kappas, 2010) with LINIS sentiment lexicon (Koltsova et al., 2016)	38.19%	39.81%	37.29%
M-BERT-Base-FiT-LINIS	41.03%	45.20%	42.73%
RuBERT-FiT-LINIS	61.64%	59.94%	**60.51%**
M-USE-CNN-FiT-LINIS	57.24%	55.72%	56.34%
M-USE-Trans-FiT-LINIS	60.12%	55.49%	56.95%

Table 8
Three-class sentiment classification on RuTweetCorp. All metrics are reported in macro-averaged evaluation measures. MNB is a simple baseline approach based on Multinominal Naive Bayes classifier with the same preprocessing as for language models.

System	Ternary			Binary		
	P	R	F_1	P	R	F_1
Current SOTA (Rubtsova, 2018)	75.95%	74.71%	75.95%	n/s	n/s	n/s
Current SOTA (Zvonarev & Bilyi, 2019)	n/s	n/s	n/s	n/s	n/s	78.1%
MNB	73.12%	73.10%	73.11%	76.11%	76.11%	76.10%
M-BERT-Base-FiT-RuTweetCorp	82.96%	83.15%	83.04%	80.11%	80.11%	80.10%
RuBERT-FiT-RuTweetCorp	83.63%	83.77%	**83.69%**	80.81%	80.81%	**80.79%**
M-USE-CNN-FiT-RuTweetCorp	81.47%	81.32%	81.34%	78.54%	78.39%	78.39%
M-USE-Trans-FiT-RuTweetCorp	83.30%	83.16%	83.17%	80.06%	79.80%	79.69%

Since the majority of previous studies utilised only part of the dataset in the research or applied custom label aggregation techniques, we identified the original sentiment classification approach described by the authors of LINIS Crowd as a current state-of-the-art solution. In order to verify its classification quality on the aggregated dataset, we reproduce the described approach, based on SentiStrength (Thelwall et al., 2010) classifier with Russian sentiment lexicon created within LINIS Crowd project (Koltsova et al., 2016). We performed division into three subsets: training subset (70%), validation subset (15%), and test subset (15%).

The results of fine-tuning Multilingual $\text{BERT}_{\text{Base}}$, RuBERT, and two versions of Multilingual USE can be found in Table 7. RuBERT achieved $F_1 = 60.51\%$, outperforming SentiStrength baseline.

4.6. RuTweetCorp

According to the original paper (Rubtsova, 2013), RuTweetCorp contains three classes: *Positive Class*, *Negative Class*, and *Neutral Class*. Positive and negative tweets were published at the official site of the dataset, and the neutral tweets were published in a separate resource. We assume that as a consequence of the publication strategy, several studies utilised only positive and negative tweets, performing binary classification (Garshina, Kalabukhov, Stepantsov, & Smotrov, 2017; Lagutina et al., 2018; Romanov, Vasilieva, Kurtukova, & Meshcheryakov, 2017; Smirnova & Shishkov, 2016). The additional nuance lies in the fact that RuTweetCorp was initially designed for the creation of a sentiment lexicon, not for sentiment classification. The dataset was collected automatically based on the Read (2005) strategy, i.e. each text was associated with the sentiment class based on the emoticons it contained. As a consequence, even a simple rule-based approach is able to demonstrate outstanding results. For example, if a model classifies a text as positive in case it contains '(' character and as negative overwise, it achieves $F_1 = 97,39\%$ in the binary classification task. To deal with the automatic sentiment analysis task, the authors recommend removing emoticons during the pre-processing stage. Thus, when analysing the literature, we took into account only those papers in which the preprocessing procedure was explicitly described and all the recommendations of the authors of the dataset were taken into account. Current state-of-the-art approach (Rubtsova, 2018) for three-class classification achieved $F_1 = 75.95\%$ using a Support Vector Machine. Current state-of-the-art approach (Zvonarev & Bilyi, 2019) for binary classification achieved $F_1 = 78.1\%$ using a Convolutional Neural Network.

In this research, we decided to perform binary (*Positive Class* and *Negative Class*) and three-class (*Positive Class*, *Neutral Class*, and *Negative Class*) classification on the preprocessed dataset without emoticons and URLs. The list of emoticons was obtained from Ekphrasis,[14] described in the paper (Baziotis, Pelekis, & Doulkeridis, 2017). We additionally removed '(' and ')' characters from texts since they commonly represent positive and negative polarity, respectively. The results of fine-tuning Multilingual $\text{BERT}_{\text{Base}}$, RuBERT and two versions of Multilingual USE can be found in Table 8. RuBERT achieved $F_1 = 86.94\%$ (10.99 point absolute improvement) in a three-class classification and $F_1 = 82.31\%$ (4.21 point absolute improvement) in binary classification, demonstrating the new state-of-the-art classification scores in both cases.

[14] https://github.com/cbaziotis/ekphrasis

Table 9
Three-class sentiment classification on RuReviews. All metrics are reported in macro-averaged evaluation measures.

System	P	R	F_1
Current SOTA (Smetanin & Komarov, 2019)	75.63%	75.31%	75.45%
M-BERT-Base-FiT-RuReviews	77.49%	77.17%	77.31%
RuBERT-FiT-RuReviews	77.65%	77.29%	**77.44**%
M-USE-CNN-FiT-RuReviews	76.96%	76.42%	76.63%
M-USE-Trans-FiT-RuReviews	77.13%	76.96%	76.94%

4.7. RuReviews

Current state-of-the-art approach (Smetanin & Komarov, 2019) applied Convolutional Neural Networks (Kim, 2014) with pre-trained Word2Vec embeddings for the sentiment classification, achieving macro-average $F_1 = 75.45\%$. At the preprocessing stage, the authors converted all characters to lowercase; replaced repeated, capitalised and elongated words with the corresponding tags; corrected elongated words; and replaced the most common emoticons with emotional tags. To identify the most effective word representations, they evaluated Word2Vec, FastText and Glove models, and selected Word2Vec since its classification quality was the highest. The proposed network architecture consisted of several layers: the embeddings layer obtained from Word2Vec, 100 parallel convolution layers with filter sizes from 1 to 5, 1-max-pooling layer for each convolutional layer, the concatenation layer, the fully-connected hidden layer, and softmax output layer.

