The Applications of Sentiment Analysis for Russian Language Texts: Current Challenges and Future Perspectives

SERGEY SMETANIN [iD]

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

e-mail: sismetanin@gmail.com

ABSTRACT Sentiment analysis has become a powerful tool in processing and analysing expressed opinions on a large scale. While the application of sentiment analysis on English-language content has been widely examined, the applications on the Russian language remains not as well-studied. In this survey, we comprehensively reviewed the applications of sentiment analysis of Russian-language content and identified current challenges and future research directions. In contrast with previous surveys, we targeted the applications of sentiment analysis rather than existing sentiment analysis approaches and their classification quality. We synthesised and systematically characterised existing applied sentiment analysis studies by their source of analysed data, purpose, employed sentiment analysis approach, and primary outcomes and limitations. We presented a research agenda to improve the quality of the applied sentiment analysis studies and to expand the existing research base to new directions. Additionally, to help scholars selecting an appropriate training dataset, we performed an additional literature review and identified publicly available sentiment datasets of Russian-language texts.

INDEX TERMS Classification, machine learning, computational linguistics, sentiment analysis, applications of sentiment analysis, Russian-language texts, public opinion.

I. INTRODUCTION

Sentiment analysis is a subsection of natural language processing whose objective is to classify a text by the sentiment it contains. By analysing a texts collection, scholars can summarise expressed sentiment and then get insights regarding different topics. For instance, sentiment analysis can be used for predicting the stock market (e.g. [1]), computing the Subjective Well-Being Index (e.g. [2]), predicting election results (e.g. [3]), and measuring reactions to particular events or news (e.g. [4]). While the applications of sentiment analysis were widely examined for English language content [5]–[7], Non-English language content, and especially Russian, received much less attention from scholars. Globally, Russian is ranked as the eighth language by the total number of speakers [8]. According to the Omnibus GFK survey [9], internet penetration in Russia exceeds 75.4% comprising 90 million Russians aged 16 or more. Russian language sites are distributed on all continents, but the largest concentration is in the Commonwealth of Independent States (CIS) and, especially, in Russia and Ukraine. According to a study by W3Techs,[1] Russian is the second most widespread language on the Internet after English. As of April 2020, 8.6% of the 10 million most popular Internet sites in the world use Russian. Thus, Russian content becomes a valuable source of data for automatic analysis, especially in the field of sentiment analysis.

Only one survey [10] by Viksna and Jekabsons directly addressed the sentiment analysis of Russian content, and several others [11]–[14] mentioned sentiment analysis of Russian in the contexts of overall comparison with globally existing approaches. There were other studies dedicated to specific aspects of sentiment analysis of Russian, for instance, the evaluation of state-of-the-art approaches [15]–[18], comparison of neural network architectures for sentiment analysis [19], [20], and comparison of publicly available Russian sentiment lexicons [21]. However, in all cases, previous studies focused on the sentiment

The associate editor coordinating the review of this manuscript and approving it for publication was Yucong Duan [iD].

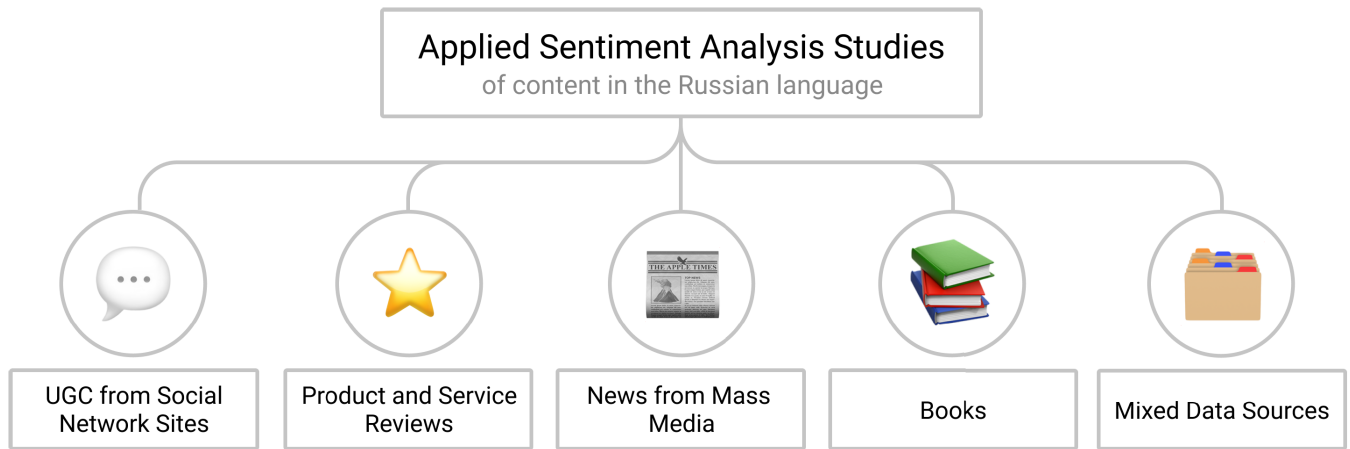[1] https://w3techs.com/technologies/overview/content_language

FIGURE 1. Proposed categories.

analysis approaches and their classification performance rather than their applications and obtained outcomes of data analysis.

In this survey, we comprehensively reviewed the applications of sentiment analysis of the Russian language content and identified current challenges and future research directions. We considered only those studies, which obtained valuable outcomes based on sentiment analysis of the real-life data, and not considered those, which just trained sentiment classifiers. By conducting the survey, we pursued several objectives simultaneously. Firstly, we hoped to provide scholars with a balanced and deep understanding of existing sentiment analysis applications dealing with Russian content. Secondly, we identified the current challenges and proposed ways to overcome them. Lastly, we introduced potential research opportunities to suggest research directions for further studies.

This survey differs from existing literature surveys in various dimensions. Firstly, targeted the applications of sentiment analysis rather than existing sentiment analysis approaches and their classification quality. We focused on the applications of sentiment analysis of the Russian language content since it was not reviewed before. Secondly, we categorised existing studies based on the utilised data source, as described in Figure 1. Thirdly, we summarised each of the surveyed papers in several directions: the problem addressed, the process of data retrieving, exploited data details, implemented data annotation process (if applied), pre-processing techniques applied (if applied), sentiment analysis approach applied, and obtained findings. Additionally, we indicated current drawbacks and future directions. The summary of analysed articles is presented in Table 2. Fourthly, we identified all publicly available sentiment analysis datasets of Russian-language texts presented in Table 1. Lastly, we summarised current challenges and proposed future research directions.

The rest of the article is organised as follows. Section 2 briefly describes the sentiment analysis task and

current approaches. Section 3 gives an overview of the proposed method for conducting this survey. Section 4 provides an overview of the identified applications of sentiment analysis for Russian-language content. Section 5 focuses on current challenges. Section 6 describes future research opportunities of identified studies. In section 7, the conclusion and main contributions of the survey are drawn.