The results of fine-tuning Multilingual $\text{BERT}_{\text{Base}}$, RuBERT, and two versions of Multilingual USE can be found in Table 9. All four models exceed the current state-of-the-art approach in terms of macro-averaged precision, recall and F_1-measure. RuBERT achieved $F_1 = 77.44\%$ (1.99% point absolute improvement), demonstrating the best classification score.

5. Discussion

5.1. Results analysis

In the majority of cases, fine-tuned RuBERT demonstrated the best classification quality in comparison with other fine-tuned language models. The closest performance was demonstrated by $\text{M-USE}_{\text{Trans}}$, which often showed almost the same results as RuBERT. $\text{M-USE}_{\text{Trans}}$ always demonstrated the higher scores than $\text{M-USE}_{\text{CNN}}$. According to the confusion matrices in Fig. 1, the most common misclassification errors were classifying neutral texts as negative or positive, and classifying negative texts as neutral. The examples of such misclassifications can be found in Table 10. Neutral sentiment is logically located between negative and neutral sentiment, so it is expected that it can be classified incorrectly. Moreover, this issue looks like a general challenge of non-binary sentiment classification. For example, Barnes et al. also reported that the most common errors come from the no-sentiment classes (i.e. neutral class in our case) (Barnes, Øvrelid, & Velldal, 2019). The deep and comprehensive analysis of these misclassification errors is a great future research direction, which will provide academics with a broader understanding of the root cause of the problem. The framework proposed by Barnes et al. can be considered as a good foundation for such kind of research. The only exception was automatically-annotated RuTweetCorp, which utilised tweets from news accounts as a source of data with a neutral sentiment. In the case of fine-grained classification on RuSentiment, speech acts were clearly separated from other classes. Predictably, the *Skip Class* was one of the most hardly classified, since it initially contained hardly interpretable posts.

In several cases, we were unable to exceed current SOTA results. The first one is the fine-tuned ELMo (Baymurzina et al., 2019) trained on RuSentiment, which is technically also a language model. The second one is BERT sentence-pair models (Golubev & Loukachevitch, 2020) trained on SentiRuEval-2016 Banks, SentiRuEval-2015 TC, and SentiRuEval-2015 Banks datasets, which are also language models. Thus, considering the obtained results, we can state that in the context of existing approaches, sentiment analysis of Russian-language texts based on the language models outperforms rule-based and basic machine learning-based approaches in terms of classification quality.

The important but expected result from our research is that the larger version of a model tends to perform better than a smaller one. At the same time, large architectures such as the BERT, USE or XLNet make the process of training and deployment difficult, requiring an extensive amount of computational resources.

5.2. Applications of sentiment analysis

The majority of existing papers, which applies sentiment analysis of Russian language texts for different research questions, used rule-based and basic machine learning approaches, and only several studies utilised neural networks (Smetanin, 2020). In this study, we showed that fine-tuned language models are commonly able to outperform basic machine learning approaches and basic neural networks. We believe that by using these fine-tuned models scholars will be able to significantly increase sentiment classification quality within their studies and therefore improve the accuracy of the research outcomes. For instance, these models can be used in public mood monitoring of social media content in Russian. While for many languages, such studies have already been conducted, the research of Russian-language content remains quite limited (Panchenko, 2014). It can be broadened and deepened in

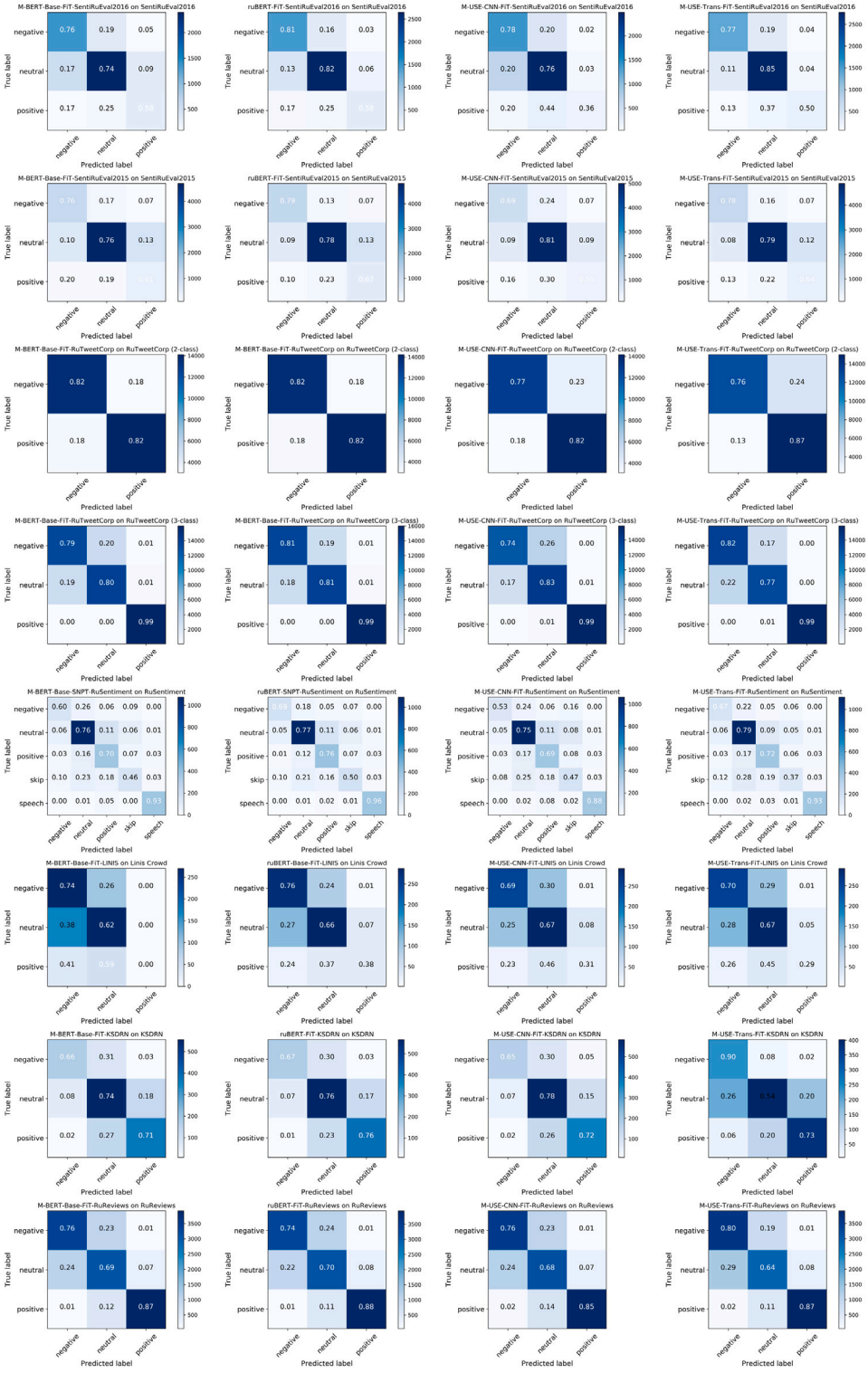

Fig. 1. Normalised confusion matrices.