II. BACKGROUND
Sentiment analysis is a subtask of natural language processing whose key objective is to classify a text in accordance with the sentiment it contains. Basic approaches usually aim at a binary classification of texts as either positive or negative. In some cases, the classification scale includes one more class of neutral texts. More advanced approaches try to identify emotional states associated with the written text, for instance, fear, anger, sadness or happiness. Alternatively, a group of approaches target associate texts with a number on the predefined scale, e.g. from -2 for negative texts to 2 for positive texts, thereby narrowing down a sentiment analysis to a regression task. Aspect-based sentiment analysis is the subsection of sentiment analysis, where the main task is to identify the sentiment towards a specific aspect associated with a given target. The sentiment analysis approaches can be broadly divided into three types.

The first one is **rule-based approaches**, which are primarily based on manually-defined classification rules and sentiment lexicons. These rules usually utilise emotional keywords and their co-occurrences in the texts with other keywords [22]–[24]. Despite the excellent performance within a narrow domain, rule-based methods suffer from a poor generalization ability. Besides, they tend to be extremely labour-intensive to create, especially in cases where there is no access to appropriate sentiment lexicons. The latter is especially relevant for the Russian language since generally, it is not as well-resourced as the English language, especially in the field of sentiment analysis. RuSentiLex [25] and LINIS Crowd [26] sentiment lexicons are the largest lexicons for the

Russian language. However, they contain information only about sentiment level from positive to negative without any emotion-specific characteristics. Therefore, there are no alternatives to such powerful English lexicons with strong emotional characteristics as SenticNet [27], SentiWordNet [28], and SentiWords [29].

The second type is **machine learning-based approaches**, which rely on automatically extracting features from text and then applying machine learning classification algorithms. Naive Bayes Classifier [30], Decision Tree [31], Logistic Regression [32], and Support Vector Machine [33] can be defined as basic algorithms for polarity classification. In recent years, deep learning techniques have captured the attention of researchers due to their ability to significantly outperform traditional methods in the sentiment analysis task [34]. This fact has also been confirmed by a chronology of SemEval competition, where leading solutions successfully used convolutional neural networks (CNNs) and recurrent neural networks (RNNs) [35]–[37] as well as transfer learning techniques [38]. One of the key characteristics of machine learning systems is an automatic feature extraction from texts. Straightforward approaches usually use the Bag of Words model for representing texts in vector space. More sophisticated systems utilise distributive semantic models to generate word embeddings, for example, Word2Vec [39], GloVe [40], and FastText [41]. There are also methods that generate embeddings on the sentence or paragraph levels and specifically target transfer learning to a variety of natural language processing task, for instance, ELMo [42], Universal Sentence Encoder (USE) [27], Bidirectional Encoder Representations from Transformers (BERT) [43], Enhanced Language Representation with Informative Entities (ERNIE) [44], and XLNet [45]. One of the major drawbacks exploiting these systems for embeddings generation is in the fact that they require large corpora of texts for training. In general, this issue is relevant to all machine learning techniques, because all supervised classification algorithms require annotated datasets for training.

The third one is **hybrid approaches**, which combine both rule-based and machine learning-based approaches. For example, Kumar and colleagues developed a hybrid Persian sentiment analysis framework, which integrated linguistic rules and CNN and LSTM modules to classify sentiment [46]. Meskele and Frasincar proposed the hybrid model for aspect-based sentiment analysis, ALDONAr, which combines a sentiment domain ontology for capturing domain information, BERT for obtaining word embeddings, and two CNN layers for enhancing sentiment classification [47]. The model achieved accuracy scores of 83.8% and 87.1% on SenEval 2015 Task 12 [48] and SemEval 2016 Task 5 [49] datasets, respectively. Similar to the rule-based approaches, language models tend to be widely used in hybrid approaches [50]–[52]. One the one hand, the combination of strengths of rule-based and machine learning-based approaches commonly allows to deliver more accurate results. One the other hand, as a consequence of

the combination, hybrid approaches also receive challenges and limitations of rule-based and machine learning-based approaches.

III. METHODOLOGY

In order to identify key publications on the applied sentiment analysis of Russian-language content, we performed a literature search in scientific databases that cover leading computer science journals and conferences: *IEEE Xplore*, *ACM Digital Library*, *ScienceDirect*, *SAGE Journals Online*, and *Springer Link*. In addition to English-language articles, we analysed high-quality articles from *Russian Science Citation Index* written in Russian in order to cover a broad set of publications. To find relevant articles and papers for our research, we applied the following search string: ((*"SENTIMENT " OR "POLARITY") AND ("ANALYSIS" OR "DETECTION" OR "CLASSIFICATION" OR "OPINION MINING" OR "TOPIC MODELING") AND ("RUSSIAN" or "RUSSIA"))*. Most of the relevant papers were derived from *ScienceDirect*, *Springer Link*, and *Russian Science Citation Index*. In addition to the comprehensive database search, we also included pre-prints by leading researchers in the field in our literature sample to cover the latest advances. Following this approach, we gathered a total of 4,041 potentially relevant publications, excluding grey literature and pre-prints. Considering first of all the most recent and cited works, we then analysed the title, keywords, and abstracts of remaining publications to further narrow down the literature sample. We limited our search to peer reviewed articles to ensure a high quality of selection. We also excluded articles that were grey literature (i.e., unfinished manuscripts, editorials, any kind of dissertations) or not applicable to our study (i.e. without applying sentiment classification model to any sentiment monitoring task). In doing so, we manually selected 32 main publications that described at least one applied sentiment monitoring approach for Russian-language content for a detailed description in this paper.

Once the literature selection process was completed, we carefully read the selected publications. Nexts, we used an open coding approach to identify the application of sentiment analysis approach as well as details including the approaches' functioning, underlying models, assumptions, and intended purposes. Then, we classified the extracted approaches into four categories based on the source of the analysed data: *User-Generated Content from Social Network Sites*, *Product and Service Reviews*, *News from Mass Media*, *Books*, and *Mixed Data Sources*. The results of our analysis are presented in the next section.

IV. THE APPLICATIONS OF SENTIMENT ANALYSIS FOR RUSSIAN LANGUAGE TEXTS

In our literature review, we identified five categories of applied sentiment monitoring studies on Russian-language content, which will be presented in the following order: *User-Generated Content from Social Network Sites*, *Product and Service Reviews*, *News from Mass Media*, *Books*, and

Mixed Data Sources. We decided to categorise approaches by the utilised data source because these approaches commonly share similar research goals, challenges, and limitations. Even even though some categories consist only of one study, we considered it important to put them into separate categories because they fundamentally differ in goals, findings and common challenges. Moreover, it should be borne in mind that the Russian language is not as well-studied as the English language in the field of sentiment monitoring, so we were limited to the existing number of research papers. We additionally divided some of these categories into subcategories that describe different directions and the purposes of the studies. A roadmap of applied sentiment monitoring studies of Russian-language content presented at Fig. 1. In general, most of the identified approaches focus on analysing social networks data in order to outline user attitudes to different topics, for instance, attitudes to ethics and migrants issues or opinions regarding conflicts in Ukraine.