Table 10
Examples of the most common misclassification errors.

Dataset	Original Text	Translated Text	Predicted Label	True Label
SentiRuEval-2016	Ну что же, в целом я приятно удивлен LTE от @Beeline_RUS, и покрытие хорошее и работает. Жаль только что мест без 3G и LTE в Москве много	Well, overall I'm pleasantly surprised by @Beeline_RUS LTE and the coverage is good and works. It's a pity that there are a lot of places without 3G and LTE in Moscow	Neutral	Positive
	@Beeline_RUS Я больше 10 дней ожидаю ответа на заявление о расторжении договора. Это нормально?	@Beeline_RUS For more than 10 days I have been waiting for an answer to the application to terminate the contract. This is normal?	Neutral	Negative
SentiRuEval-2015	хм, судя по карте у мтса нет связи в Тунгале и Исе -_- а у билайна есть... Не круто...	hmm, according to the map, mts has no connection in Tungala and Isa - _- but the beeline has... Not cool...	Neutral	Positive
	Сеть beeline начала глючить	Beeline network started to fail	Neutral	Negative
RuTweetCorp	Вообще нормально просыпаться рано и делать всевозможные дела	It's generally okay to wake up early and do all sorts of things	Negative	Neutral
	Спала почти 15 часов, потом еще днем часа 3, а все потому, что только когда я сплю мне не болит живот	I slept for almost 15 h, then another 3 h in the afternoon, and all because only when I sleep my stomach does not hurt	Neutral	Negative
Linis Crowd	Ну никак не могу объяснить :) Просто хочется. Это да. С хотелками бороться бесполезно. Ну тоже конечно аргумент, но жиденький какой то. Кхм. Это если сравнивать на прилчных боеприпасах. Кои у нас в дефиците. А ежели кормить барнаулом каким? (хихикая) Ругеровские футболки и не-сайгу жене.	Well, I just can't explain :) I just want to. This is yes. It's useless to fight with Wishlist. Well, also of course an argument, but some kind of thin. Ahem. This is when compared to decent ammunition. Which are in short supply. And if you feed Barnaul what? (chuckling) Ruger T-shirts and a non-saiga wife.	Neutral	Positive
	Одна из наиболее распространенных сегодня версий кулак, это крепкий хозяйственник, который держит все свое хозяйство в кулаке. Но вначале XX века больше была распространена другая версия. Одним из основных способов обогащения кулака дача денег или зерна в рост. То есть: кулак дает деньги своим односельчанам, или дает зерно, посевной фонд бедным односельчанам. Дает с процентами, довольно приличными. За счет этого он этих односельчан разоряет, за счет этого он становится богаче. Как этот кулак получал свои деньги или зерно обратно? Вот он дал, допустим, зерно в рост это происходит, например, в Советском Союзе в 20-е годы, то есть до раскулачивания.	One of the most common versions of the Kulak today, it is a tough business executive who keeps his entire household in a fist. But at the beginning of the twentieth century, another version was more widespread. One of the main ways of Kulak enrichment is giving money or grain to grow. That is: Kulak gives money to its fellow villagers, or gives grain, a sowing fund to poor fellow villagers. Gives with pretty decent interest. Due to this, he ruins these fellow villagers, due to this he becomes richer. How did this Kulak get his money or grain back? So he gave, let's say, grain for growth, this happens, for example, in the Soviet Union in the 20s, that is, before dekulakization.	Neutral	Negative
RuReviews	качество хорошее. заказ пришел довольно быстро. продавца рекомендую.	good quality. The order came pretty quickly. I recommend the seller.	Neutral	Positive
	Качество толстовки не соответствует описанию. СИНТЕТИКА. Поверхность ткани блестит.	The quality of the sweatshirt does not match the description. SYNTHETICS. The surface of the fabric is shiny.	Neutral	Negative

terms of analysed data volume, quality of sentiment classification model and methodology of social sentiment indexes calculation. Additionally, some studies were dedicated to developing the sentiment monitoring systems for analysis of Russian-language social media, but the authors commonly did not report any results of sentiment monitoring. In addition to sentiment component of texts, these applied studies may also incorporate demographic information, e.g. geolocation, gender, and age group of users. Depending on the source of the data, this information can be directly obtained from the user profiles. However, in some cases, this information can be partially hidden or completely unavailable, so researches may also apply machine learning methods to identify gender or age group based on texts (Litvinova, Sboev, & Panicheva, 2018; Panicheva, Mirzagitova, & Ledovaya, 2017; Sboev, Litvinova, Gudovskikh, Rybka, & Moloshnikov, 2016).

The applications of sentimental analysis, in most cases, do not occur in isolation. Usually, it is one of the components of data processing and decision-making systems, where the data collection system is also an important component. Moreover, taking into account a rapidly increasing role of data in our everyday lives, individuals and businesses need to be assured of provenance and quality of data. Distributed ledgers have the potential to help provide this assurance through delivering an immutable log of what happens to that data and ensuring a high level of data storage security for the reliability of subsequent analysis and decision making. To improve the reliability of data storage, solutions based on Distributed Ledger Technologies (DLT) can be used. DLT allows to