In the recent decade, a wide range of social network sites has become integral to modern ways of social engagement [53], which can be perceived as an essential and openly available source of public opinion, or at least its reasonable proxy [54]. As the most common data source, user-generated content from social networks was explored in three directions: attitudes to different topics, social sentiment indexes and specifics of user interaction with content expressing different sentiment. Attitudes to different topics have been examined from various perspectives such as identifying of attitudes towards migrants and ethnic groups (e.g. [55]), examining expressed sentiment during Ukrainian Crisis (e.g. [56]), measuring the level of social tension (e.g. [57]) or by focusing on the discourse on other significant topics (e.g. [58]). These approaches commonly utilised a combination of topic modelling and sentiment analysis techniques to extract relevant topics and corresponding sentiments. A significant share of studies (e.g. [59]–[67]), which applied topic modelling techniques without a subsequent polarity classification (and, therefore, were not considered in this study), outlined the usage of sentiment analysis as a further way of development. The other subcategory of studies (e.g. [68]) computed social sentiment indexes based on the opinions expressed in social networks to get a sort of alternative to the traditional Index of Subjective Well-Being. Besides, some studies (e.g. [69]) also examined patterns of user interaction with content depending on its sentiments. One of the critical challenges of these studies was to retrieve a representative data sample and to filter relevant texts for the analysis.

The next most common data source is product and service reviews, which were analysed in terms of characteristics of reviewers (e.g. [70]), characteristics of products and services (e.g. [71]), and characteristics of merchants (e.g. [72]). In contrast to the analysis of UGC content from social networks, there were no challenges regarding access to historical data. Since review platforms commonly provide users with an ability to assign a rating to their reviews, technically there is

no need to create a sentiment classification model, because reviews' sentiment classes are known beforehand. However, some studies implemented sentiment classification models just in the context of academic interest anyway.

While UGC content in social networks and user reviews commonly represents subjective texts, since authors express their opinions freely, the situation is different in the context of news analysis. Generally, journalists try to avoid making judgments and overt partiality and steer clear of doubt and ambiguity, since objectivity, or at least widely acceptable neutrality, is their philosophical basis [73]. As a consequence, a journalist often abstains from using negative or positive vocabulary, but resort to other ways to express their opinion [74]. News from mass media became the third most common data source, which was studied in two directions: the analysis of sentiments expressed in news articles (e.g. [75]) and constructing economic and business forecasts based on the sentiment of the news (e.g. [76]). In contrast with the analysis of social networks content, there were no challenges regarding the access to the historical data, because mass media commonly holds no restrictions to the access to all published data. However, some studies attempted to identify the public attitude to specific topics, which, to our mind, required further elaboration. Mass media, of course, are supposed to be a proxy of public opinion. However, in some cases, the publisher's policy may affect the presentation form, so the news does not always represent the society opinion.

Slightly less attention of the scholars was given to most new research directions, the sentiment analysis of textbooks, studies on which appeared only in 2019. These studies were focused on the comparison of the sentiment expressed in different textbooks (e.g. [77]) and on the influence of the sentiment of textbooks on the educational process (e.g. [78]). The critical challenge for this group of studies lies in the absence of the sentiment lexicons and training datasets within the target domain of educational textbooks. Moreover, in case of analysis texts on the document-level, it is becoming hard to associate texts with the corresponding sentiment class, since texts in textbooks are long, and throughout the texts authors may express different emotions.

To cover a broader range of opinions, some studies also utilised mixed data sources. In this group, scholars commonly examined attitudes to different topics such as Ukrainian Crisis (e.g. [79]) or media coverage of Alexei Navalny (e.g. [80]). Since these studies utilised a mixture of data sources, potentially this type of data can be used in all possible research directions. However, as an addition to a broad range of expressed opinions, authors also received all source-specific challenges and limitations.

The overview of the identified approaches is presented in Table 1. Based on the year-wise distribution of articles identified for the survey, it can be observed that applied studies on sentiment analysis of the Russian-language content proliferated during 2014-2016 years and reached a maximum number of studies in 2017. The number of articles published in the same journals and conference proceedings varies

TABLE 1. Overview of the identified studies.

Category	Purpose	Description	Reference	SA Approach	SA Level
UGC	Attitudes to topics	Identifying attitudes towards ethnic groups and migrants	[81]	ML (Logit)	DL
			[82]	ML (Logit)	DL
			[83]	ML (Logit)	DL
			[84]	RB (SentiStrength)	DL
			[55]	ML (SVM)	DL
		Identifying attitudes during the Ukrainian Crisis	[85]	RB (custom)	DL
			[86]	RB (POLYARNIK)	DL
			[87]	RB (SentiMental)	DL
			[88]	UNK (IQBuzz)	DL
			[56]	RB (custom)	DL
		Measuring a level of social tension	[89]	ML (SVM)	DL
			[57]	RB (SentiStrength)	DL
		Examining reaction to the meteor explosion in Chelyabinsk	[58]	n/s	DL
		Measuring reaction to Sochi 2014 Olympics	[90]	RB (SentiStrength)	DL
		Examining mass protests in Russia between 2011 and 2012	[91]	RB (SentiStrength)	DL
		Distribution of emotions in Saint Petersburg	[92]	ML (NBC)	DL
	Social Sentiment Index	Constructing the Index of Subjective Well-Being	[93]	RB (custom)	WL, DL
			[68]	ML (GBM)	DL
	User Behaviour	Measuring of the impact of sentiment on the mechanisms of feedback from the audience.	[69]	ML (BiGRU)	DL
Reviews	Characteristics of Reviewers	Identifying reasons why employees leave Russian companies	[70]	n/s	DL
	Characteristics of Products and Services	Evaluating road pavement assessment of the Northwestern Federal District of Russia	[71]	ML (NB, SGD)	DL
	Characteristics of Merchants	Identifying sellers' product quality	[72]	ML (RNTN)	DL
News	Content of News	Identifying hot topics and polarity of media coverage of news	[94]	RB (custom)	DL
			[95]	RB (custom)	DL
		Examining sentiment coverage of technologies and innovations mentioned in the mass media	[96]	RB (custom)	DL
		Comparing the networked issue agendas of Vladimir Putin and Alexey Navalny	[75]	UNK (Medialogia)	DL
	Economic and Business Forecasts	Constructing a high-frequency indicator of economic activity in Russia	[76]	ML (SVM)	DL
Books	Content of books	Comparing the sentiment expressed in Russian textbooks on Social Studies and History	[77]	RB (custom)	WL
	Educational process	Measuring the correlation between the sentiment of educational texts, a subjective assessment by the international students, and the real success of educational process.	[78]	ML (n/s)	DL
Mixed	Attitudes to topics	Identifying attitudes during the Ukrainian Crisis	[97]	UNK (Crimson Hexagon)	DL
			[79]	UNK (Crimson Hexagon)	DL
		Analysing the intensity and sentiment of the media coverage of Alexei Navalny	[80]	UNK (Medialogia)	DL

RB	–	rule-based approaches	n/s	–	not specified	SGB	–	Stochastic Gradient Descent	Logit	–	Logistic Regression
ML	–	machine learning-based approaches	NB	–	Naive Bayes	RNTN	–	Recursive Neural Tensor Network	DC	–	document-level
UNK	–	unknown approaches	MNB	–	Bidirectional Gated Recurrent Units	SVM	–	Support Vector Machine	WC	–	word-level

slightly, while there are only seven journals and proceedings where more than one of the analysed articles were published. In the conference proceedings "Digital Transformation and Global Society" was published a maximum number of identified articles.