develop and deploy shared, digital infrastructures for applications by enabling the operation of a highly available, append-only distributed database in an untrustworthy environment (Lamport, Shostak, & Pease, 1982). This is achieved partly through separated storage and computing devices (referred to as nodes), which maintain a local replication of the ledger (Kannengießer, Lins, Dehling, & Sunyaev, 2020). In turn, nodes are controlled and maintained by organisations or individuals. An untrustworthy environment is characterised by the arbitrary occurrence of Byzantine failures (Lamport et al., 1982), such as network delays, unreachable or crashed nodes, and malicious behaviour of nodes (Sunyaev, 2020). The data in the ledger are appended in the form of transactions and is further stored in a chronologically-ordered sequence. Each transaction contains meta-data (e.g., timestamp) and a digital representation of certain assets (e.g., program code of a smart contract). When a node receives a new transaction, it validates the transaction by the selected consensus protocol, e.g. by involving more heterogeneous nodes as part of proof-of-activity (Zhidanov et al., 2019). Analysis of the literature has shown that the use of DLT generally increases the level of security (Natoli & Gramoli, 2017; Weber et al., 2017). However, it affects other parameters of the system, for example, performance (Kannengießer, Lins, Dehling, & Sunyaev, 2019; Kasahara & Kawahara, 2019; Smetanin, Ometov, Komarov, Masek, & Koucheryavy, 2020). Thus, to assess the performance of DLT-based sentimental analysis system, it is also necessary to simulate and evaluate the parameters of the systems that are used at the initial level to collect and store datasets for subsequent analysis (Smetanin et al., 2020), which is one of the directions for future research and is important component in assessing the effectiveness of sentiment analysis systems.

The systems combining distributed ledger technologies (e.g. blockchain) and sentiment analysis techniques are in the initial phase, but they tend to have great prospects. In general, data management tends to be one of the most indisputable properties of the blockchain since this technology provides the great benefits in the context of secure and privacy (Casino, Dasaklis, & Patsakis, 2019). Such benefits as safety, privacy, automation and accountability, that blockchain offers for processing public records, could make government services more efficient (Casino et al., 2019). This is especially relevant for integrating social, physical, and business infrastructures in a smart city context, where blockchain is able to serve as a highly-secure communication platform. For example, Banerjee et al. proposed a decentralised policy feedback system for privacy and governance using Blockchain and sentiment analysis for smart city applications (Banerjee, Mondal, Deb, & Ghosh, 2020). This system is a clear illustration of Blockchain and sentiment analysis combination for proper governance and framing of policies, where the governed can express their opinions without the risk of their identities being misused. In this case, the Blockchain system works in tandem with sentiment analysis to not just ensure fair feedback to a framed policy but also to help in maintaining privacy. From the business point of view, blockchain holds a significant potential to become a major source of disruptive innovations in automating, improving, and optimising business processes (Casino et al., 2019). In particular, blockchain-based systems are expected to increase the level of accountability and transparency in supply chain networks. Following this idea, Mao et al. introduced a credit evaluation system that adopts blockchain technology to strengthen the supervision and management of traders in the food supply chain (Mao, Wang, Hao, & Li, 2018). In this system, both transaction and credit evaluation of traders are grouped and stored in blocks. The credit evaluation component relies on the LSTM neural network to analyse the sentiment of credit evaluation text directly and generate a credit evaluation score. The Blockchain component ensures the authenticity of the information of transaction and credit evaluation about traders in the food supply chain. Thus, the combination of these two components allows to strengthens the effectiveness of the management and supervision by generating and feeding back the credit evaluation result to regulators.

6. Conclusion

In this paper, we conducted fine-tuning experiments to identify classification baselines for sentiment analysis in Russian using Multilingual Bidirectional Encoder Representations from Transformers (Devlin et al., 2019), RuBERT (Kuratov & Arkhipov, 2019) and two versions of Multilingual Universal Sentence Encoder (Yang et al., 2020), the results are provided in Table 11. As a source data for experiments, we utilised seven sentiment datasets in Russian: SentiRuEval-2016 (Lukashevich & Rubtsova, 2016), SentiRuEval-2015 (Loukachevitch et al., 2015), RuSentiment (Rogers et al., 2018), Kaggle Russian News Dataset (Kaggle, 2017), LINIS Crowd (Koltsova et al., 2016), RuTweetCorp (Rubtsova, 2013), and RuReviews (Smetanin & Komarov, 2019).

The practical and academic contribution of this study is fourfold. Firstly, we identified the most commonly used sentiment analysis datasets of the Russian language texts. Secondly, for each of these datasets, we identified the current state-of-the-art sentiment analysis approach. Thirdly, we examined modern language models and outlined those of them which officially supports the Russian language. Finally, we fine-tuned language models on the selected datasets and achieved new state-of-the-art classification results on the half of sentiment analysis datasets. Considering the obtained results, we can state that in the context of existing approaches, sentiment analysis of the Russian language texts based on the language models outperforms rule-based and basic machine learning-based approaches in terms of classification quality. To provide further sentiment analysis studies with strong classification baselines, we made pre-trained Multilingual BERT-based, RuBERT-based, and Multilingual USE-based models publicly available[15] to the research community.

Future research could be focused on the usage of fine-tuned language models on applied tasks, e.g. on monitoring of sentiment index of social media content in Russian. Since fine-tuned models demonstrated the news SOTA results in most cases, they are potentially able to significantly increase sentiment classification quality and therefore improve the accuracy of the sentiment analysis outcomes. Within this direction, it can be extremely interesting not only to analyse the emotional component of the texts but also to automatically determine the age group and gender of the authors (e.g. based on public profile data or based on texts features) in order to obtain a more comprehensive picture of monitoring. Moreover, future research could be also focused on the pre-training of language models which currently does not support Russian language and future fine-tuning these models on sentiment analysis datasets.

[15] https://github.com/sismetanin/sentiment-analysis-in-russian

Table 11
Classification quality of fine-tuned models.