The proportion between rule-based (40.63%) and machine learning-based approaches (37.5%) was almost equal, with a slight predominance of the first one. Among rule-based approaches, scholars usually employed custom rule-based models or applied SentiStrength [22], which became an absolute leader in terms of usage frequency in the context of third-party out-of-the-box solutions. As for machine learning-based approaches, Logistic Regression [32], Support Vector Machine [33], and Naive Bayes [30] were the most common choices. Most attention was given to basic machine learning approaches, while applications of neural networks account only to 16.7% among machine learning-based approaches. Until 2018, the share of rule-based approaches was higher or at the same level as the share of machine learning-based approaches. However, since 2019, the percentage of machine learning-based

approaches has significantly exceeded the share of rule-based approaches. Additionally, 15.6% of identified studies used third-party cloud services for sentiment analysis, e.g. Medialogia, IQBuzz, and Crimson Hexagon. In these cases, we were unable to identify sentiment analysis approaches, since these services did not disclose information about implemented classification algorithms.

It was also found that several of the identified studies suffer from a set of methodological drawbacks, including lack of the description of preprocessing, data annotation and training phrases as well as classification quality metrics. In some cases, validation of the classification model on the dataset from the target domain was not performed at all. This is particularly relevant for sentiment analysis using rule-based approaches or third-party services since, in this case, scholars usually did not perform manual annotation of texts collection, and therefore were not able to evaluate classification quality.

A substantial proportion of studies utilised local Russian social network sites and data aggregation platforms, a brief description of the most common local and

international resources and corresponding usage statistics is provided below.

- **VKontakte**[2] is a Russian online social media site available in more than 90 languages, but it is predominantly used by Russian-speakers. According to Deloitte's report [98], VKontakte is the most popular internet resource in Russia, which is used by up to 70% of Russia's population. VKontakte is widespread among the youngest audience of 16-24 years old, and as the age of respondents increases, their significance decreases.

- **YouTube**[3] is a video hosting site with social network features, which provides users with services of storage, delivery, and display of video. According to Deloitte's report [98], YouTube is the second most popular internet resource in Russia, which is used by up to 62% of Russia's population. YouTube is widespread among the youngest audience of 16-24 years old, and as the age of respondents increases the share of age groups varies between 58% and 64%.

- **Twitter**[4] is an international social network for public microblogging. According to Deloitte's report [98], Twitter is in the top-10 most popular internet resources in Russia, which uses up to 5% of Russia's population. The use of Twitter is almost evenly distributed in age groups from 25 to 65+ years with the peak in the 55-64 years age group.

- **LiveJournal**[5] is an international blog platform for maintaining online diaries, as well as separate personal blogs. According to Deloitte's report [98], LiveJournal is in the top-10 most popular internet resources in Russia, which is used by up to 3% of Russia's population. The use of LiveJournal is more common among Russians aged 35-44 years, as well as older generations.

- **Medialogia**[6] is a Russian company developing an automatic system for monitoring and analysing media and social networks in real-time. Medialogia automatically processes 500 thousand media messages and 100 million social media messages per day. The system aggregates content from 52,000 media sources and 900 million social network accounts.

- **IQBuzz**[7] is the monitoring service, which processes information from more than 10,000 sources of online media, Facebook, Twitter, VKontakte, My World, Instagram, 4sq, LiveJournal, LiveInternet, Google+, YouTube, RuTube and other sources. The system allows to automatically identify positive and negative messages, control duplicate messages, and provides a powerful search on the history of messages.

In the following sections, we describe the identified studies and their outcomes. We provide the views and findings of

the authors of the identified studies and not necessarily our position.

A. USER-GENERATED CONTENT FROM SOCIAL NETWORK SITES

Nowadays, a wide range of social network sites has become integral to modern ways of social engagement [53]. The user-generated content can be perceived as an important and openly available source of public opinion, or at least its reasonable proxy, which can supplement and sometimes substitute opinion polls [54]. The applied sentiment analysis studies on user-generated content from social network sites can be broadly divided into three subcategories: studies examining user attitudes to different topics, studies computing social sentiment indexes, and studies exploring user behaviour patterns of interaction with content.

1) ATTITUDES TO TOPICS

Among studies dedicated to the Russian-language content, interethnic and migrant issues, as well as the Ukrainian Crisis, were the most commonly explored research directions. Considerable attention was also given to social tension analysis and other topics examination.

a: ETHNIC GROUPS AND MIGRANTS

Interethnic relations, as well as migrants issues and related studies, have been deeply examined through well-developed methodologies of social science. Nevertheless, the rapid development of the Internet and natural language processing techniques allowed researchers to apply new options to conduct studies. Social media allows both individuals and groups to cooperate or engage in conflict, being highly visible for the Internet users. The dissemination of judgments about migrants issues and ethnic groups on the Internet may potentially progress faster and reach a more broad audience than before the Internet epoch [54]. Moreover, academic studies have proven that negative online content impacts on the offline interethnic conflicts [99] and hate crimes [100]. Thus, the task of ethnicity and migrants' issues analysis based on online content is becoming increasingly important as the internet technology develops.

The research conducted by Bodrunova *et al.* explored the attitudes of the Russian speaking online community to migrants in public discourse [81]. The authors collected a dataset of 363,579 posts from the 4th of February to the 19th of May 2013, by the top 2,000 Russian bloggers. Using previously reported strategy [59], [101], they applied topic modelling to identify relevant discussions, using Latent Dirichlet Allocation (LDA) [102]. Next, they performed a series of manual annotations of selected discussions, including sentiment annotation. Finally, they trained Binomial Logistic Regression [32] for a range of text classification tasks, including the sentiment classification. It was found that all migrants were perceived negatively, especially North Caucasians, in comparison with Central Asians and Americans. At the same time, neither Europeans or Americans

---

[2] https://vk.com/

[3] https://www.youtube.com/

[4] https://twitter.com/

[5] https://www.livejournal.com/

[6] https://www.mlg.ru/

[7] http://iqbuzz.pro/

were covered positively. While Europeans, North Caucasians, and Americans were never described as victims, they were commonly described as aggressors. Central Asians were described by users as an alien with a positive connotation, while North Caucasians and Americans were alien with a negative connotation. In general, Europeans were perceived as neither alien nor partners, Americans were described as dangerous, and Jews were perceived as completely non-dangerous. Finally, the authors claimed that the mental, post-Soviet divisions do not fully correspond to the current formal national borders since previously close groups are already perceived as distant nations with their political agendas. One of the main drawbacks of this work is the lack of data annotation quality assessment and classification metrics specification.

In the paper [82], Koltsova et.al. measured the overall volume of ethnicity-related discussion in the Russian content of social media sites, adapting developed techniques from previous works [103], [104]. To construct an initial corpus of texts, they developed a comprehensive list of ethnonyms and related bigrams, embracing 97 ethnic groups from Post-Soviet territories, collecting 2,660,222 texts. Next, the authors created a training dataset via manual annotation of 7181 texts, where each text was annotated by three assessors in several aspects, including the presence of intergroup conflict, positive intergroup contact, and overall positive and negative sentiment. To perform the sentiment classification, the Logistic Regression [32] was trained on the annotated dataset, achieving $F_1 = 0.75$ for a positive sentiment and $F_1 = 0.68$ for a negative sentiment. As a result of the research, the authors identified that attention to ethnic groups ranged highly by a certain group and a region. Based on this research, Koltsova et.al. improved the quality of results and increased the number of prejudice aspects that could be detected in the next paper [83]. To begin with, the authors enlarged the previous dataset for hand coding from 7,181 to 14,998 unique texts and performed annotation procedure by at least three independent persons. Next, they trained Logistic Regression to classify text into three classes (positive, neutral and negative attitude) using the best hyperparameters from the previous research and achieved significant improvement in classification metrics. The average values for sentiment accounted for $P = 0.67$, $R = 0.55$, and $F_1 = 0.58$.

The paper [84], by Nagornyy, examined the topic structure of ethnic discussions in Russian-speaking social media. Based on a comprehensive list of over 4,000 words related to ethnic discussions, the authors collected 2,659,849 texts from January 2014 to December 2016 from VKontakte using IQBuzz. For the topic modelling, the author applied ISLDA [26], that is a modification of the LDA algorithm developed by the Higher School of Economics Laboratory for the Internet Studies. To compute the sentiment score, Nagornyy utilised SentiStrength [22] with LINIS Crowd Russian-language sentiment lexicon [26]. For each topic, the polarity index was calculated as the sum of products of probabilities of this topic in the text and the corresponding

sentiment value divided by the overall salience of the topic. By analysing the topic profile of ethnic discussions, obtained using LDA, Nagornyy identified the most negative and salient topics. Thus, the primary piece of ethnic discussions was occupied by topics about Ukraine-Russia, taking into account the recent conflict between these countries. As a consequence, it was difficult to divide ethnic topics from political ones because of the current conflict, which influenced the polarity of discussions on the Internet. The most negative topics were formed around Uzbek ethnicity and Turkish-Armenian relations in the contexts of the Armenian Genocide. However, this study holds several weaknesses. Firstly, one of the drawbacks lies in the uncertainty about the exact way of data collection. Even though IQBuzz declares that it tracks all the mentions on the Internet, there is no way to validate this allegation without full access to VKontakte posts. Secondly, the classification metrics were not measured on the target text collection, so it is hard to verify the quality of the classified sentiment.

Borodkina and Sibirev, from Saint Petersburg University, studied the discussion on international migration issues in the Russian-language Twitter and different kinds of social problems associated with the migration [55]. To perform analysis, the authors utilised 13,200 tweets published between November 2017 and February 2018, which were initially selected based on the subject "migration" and its associated keywords. Next, the authors measured the similarity level between tags on the basis of cosine similarity and applied Pareto's principle to remove the insignificant, weak links from the network graph. For sentiment classification, the Support Vector Machine [33] classifier was trained. Finally, to identify the relationship between the characteristics (e.g. sentiment, the peculiarities of the content) the correspondence analysis techniques were used. It was found that, while the attitude to migrants among Russian users living in different countries is quite similar, a significant amount of these users hold a negative attitude towards international migrants. The key topics discussed in Russian-language Twitter, towards the migrants' issues are cultural and security risks connected with terrorist attacks and illegal migration, human rights in general, and the violation of the rights of the immigrants' in the social and economic spheres in Russia. However, this study has several minor drawbacks. The sentiment analysis approach was briefly described without specification of the pre-processing stage, training hyperparameters and final classification quality of the trained model. Additionally, basic Twitter API provides only partial access to all tweets, so the representativeness of the data sample should be further validated.

Thus, in the context of examining ethnic and migrant issues, the researchers primarily explored user-generated content from social network sites, using a combination of topic modelling and sentiment analysis techniques. The concept of ethnicity has been much theorised about in the academic literature. However, from the computational linguistics perspective, the task of ethnicity identification in user texts

narrows down to the task of identification of the ethnic markers usage by text authors [54]. So, to identify relevant texts, the researchers commonly compiled a list of ethnic status markers and searched for texts, which contained these markers. However, the retrieval of representative data is a hard task to accomplish, since not all platforms provide full access to all available information. Next, the document-level or aspect-level sentiment analysis is commonly performed to gain insight from the data. Since negative statements may contain identity-based attacks, as well as abuse and hate speech, they may be subject to censorship under the user agreement of the analysed social network site and the law. In the case of Russia, *The Criminal Code of the Russian Federation* has a regulation policy regarding publicly broadcast appeals for radical action, which is supposed to affect the volume of strong negative statements in both online and offline discussions. As a consequence, this nuance should be clearly identified in the limitations section.

b: UKRAINIAN CRISIS

Relations between Russia and Ukraine became tense after the Ukrainian revolution in 2014, followed by the annexation of Crimea and the uprising of separatists in the Donetsk People's Republic and Luhansk People's Republic, who advocated independence from Ukraine. Since a wide range of social networks has become integral to modern ways of social engagement [53], a series of computational linguistics studies attempted to examine the feasibility of using online discourse to analyse expressed opinions and characteristics of discourse participants. According to the Ukrainian Census in 2001, 67.5% of the population consider Ukrainian as the native language, while 29.6% of the population considers Russian as the native language. As a consequence, in addition to or instead of Ukrainian, researchers commonly analysed the Russian-language content.

Duvanova's research group studied the effects of Ukraine's military conflict with Russia in the online social connections among all Ukrainian regions [85]. The authors used VKontakte as a data source, since it is the most visited social network site in Ukraine. At the first stage, they identified the list of relevant communities using a list of predefined keywords. Based on the created list of 14,777 communities, the datasets of 19,430,445 wall posts and 62,193,711 comments were collected using social media monitoring software presented in Semenov and Veijalainen [105] and Semenov *et al.* [106]. To classify texts into positive and negative classes, the authors applied a rule-based approach with the vocabulary of 8,863 positive and 24,299 negative words in both Russian and Ukrainian. According to the results of the analysis, discussions in Ukraine became more polarised in response to war casualties. Regions located in the Eastern part of Ukraine demonstrated simultaneous increases in both negative and positive sentiments. However, in other parts of the country, war casualties had no observable effect on the intensity of sentiment expressions. Thus, war casualties enticed a strong emotional response in Ukraine's regions but

did not inevitably enlarge the level of social cohesion in internal communications between regions. However, the authors did not provide details of the preprocessing and training stages as well as classification metrics.