Dataset	Measure	Current SOTA	M-BERT-*	RuBERT-*	M-USE-CNN-*	M-USE-Trans-*
SentiRuEval-2016 TC (Lukashevich & Rubtsova, 2016)	$\mathbf{F_1}$	68.42%	66.29%	**70.68%**	63.64%	68.27%
	macro $\mathbf{F_1^{PN}}$	66.07%	61.78%	**66.40%**	58.97%	62.77%
	micro $\mathbf{F_1^{PN}}$	74.11%	72.45%	**76.71%**	71.31%	75.00%
SentiRuEval-2016 Banks (Lukashevich & Rubtsova, 2016)	$\mathbf{F_1}$	**74.06%**	65.31%	72.83%	66.71%	72.40%
	macro $\mathbf{F_1^{PN}}$	**69.53%**	58.00%	65.89%	58.73%	65.04%
	micro $\mathbf{F_1^{PN}}$	**71.76%**	60.52%	68.43%	62.41%	68.21%
SentiRuEval-2015 TC (Loukachevitch et al., 2015)	$\mathbf{F_1}$	**68.54%**	60.47%	64.39%	60.57%	64.28%
	macro $\mathbf{F_1^{PN}}$	**63.47%**	53.16%	57.76%	52.37%	57.60%
	micro $\mathbf{F_1^{PN}}$	**67.51%**	57.03%	61.38%	57.76%	61.18%
SentiRuEval-2015 Banks (Loukachevitch et al., 2015)	$\mathbf{F_1}$	**79.51%**	67.65%	70.58%	66.32%	69.62%
	macro $\mathbf{F_1^{PN}}$	**67.44%**	56.97%	60.95%	54.74%	59.12%
	micro $\mathbf{F_1^{PN}}$	**70.09%**	59.32%	63.33%	57.61%	62.17%
RuSentiment (Rogers et al., 2018)	$\mathbf{F_1}$	n/s	71.37%	**72.03%**	66.27%	68.60%
	weighted $\mathbf{F_1}$	**78.50%**	75.13%	75.71%	71.05%	73.42%
Kaggle Russian News Dataset (Kaggle, 2017)	$\mathbf{F_1}$	70.00%	71.36%	**73.63%**	71.27%	72.66%
LINIS Crowd (Koltsova et al., 2016)	$\mathbf{F_1}$	37.29%	42.73%	**60.51%**	56.34%	56.95%
RuTweetCorp Trinary (Rubtsova, 2013)	$\mathbf{F_1}$	75.95%	83.04%	**83.69%**	81.34%	83.17%
RuTweetCorp Binary (Rubtsova, 2013)	$\mathbf{F_1}$	78.1%	80.10%	**80.79%**	78.39%	79.69%
RuReviews (Smetanin & Komarov, 2019)	$\mathbf{F_1}$	75.45%	77.31%	**77.44%**	76.63%	76.94%

CRediT authorship contribution statement

Sergey Smetanin: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization, Project administration. **Mikhail Komarov:** Supervision, Resources, Writing - review & editing.

Acknowledgement

Project group was funded by the Graduate School of Business National Research University Higher School of Economics. This research was supported in part through resources of supercomputer facilities provided by NRU HSE.

References

Adaskina, Y. V., Panicheva, P., & Popov, A. (2015). Syntax-based sentiment analysis of tweets in Russian. In *Computational linguistics and intellectual technologies. Papers from the annual international conference dialogue 2015* (pp. 1–11).

Akbik, A., Blythe, D., & Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th international conference on computational linguistics COLING* (pp. 1638–1649).

Alekseev, A., & Nikolenko, S. (2016). User profiling in text-based recommender systems based on distributed word representations. In *International conference on analysis of images, social networks and texts* (pp. 196–207). Springer, http://dx.doi.org/10.1007/978-3-319-52920-2_19.

Alimova, I., Tutubalina, E., Alferova, J., & Gafiyatullina, G. (2017). A machine learning approach to classification of drug reviews in Russian. In *2017 ivannikov ISPRAS open conference* (pp. 64–69). IEEE, http://dx.doi.org/10.1109/ISPRAS.2017.00018.

Arkhipenko, K., Kozlov, I., Trofimovich, J., Skorniakov, K., Gomzin, A., & Turdakov, D. (2016). Comparison of neural network architectures for sentiment analysis of russian tweets. In *Computational linguistics and intellectual technologies. Papers from the annual international conference dialogue 2016* (pp. 50–59).

Banerjee, A., Mondal, S., Deb, A., & Ghosh, S. (2020). Decentralized policy feedback system for privacy and governance using blockchain and sentiment analysis for smart city applications. In *2020 international conference on computer science, engineering and applications* (pp. 1–6). http://dx.doi.org/10.1109/ICCSEA49143.2020.9132877.

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* (pp. 12–23). Florence, Italy: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/W19-4802.

Bartunov, S., Kondrashkin, D., Osokin, A., & Vetrov, D. (2016). Breaking sticks and ambiguities with adaptive skip-gram. In *Artificial intelligence and statistics* (pp. 130–138).

Basile, A., Franco-Salvador, M., Pawar, N., Štajner, S., Chinea Rios, M., & Benajiba, Y. (2019). SymantoResearch at SemEval-2019 task 3: Combined neural models for emotion classification in human-chatbot conversations. In *Proceedings of the 13th international workshop on semantic evaluation* (pp. 330–334). Minneapolis, Minnesota, USA: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/S19-2057.

Baymurzina, D., Kuznetsov, D., & Burtsev, M. (2019). Language model embeddings improve sentiment analysis in Russian. *Computational Linguistics and Intellectual Technologies*, *18*, 53–63, Papers from the Annual International Conference Dialogue 2019.

Baziotis, C., Pelekis, N., & Doulkeridis, C. (2017). DataStories at SemEval-2017 task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th international workshop on semantic evaluation* (pp. 747–754). Vancouver, Canada: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/S17-2126.

Camacho-Collados, J., & Pilehvar, M. T. (2018). From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, *63*(1), 743–788. http://dx.doi.org/10.1613/jair.1.11259.

Carosia, A. E. O., Coelho, G. P., & Silva, A. E. A. (2020). Analyzing the Brazilian financial market through portuguese sentiment analysis in social media. *Applied Artificial Intelligence*, *34*(1), 1–19. http://dx.doi.org/10.1080/08839514.2019.1673037.

Casino, F., Dasaklis, T. K., & Patsakis, C. (2019). A systematic literature review of blockchain-based applications: Current status, classification and open issues. *Telematics and Informatics*, *36*, 55–81. http://dx.doi.org/10.1016/j.tele.2018.11.006.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., et al. (2018). Universal sentence encoder. arXiv preprint arXiv:1803.11175.

Chatterjee, A., Narahari, K. N., Joshi, M., & Agrawal, P. (2019). SemEval-2019 task 3: EmoContext contextual emotion detection in text. In *Proceedings of the 13th international workshop on semantic evaluation* (pp. 39–48). Minneapolis, Minnesota, USA: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/S19-2005.

Chetvirokin, I., Braslavskiy, P., & Loukachevitch, N. (2012). Sentiment analysis track at ROMIP 2011. *Computational Linguistics and Intellectual Technologies*, *2*, 1–14, Papers from the Annual International Conference Dialogue 2012.