In the paper [86], Volkova *et al.* studied public opinion, expressed through social network VKontakte, during the Russian-Ukrainian crisis. By filtering relevant posts, using predefined keywords, the authors collected a dataset of 5,97,247 VKontakte posts, which were published by active users from September 2014 to March 2015. For targeted opinion prediction, they used the opinion classification system, POLYARNIK [107], which is based on morphological and syntactic rules, affective lexicons and supervised models [108]. For emotion classification, the authors created the dataset from an independent sample of crisis-related discourse from Twitter. Based on the approaches [109], [110], the authors bootstrapped noisy hashtag annotations for six basic Ekman's emotions [111]. They then performed manual re-validation of the automatic annotation by a native speaker of Russian and Ukrainian. In total, they collected 5,717 tweets annotated with anger, joy, fear, sadness, disgust, surprise, and 3,947 tweets that did not express any emotions. To predict emotions, a two-step approach was used. At the first step, texts were classified as emotional or non-emotional. At the second stage, emotional texts were categorised into six emotional classes using a logistic regression model [32] based on stylistic, lexical, and binary unigram features from stemmed words. The weighted F_1-measure of the emotional classification model achieved up to 58%. According to the results of text classification, the share of positive opinions, expressed toward Euromaidan, was higher in Ukraine than in Russia. In comparison, the percentage of positive impressions shown towards Putin and Crimea was more pronounced in Russia than in Ukraine. Additionally, some of the findings contradicted the common misconceptions broadcasted by the media. For example, there were posts from Russia, expressing sentiment in favour of the USA and against Putin, and posts, from Ukraine, showing support for Putin but not for Euromaidan. The main weakness in their study is that they applied POLYARNIK for sentiment analysis without examination of the classification performance on the target domain texts. Moreover, they applied the model for emotion recognition, trained on Twitter data, to posts from VKontakte, which has different linguistic characteristics, at least in terms of the average length of texts. Also, this raises many questions about the quality of annotation by only one assessor, since it is not possible to measure inter-rater agreement metrics.

Using the Russian-Ukrainian conflict in 2014 as a case study, Rumshisky *et al.* analysed the temporal dynamics of the political conflict as reflected in social media [87]. In contrast with Volkova's study [86], the authors did not rely on noisy location-based data to create a corpus for the analysis. However, instead, they used self-labelled user groups relevant to the crisis. Analysing VKontakte data, the authors manually selected 51 Antimaidan-related groups with 1,942,918 unique users and 47 Euromaidan-related

groups with 2,445,661 individual users and then retrieved all posts from these group walls. Next, the collection of posts was extended with posts from the walls of active users and the users who liked these posts. Only those posts, which included at least one of the predefined political keywords, were added to the collection. To predict the sentiment of Russian texts, they used an upgraded version of SentiMental library,[8] which is a dictionary-based sentiment analysis system. The results of the study confirmed that negative sentiment tends to accompanies the intensification of the conflict. During the analysed period, the relationship between the dominating sentiment and the random walk controversy measure was investigated. As a dispute arises, both the standard deviation of overall sentiment, expressed by the opposing groups, and the random walk controversy measure will increase simultaneously. The principal defect in their study is that they did not provide any details about the preprocessing and training phrases.

Zaezjev proposed an approach for studying the process of political mobilisation based on social media content analysis [88], using the Ukrainian Euromaidan revolution of 2013–2014 as a case study. The author focused on the very first days of protests, specifically from the 21st of February 2013 to the 22nd of February 2014, and analysed posts from the most popular social networks in Ukraine – VKontakte and Facebook. Zaezjev defined a set of relevant keywords, based on the Godbole's general guideline [112] and then collected more than 124,000 messages via IQBuzz. Using IQBuzz tonality recognition algorithms, the texts were classified into one of the following sentiment categories: negative, neutral, positive, and mixed. Based on the assumption that Euromaidan supporters would, instead, express positive sentiment to it, the authors removed all non-positive messages from the created collection. Next, the authors filtered the texts collection by a list of predefined keywords, so the final dataset of 4,255 messages was further utilised. By analysing the collected data, it was found that on the very first night of protests, social network sites were predominantly used as a tool of political mobilisation, and later, they became a tool of media coverage. The main drawback of this study is that sentiment classification metrics were not reported, so it is challenging to validate the accuracy of the outcomes.

Ukrainian top bloggers' discourse, in respect of the Donbas territory and population in 2009-2018, was studied by Tokarev from the Moscow State Institute of International Relations [56]. The author analysed the semantics, frequency and emotionality of discourses on the people and territory of Donbass by examining the Ukrainian Facebook segment. The study was carried out in several stages. Firstly, opinion leaders on Facebook were identified, and their public posts from the 1st of January 2009 to the 15th of February 2018 were downloaded. Next, only those posts which were devoted to the Donbas were extracted. Then, using a set of predefined keywords from the discourse, they identified the posts

dedicated to the Donbass. At the next stage, a sentiment vocabulary was created and was further utilised to differentiate discourses according to their degree of emotionality. Jointly with the help of volunteers, the dictionary of 566 marker words for territory and population was created, where each word was presented in both Russian and Ukrainian languages. After that, a team of 69 assessors annotated the vocabulary into five classes: positive, neutral positive, neutral, neutral negative, and negative. At the final stage, an assessment of the sentiment expression degrees was performed, and the dynamics of discourses were examined. The collected corpus of 1,069,687 posts in seven languages, from 376 top bloggers, was analysed. They showed that the beginning of discussions on the territory and population was at the turn of 2013-2014. Before that, the frequency of mentions of Donbass was almost zero. There was a significant negative discourse regarding the population, while negative discussion concerning the territory was practically absent. The main one is the neutral discourse. The number of positive and negative discourses, with respect to the territory, was much less than with respect to the population. That allows concluding about the high degree of uncertainty among top bloggers in relation to the territory and the low probability of the discourse transition from neutral to positive. Their approach suffers from the same pitfalls as Zaezjev's study [88], since sentiment classification metrics were not reported.

Therefore, in the process of studying the Ukrainian Crisis, in addition to sentiment information, researchers commonly utilised the geolocation of the posts to examine users' territorial affiliation. To identify relevant texts, researchers also compiled a list of target conflict markers and searched for texts, containing these markers. The challenges of retrieving representative data and outlining a comprehensive description of limitations remain the same as for the analysis of ethnic groups and migrants' issues.

c: SOCIAL TENSION

The processes that can be observed in modern Russian society shape a real need for putting the social relations conflicts in a particular framework [113]. Taking into account the widespread uptake of social networks, which brought both benefits and risks for civil society [114], the analysis of online content should be given due and appropriate consideration, specifically looking for social tension. Online social tension is possible to be measured through indices and metrics, and then using these values for monitoring of tension peaks, affording a form of 'anticipatory governance' [115].

Donchenko et.al. analysed user comments on social tension topics, discussed on VKontakte, for the period from January to June 2017 [89]. Based on the predefined list of popular topics related to social tension problems, they collected users' posts using preset search queries via VKontakte API. Collected texts were preprocessed using the following techniques: stemming, removing punctuation, replacing standard abbreviations or slang words with corresponding actual words. For text categorisation into social tension

---

[8]https://github.com/Wobot/Sentimental

topics, the authors trained the Support Vector Machine (SVM) [33] model with TF-IDF [116] vectorisation to classify texts into topics relevant to the level of unemployment, corruption crimes, and the growth of prices on consumer goods. Additionally, they used the SVM model to classify sentiment polarity of the texts. It was found that protest moods were predominantly concentrated in the centres of highly populated regions. One of the main drawbacks of this work is the lack of data annotation quality assessment and sentiment classification metrics specification.

Koltsova and Nagornyy examined which topics were defined as social problems by reviewing the comments of the audiences of the regional mass media in Russia [57]. They collected a dataset of 33,887 news items and 258,107 comments from local online media sources in Omsk (Gorod55,[9] BK55,[10] NGS Omsk,[11] and Omsk-Inform[12]) during the period from September 2013 to September 2014. To extract topics from news texts, the authors applied the Gensim [117] implementation of the Latent Dirichlet Allocation algorithm [102] with a metric by Arun, Suresh, Madhavan and Murthy [118]. To classify the sentiment polarity of comments, authors used SentiStrength [22] with PolSentiLex lexicon. Koltsova and Nagornyy found that such topics as entertainment, culture, sports and holidays had the most positive emotions, while the most negative emotions were related to crimes and disasters. Based on these findings, the researchers calculated the importance and polarisation index of each topic. A fundamental problem with the usage of SentiStrength in this study is that the authors did not report classification metrics on the target domain data, so it is challenging to validate the accuracy of the outcomes.

For the identification of social tension topics, the authors applied two types of approaches. The first one was based on the filtering data, based on a set of predefined keywords, and the second was based on the unsupervised clustering of the whole data sample and then extracting social tension topics. In the case of utilising content from social networks, the authors shared the same challenge of retrieving representative data. However, this challenge becomes irrelevant in the case of analysing data from news sites, since they commonly hold no restrictions to the access to published data. Since harsh statements may accompany social tension discourse, they may be subject to censorship under the user agreements of the analysed social media sites and the law.

d: OTHER TOPICS

Besides the analysis of conflict discourse, a series of studies examined attitudes towards topics in different domains.

Rulyova explored the reaction of Russian Twitter and YouTube users' to a meteor explosion in Chelyabinsk that took place in February 2013 [58]. The meteor was the largest

celestial body that entered Earth's atmosphere in the last 100 years. Therefore, it, expectedly, provoked a heated discussion in both the offline and online communication platforms. For the research, the author collected 495 tweets, by filtering texts with the hashtag ''meteorite'', published in the period from 15 to 20 February 2013, and an unspecified number of the YouTube videos. The emphasis was placed on the comparative analysis of YouTube and Twitter content, in the context of differences between primary and secondary speech genres [119]. However, a particular interpretation of sentiment and emotion in texts was also carried out. Rulyova found that YouTube content provides more useful data for exploring sentiment in contrast with Twitter content. The author drew on genre analysis and the mixture of linguistic and semiotic analysis, which meant that Rulyova analysed the text and how it was represented. The author supposed that this observation may have a few reasons. First, the instantaneity is relative and not absolute, especially when comparing spoken and written ways of communication. Second, the YouTube and Twitter users tend to belong to a different social group, and as a consequence, they may have different sentiment expression patterns. In general, even though the research is only indirectly related to the sentiment aspect of texts, the author was one of the first who examined differences between different types of Russian-language content. However, the methodology of sentiment comparison and YouTube data collection is not described in details. During the Twitter data collection, a basic filtering approach was utilised, ignoring a vast amount of data without direct mention of the 'meteorite' hashtag. Moreover, without using Historical API,[13] the Twitter search tool provides only partial access to all publicly available tweets.

Kirilenko and Stepchenkova conducted a comparative study of Twitter discourse on the Sochi 2014 Olympics in Russian and English [90]. Over 400,000 tweets were collected through API Twitter search for 6-month period windowing 2014 Sochi Olympic Games [120] and further utilised for cluster analysis and analysis of the sentiment on the Games. The authors evaluated Deeply Moving [121], Pattern, and SentiStrength [22] approaches on a manually labelled dataset of 600 English tweets and 3,000 Russian tweets. Based on the evaluation results, SentiStrength was selected for further utilisation in the research, since it achieved the highest scores for English and Russian datasets. It was found that while the positive attitudes expressed in the tweets about the Sochi Olympics improved throughout the course of the Games, this improvement was practically significant only for the hosts' segment of the sample. However, the authors did not provide classification metrics for the evaluated models and a description of the preprocessing stage.

Spaiser et al. examined a wave of mass protests surrounding the Duma and presidential elections happening in Russia between 2011 and 2012 [91]. Their analysis was based on Russian political discourse on Twitter collected between the

---

[9]https://gorod55.ru/
[10]https://m.bk55.ru/
[11]https://ngs55.ru/
[12]https://www.omskinform.ru/

[13]https://developer.twitter.com/en/docs/tutorials/choosing-historical-api

17th of March 2011 and the 12th of March 2012 via Twitter Streaming API. The retrieved tweets were filtered for Russian language and political content using a predefined list of Russian political keywords. After the filtering procedure, the dataset of 690,297 Russian language tweets with political content was compiled. To identify the pro-Putin camp and the opposition camp in the dataset, they used a predefined list of keywords in combination with SentiStrength [22], and then classified 1,000 most active users by the average sentiment score of their tweets as either belonging to one of the camps. By comparing manual annotation of 100 users into different camps with automatic classification, the researchers reported that approximately 70% of political camps were classified correctly. At the final stage, they performed a qualitative research method approach [122] and manually coded the key extracted n-grams. As one of the primary outcomes, they found evidence for Twitter discourse that matches initially with significant support for the opposition and later with declining in oppositional mobilization and growing support for Putin. However, this study holds several weaknesses. Firstly, one of the drawbacks lies in the uncertainty about the representativeness of the data sample, since Twitter Streaming API provides only partial access to all published data. Secondly, the classification metrics were not measured on the target text collection, so it is hard to validate the quality of classified sentiment.