Chetvirokin, I., & Loukachevitch, N. (2013). Sentiment analysis track at ROMIP 2012. *Computational Linguistics and Intellectual Technologies*, *2*, 40–50, Papers from the Annual International Conference Dialogue 2013.

Chidambaram, M., Yang, Y., Cer, D., Yuan, S., Sung, Y., Strope, B., et al. (2019). Learning cross-lingual sentence representations via a multi-task dual-encoder model. In *Proceedings of the 4th workshop on representation learning for NLP* (pp. 250–259). Florence, Italy: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/W19-4330.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 670–680). Copenhagen, Denmark: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D17-1070.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Long and Short Papers*, *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N19-1423.

Enikeeva, E., & Popov, A. (2018). Developing a Russian database of regular semantic relations based on word embeddings. In *The XVIII EURALEX international congress* (p. 134).

Gao, Z., Feng, A., Song, X., & Wu, X. (2019). Target-dependent sentiment classification with BERT. *IEEE Access*, *7*, 154290–154299. http://dx.doi.org/10.1109/ACCESS.2019.2946594.

Garshina, V., Kalabukhov, K., Stepantsov, V., & Smotrov, S. (2017). Development of the system of sentiment analysis of the text. *Proceedings of Voronezh State University. Series: Systems analysis and information technologies*, *3*, 185–194.

Georgiadou, E., Angelopoulos, S., & Drake, H. (2020). Big data analytics and international negotiations: Sentiment analysis of Brexit negotiating outcomes. *International Journal of Information Management*, *51*, Article 102048. http://dx.doi.org/10.1016/j.ijinfomgt.2019.102048.

Golubev, A., & Loukachevitch, N. (2020). Improving results on Russian sentiment datasets. In A. Filchenkov, J. Kauttonen, & L. Pivovarova (Eds.), *Artificial intelligence and natural language* (pp. 109–121). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-030-59082-6_8.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. In *Proceedings of the 56th annual meeting of the association for computational linguistics (vol. 1)* (pp. 328–339). Melbourne, Australia: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P18-1031.

Iacus, S., Porro, G., Salini, S., & Siletti, E. (2020). An Italian composite subjective well-being index: The voice of Twitter users from 2012 to 2017. *Social Indicators Research*, 1–19. http://dx.doi.org/10.1007/s11205-020-02319-6.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., et al. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, *5*, 339–351. http://dx.doi.org/10.1162/tacl_a_00065.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th conference of the European chapter of the association for computational linguistics: Volume 2, short papers* (pp. 427–431). Association for Computational Linguistics.

Kaggle (2017). Sentiment analysis in Russian | kaggle. URL https://www.kaggle.com/c/sentiment-analysis-in-russian.

Kannengießer, N., Lins, S., Dehling, T., & Sunyaev, A. (2019). Mind the gap: trade-offs between distributed ledger technology characteristics. arXiv preprint arXiv:1906.00861.

Kannengießer, N., Lins, S., Dehling, T., & Sunyaev, A. (2020). Trade-offs between distributed ledger technology characteristics. *ACM Computing Surveys*, *53*(2), http://dx.doi.org/10.1145/3379463.

Karyaeva, M., Braslavski, P., & Kiselev, Y. (2018). Extraction of hypernyms from dictionaries with a little help from word embeddings. In *Analysis of images, social networks and texts* (pp. 76–87). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-030-11027-7_8.

Kasahara, S., & Kawahara, J. (2019). Effect of Bitcoin fee on transaction-confirmation process. *Journal of Industrial & Management Optimization*, *15*(1), 365. http://dx.doi.org/10.3934/jimo.2018047.

Khodak, M., Risteski, A., Fellbaum, C., & Arora, S. (2017). Automated WordNet construction using word embeddings. In *Proceedings of the 1st workshop on sense, concept and entity representations and their applications* (pp. 12–23). http://dx.doi.org/10.18653/v1/W17-1902.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 1746–1751). Doha, Qatar: Association for Computational Linguistics, http://dx.doi.org/10.3115/v1/D14-1181.

Kirekov, S., & Krajvanova, V. (2018). Comparative analysis of image classification and sentiment analysis tasks using neural networks. *Polzunovsky vestnik*, (4), 172–177.

Koltsova, O., Alexeeva, S., & Kolcov, S. (2016). An opinion word lexicon and a training dataset for Russian sentiment analysis of social media. *Computational Linguistics and Intellectual Technologies*, 227–287, Papers from the Annual International Conference Dialogue 2016.

Kudo, T., & Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 conference on empirical methods in natural language processing: System demonstrations* (pp. 66–71). Brussels, Belgium: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D18-2012.

Kuratov, Y., & Arkhipov, M. (2019). Adaptation of deep bidirectional multilingual transformers for Russian language. *Computational Linguistics and Intellectual Technologies*, *18*, 333–340, Papers from the Annual International Conference Dialogue 2019.

Kutuzov, A., & Andreev, I. (2015). Texts in, meaning out: Neural language models in semantic similarity task for Russian. In *Computational linguistics and intellectual technologies. Papers from the annual international conference dialogue 2015 (vol. 2)* (pp. 113–144).

Kutuzov, A., & Kuzmenko, E. (2017). WebVectors: A toolkit for building web interfaces for vector semantic models. In *Analysis of images, social networks and texts* (pp. 155–161). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-319-52920-2_15.

Lagutina, K., Larionov, V., Petryakov, V., Lagutina, N., & Paramonov, I. (2018). Sentiment classification of Russian texts using automatically generated thesaurus. In *2018 23rd conference of open innovations association* (pp. 217–222). http://dx.doi.org/10.23919/FRUCT.2018.8588096.

Lamport, L., Shostak, R., & Pease, M. (1982). The byzantine generals problem. *ACM Transactions on Programming Languages and Systems*, *4*(3), 382–401.

Li, X., Fu, X., Xu, G., Yang, Y., Wang, J., Jin, L., et al. (2020). Enhancing BERT representation with context-aware embedding for aspect-based sentiment analysis. *IEEE Access*, *8*, 46868–46876. http://dx.doi.org/10.1109/ACCESS.2020.2978511.

Litvinova, T., Sboev, A., & Panicheva, P. (2018). Profiling the age of Russian bloggers. In *Artificial intelligence and natural language* (pp. 167–177). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-030-01204-5_16.

Liu, Q., Kusner, M. J., & Blunsom, P. (2020). A survey on contextual embeddings. arXiv preprint arXiv:2003.07278.

Liu, R., Shi, Y., Ji, C., & Jia, M. (2019). A survey of sentiment analysis based on transfer learning. *IEEE Access*, *7*, 85401–85412. http://dx.doi.org/10.1109/ACCESS.2019.2925059.

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. Papers from the annual international conference dialogue 2015 (vol. 2)* (pp. 3–13).

Loukachevitch, N., & Levchik, A. (2016). Creating a general Russian sentiment lexicon. In *Proceedings of the tenth international conference on language resources and evaluation* (pp. 1171–1176). Portorož, Slovenia: European Language Resources Association (ELRA).

Loukachevitch, N., & Parkhomenko, E. (2018). Recognition of multiword expressions using word embeddings. In *Russian conference on artificial intelligence* (pp. 112–124). Springer, http://dx.doi.org/10.1007/978-3-030-00617-4_11.

Lukashevich, N., & Rubtsova, Y. R. (2016). SentiRuEval-2016: overcoming time gap and data sparsity in tweet sentiment analysis. In *Computational linguistics and intellectual technologies. Papers from the annual international conference dialogue 2016* (pp. 416–426).

Maas, A., Daly, R., Pham, P., Huang, D., Ng, A., & Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies (vol. 1)* (pp. 142–150). Stroudsburg, PA, USA: Association for Computational Linguistics.

Malykh, V., Alekseev, A., Tutubalina, E., Shenbin, I., & Nikolenko, S. (2019). Wear the right head: Comparing strategies for encoding sentences for aspect extraction. In *Analysis of images, social networks and texts* (pp. 166–178). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-030-37334-4_15.

Mao, D., Wang, F., Hao, Z., & Li, H. (2018). Credit evaluation system based on blockchain for multiple stakeholders in the food supply chain. *International Journal of Environmental Research and Public Health*, *15*(8), 1627. http://dx.doi.org/10.3390/ijerph15081627.

McCann, B., Bradbury, J., Xiong, C., & Socher, R. (2017). Learned in translation: Contextualized word vectors. In *Advances in neural information processing systems* (pp. 6294–6305).

Meškel, D., & Frasincar, F. (2020). ALDONAr: A hybrid solution for sentence-level aspect-based sentiment analysis using a lexicalized domain ontology and a regularized neural attention model. *Information Processing & Management*, *57*(3), Article 102211. http://dx.doi.org/10.1016/j.ipm.2020.102211.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th international conference on neural information processing systems (vol. 2)* (pp. 3111–3119). USA: Curran Associates Inc.

Mozetič, I., Grčar, M., & Smailović, J. (2016). Multilingual Twitter sentiment classification: The role of human annotators. *PLoS One*, *11*(5).

Natoli, C., & Gramoli, V. (2017). The balance attack or why forkable blockchains are ill-suited for consortium. In *2017 47th annual IEEE/IFIP international conference on dependable systems and networks* (pp. 579–590). IEEE, http://dx.doi.org/10.1109/DSN.2017.44.

Nenashev, M. (2019). Sentiment analysis of news articles. In *Proceedings of the L international scientific conference on control processes and stability* (pp. 326–330).

Panchenko, A. (2014). Sentiment index of the Russian speaking Facebook. *Computational Linguistics and Intellectual Technologies*, *2*, 506–517, Papers from the Annual International Conference Dialogue 2014.

Panchenko, A., Lopukhina, A., Ustalov, D., Lopukhin, K., Arefyev, N., Loukachevitch, N., et al. (2018). Russe'2018: A shared task on word sense induction for the Russian language. In *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii* (pp. 547–564).

Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting of the association for computational linguistics* (pp. 115–124). Ann Arbor, Michigan: Association for Computational Linguistics, http://dx.doi.org/10.3115/1219840.1219855.

Panicheva, P., Mirzagitova, A., & Ledovaya, Y. (2017). Semantic feature aggregation for gender identification in Russian Facebook. In *Conference on artificial intelligence and natural language* (pp. 3–15). Springer, http://dx.doi.org/10.1007/978-3-319-71746-3_1.

Pei, S., Wang, L., Shen, T., & Ning, Z. (2019). DA-BERT: Enhancing part-of-speech tagging of aspect sentiment analysis using BERT. In P.-C. Yew, P. Stenström, J. Wu, X. Gong, & T. Li (Eds.), *Advanced parallel processing technologies* (pp. 86–95). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-030-29611-7_7.

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 1532–1543). Doha, Qatar: Association for Computational Linguistics, http://dx.doi.org/10.3115/v1/D14-1162.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., et al. (2018). Deep contextualized word representations. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (pp. 2227–2237). New Orleans, Louisiana: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N18-1202.

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., et al. (2016). SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation* (pp. 19–30). San Diego, California: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/S16-1002.

Popov, D., Pugachev, A., Svyatokum, P., Svitanko, E., & Artemova, E. (2019). Evaluation of sentence embedding models for natural language understanding problems in Russian. In *International conference on analysis of images, social networks and texts* (pp. 205–217). Springer, http://dx.doi.org/10.1007/978-3-030-37334-4_19.

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, http://dx.doi.org/10.1007/s11431-020-1647-3.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. OpenAI Blog.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog.

Read, J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL student research workshop* (pp. 43–48). Ann Arbor, Michigan: Association for Computational Linguistics.

Rodina, J., Bakshandaeva, D., Fomin, V., Kutuzov, A., Touileb, S., & Velldal, E. (2019). Measuring diachronic evolution of evaluative adjectives with word embeddings: The case for english, norwegian, and Russian. In *Proceedings of the 1st international workshop on computational approaches to historical language change* (pp. 202–209). Florence, Italy: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/W19-4725.

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* (pp. 755–763). Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Romanov, A., Vasilieva, M., Kurtukova, A., & Meshcheryakov, R. (2017). Sentiment analysis of text using machine learning techniques. In *Proceedings of the R. Piotrowski's readings in language engineering and applied linguistics* (pp. 86–95).

Rubtsova, Y. (2013). A method for development and analysis of short text corpus for the review classification task. In *Proceedings of conferences digital libraries: Advanced methods and technologies, digital collections* (pp. 269–275).

Rubtsova, Y. (2018). Reducing the deterioration of sentiment analysis results due to the time impact. *Information*, *9*, 184. http://dx.doi.org/10.3390/info9080184.

Ruder, S., Peters, M. E., Swayamdipta, S., & Wolf, T. (2019). Transfer learning in natural language processing. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Tutorials* (pp. 15–18). Minneapolis, Minnesota: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N19-5004.

Ruseti, S., Sirbu, M.-D., Calin, M. A., Dascalu, M., Trausan-Matu, S., & Militaru, G. (2020). Comprehensive exploration of game reviews extraction and opinion mining using NLP techniques. In X.-S. Yang, S. Sherratt, N. Dey, & A. Joshi (Eds.), *Fourth international congress on information and communication technology* (pp. 323–331). Singapore: Springer Singapore, http://dx.doi.org/10.1007/978-981-15-0637-6_27.

Rusnachenko, N., & Loukachevitch, N. (2018). Extracting sentiment attitudes from analytical texts via piecewise convolutional neural network. In *Proceedings of XX international conference on data analytics and management in data intensive domains* (pp. 186–192).

Rybakov, V., & Malafeev, A. (2018). Aspect-based sentiment analysis of Russian hotel reviews. In *Supplementary proceedings of the seventh international conference on analysis of images, social networks and texts* (pp. 75–84).

Sboev, A., Litvinova, T., Gudovskikh, D., Rybka, R., & Moloshnikov, I. (2016). Machine learning models of text categorization by author gender using topic-independent features. *Procedia Computer Science*, *101*, 135–142. http://dx.doi.org/10.1016/j.procs.2016.11.017, 5th International Young Scientist Conference on Computational Science, YSC 2016, 26–28 October 2016, Krakow, Poland.

Shalkarbayuli, A., Kairbekov, A., & Amangeldi, Y. (2018). Comparison of traditional machine learning methods and Google services in identifying tonality on Russian texts. *Journal of Physics: Conference Series*, *1117*, Article 012002. http://dx.doi.org/10.1088/1742-6596/1117/1/012002.

Sharma, U., Datta, R. K., & Pabreja, K. (2020). Sentiment analysis and prediction of election results 2018. In R. K. Shukla, J. Agrawal, S. Sharma, N. S. Chaudhari, & K. K. Shukla (Eds.), *Social networking and computational intelligence* (pp. 727–739). Singapore: Springer Singapore, http://dx.doi.org/10.1007/978-981-15-2071-6_61.

Smetanin, S. (2020). The applications of sentiment analysis for Russian language texts: Current challenges and future perspectives. *IEEE Access*, *8*, 110693–110719. http://dx.doi.org/10.1109/ACCESS.2020.3002215.

Smetanin, S., & Komarov, M. (2019). Sentiment analysis of product reviews in Russian using convolutional neural networks. In *2019 IEEE 21st conference on business informatics (vol. 1)* (pp. 482–486). http://dx.doi.org/10.1109/CBI.2019.00062.

Smetanin, S., Ometov, A., Kannengießer, N., Sturm, B., Komarov, M., & Sunyaev, A. (2020). Modeling of distributed ledgers: Challenges and future perspectives. In *2020 IEEE 22nd conference on business informatics (vol. 1)* (pp. 162–171). http://dx.doi.org/10.1109/CBI49978.2020.00025.

Smetanin, S., Ometov, A., Komarov, M., Masek, P., & Koucheryavy, Y. (2020). Blockchain evaluation approaches: State-of-the-art and future perspective. *Sensors*, *20*(12), 3358. http://dx.doi.org/10.3390/s20123358.

Smirnova, O., & Shishkov, V. (2016). The choice of the topology of neural networks and their use for the classification of small texts. *International Journal of Open Information Technologies*, *4*(8), 50–54.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., et al. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1631–1642). Seattle, Washington, USA: Association for Computational Linguistics.

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune BERT for text classification? In *Chinese computational linguistics* (pp. 194–206). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-030-32381-3_16.

Sunyaev, A. (2020). Distributed ledger technology. In *Internet computing: Principles of distributed systems and emerging internet-based technologies* (pp. 265–299). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-030-34957-8_9.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, *61*(12), 2544–2558. http://dx.doi.org/10.1002/asi.21416.

Tutubalina, E., Alimova, I., Miftahutdinov, Z., Sakhovskiy, A., Malykh, V., & Nikolenko, S. (2020). The Russian Drug Reaction Corpus and neural models for drug reactions and effectiveness detection in user reviews. *Bioinformatics*, http://dx.doi.org/10.1093/bioinformatics/btaa675, btaa675.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

Wang, C., Xu, Y., & Wang, Q. (2018). *Novel approaches to sentiment analysis for stock prediction*. Stanford University.

Weber, I., Gramoli, V., Ponomarev, A., Staples, M., Holz, R., Tran, A. B., et al. (2017). On availability for blockchain-based systems. In *2017 IEEE 36th symposium on reliable distributed systems* (pp. 64–73). IEEE, http://dx.doi.org/10.1109/SRDS.2017.15.

Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., et al. (2020). Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th annual meeting of the association for computational linguistics: System demonstrations* (pp. 87–94). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.acl-demos.12.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, 5754–5764.

Yang, L., Li, Y., Wang, J., & Sherratt, R. S. (2020). Sentiment analysis for E-commerce product reviews in chinese based on sentiment lexicon and deep learning. *IEEE Access*, *8*, 23522–23530. http://dx.doi.org/10.1109/ACCESS.2020.2969854.

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. In *Proceedings of the 28th international conference on neural information processing systems (vol. 1)* (pp. 649–657). Cambridge, MA, USA: MIT Press.

Zhidanov, K., Bezzateev, S., Afanasyeva, A., Sayfullin, M., Vanurin, S., Bardinova, Y., et al. (2019). Blockchain technology for smartphones and constrained IoT devices: A future perspective and implementation. In *Proc. of 21st conference on business informatics (vol. 2)* (pp. 20–27). IEEE, http://dx.doi.org/10.1109/CBI.2019.10092.

Zvonarev, A., & Bilyi, A. (2019). A comparison of machine learning methods of sentiment analysis based on Russian language Twitter data. In *Proceedings of the 11th majorov international conference on software engineering and computer systems*. Saint Petersburg, Russia: ITMO University.