Nenko and Petrova conducted a comparative analysis of the distribution of emotions in Saint Petersburg based on user-generated comments on urban venues from Google Places and data from public participation geographic information system Imprecity[14] [92]. The dataset contained 1,800 emotional marks from Imprecity and 2,450 geolocated comments from Google Places. A subset of comments was marked by two experts into six emotions and then processed with Naïve Bayes Classifier [123]. Based on the results of sentiment analysis and Impercity dataset, the authors created negative and positive emotional heatmaps for Saint Petersburg. A shared trend for the heatmaps is the concentration of positive and negative emotions in the historic city centre in the south of the city, at the western extremity of the semi-central Vasilyevskiy island, and the centre of the Petrogradskiy island. However, the authors did not provide preprocessing techniques and classification metrics.

Thus, by measuring attitudes towards different events or places, the researchers faced the same challenges of retrieving representative data and outlining a comprehensive description of limitations. Besides, in the majority of studies, the major drawback was in the absence of sentiment analysis model evaluation on the target-domain texts, so it is hard to validate the quality of classified sentiment.

2) SOCIAL SENTIMENT INDEX

To measure the level of happiness and life satisfaction, for instance, Subjective Well-Being (SWB) [124], the traditional psychological approaches rely on self-report scales, which have a series of drawbacks. For example, such nuances as a limited number of samples they can access, a high cost of respondents acquisition, and sensitivity to participants' memory make it a hard task to accomplish to present the real-time status of respondents [125]–[127]. As an alternative, scholars have attempted to measure a variety of social sentiment indexes through sentiment analysis of these content, because of a wide range of opinions expressed in user-generated content from social networks [2], [127]–[133].

In the paper [93], Panchenko calculated the sentiment index of the Russian speaking Facebook, which was measured as the average emotional level in a corpus. For the analysis, the author used a corpus of 573,000,000 anonymised Facebook posts and comments published in the period from the 5th of August 2006 to the 13th of November 2013, which were provided for research purposes by Digsolab LLC.[15] The authors filtered the entire collection for text written in Russian using langid.py module [134]. The social sentiment analysis index was computed based on the vocabulary-based approach [135], [136], similar to the Dodds's approach [129]. The author developed a custom sentiment dictionary of 1,511 terms, which was labelled by two annotators into positive and negative classes. To measure classification quality, Panchenko evaluated the vocabulary-based approach on the ROMIP 2012 dataset [15] and reported macro-averaged F_1-score up to 0.383 and accuracy score up to 0.465 on movie reviews dataset. To measure the sentiment, the authors proposed four indexes: Word Sentiment Index, Word Emotion Index, Text Sentiment Index, and Text Emotion Index. First two of them rely on the sentiment of words, while two others exploit the sentiment classification of texts. According to the results of the analysis, positive content predominates over a negative one. The maximum index values coincide with the national holidays, while index minimum values with commemoration days and national tragedies. Generally, users expressed positive sentiment terms in 3.8 times more than the negative ones. Users tended to show less emotional terms in posts and more negative and positive terms in comments. The most important limitation lies in the fact that the authors validated classification quality on the movies review dataset and applied it to the general domain texts, so it is challenging to verify the accuracy of the outcomes. Additionally, the internal process of data collection by Digsolab LLC has not been specified, and as a consequence, the question of its representativeness arises.

Shchekotin et.al. proposed a new method of subjective assessment of well-being based on the data of user online activity in social media using VKontakte data [68]. Based on Gavrilova's model of quality of life indicators [137], the authors selected a part of the indicators for monitoring in their study. Guided by geographical and socio-economic representativeness, they selected 43 regions of Russia out of 85. Then in each of these 43 regions, the authors identified

---

[14]http://www.imprecity.ru/

[15]https://digsolab.ru/

the three largest cities and selected ten communities on the VKontakte social network that bring together the inhabitants of these cities, i.e. the so-called urban communities. Next, they retrieved content published in these communities from the 1st of January to the 31st of December 2018 using social media data collection and analysis platform of the University Consortium of Big Data Researchers,[16] developed by Tomsk State University. At the next stage, they removed irrelevant data, e.g. advertising posts, as well as topics that are beyond the scope of this study containing information about vacancies, sports and cultural events. Irrelevant data filtering was carried out in two stages: manual screening of about 60,000 messages and automatic cleaning, the algorithm of which was trained based on manually cleared messages. The total number of posts left after filtering amounted to approximately 1,700,000. During the cleaning process, posts were manually annotated by 19 topics and by three sentiment classes (positive, negative and neutral). At the stage of data preprocessing, characters that are not in the English and Russian alphabet were removed, all words were reduced to the word basis using stemming, and rare words were removed. Based on the obtained data, a series of machine learning algorithms were trained. A gradient boosting algorithm from LightGBM [138] showed the best classification quality, achieving the accuracy score up to 68% for the category classification and 79% for the sentiment classification. To build the Index of Subjective Well-Being [124], [139] for each of the regions, the authors proposed a calculation method based on online activity indicators. The results of the study showed that the topics related to the development of regional infrastructure are most actively discussed in the chosen regions in a positive tone. The least positive activity is associated with assessing the overall emotional state and freedom of the media. Talking about negative assessments, the most significant online activity is observed in connection with the topic of security, i.e. an evaluation of the activities of law enforcement and other government agencies related to ensuring security in the region. The lowest negative indicators of online activity are again typical of the general emotional state and freedom of the media. Additionally, the authors compiled a comprehensive list of limitations, including data sample representativeness, the audience of the analysed social network, and the potential influence of bots. However, in the process of preparing the training collection, the authors did not indicate the distribution by sentiment classes. If datasets were unbalanced, the quality of classification is recommended to measure in more sophisticated metrics, for example, in Precision, Recall and F-measure.

3) USER BEHAVIOUR
Content from social networks can be a valuable source of information not only about the attitudes towards different topics but also about user behaviour patterns during interaction with the content.

Svetlov and Platonov determined the impact of sentiment on the mechanisms of feedback from the audience [69]. As a source of data, they utilised 46,293 posts and 2,197,063 comments from the most popular accounts of Russian politicians on the social network VKontakte in the period from January 2017 to April 2019. Svetlov and Platonov trained Bidirectional Gated Recurrent Unit [140] on RuTweetCorp [141] and RuSentiment [142], achieving macro-averaged $F_1 = 0.91$ and $F_1 = 0.77$, consequently. Based on the results of the sentiment analysis, the authors identified several reacting patterns. To begin with, posts classified as positive have a higher number of views and likes from users, while posts classified as negative have a higher number of reposts and comments. However, using training data from one domain and applying to another domain raises many questions. RuTweetCorp is a collection of short posts from Twitter and RuSentiment is a collection of general-domain posts from VKontakte, while target posts of this study were from the political domain. A good solution to this uncertainty may be to manually annotate a small dataset in the target domain and test the trained model on it.

B. PRODUCT AND SERVICE REVIEWS
In the Internet era, product and service reviews have became a powerful expression of social proof that push people to purchase commodities in a variety of e-commerce platforms [143]. Reviews tend to be a valuable source of information, not only for customers and merchants but also for the research community. In this section, we divided the literature into subsections based on the target object of the analysis: the characteristics of customers, the characteristics of products and services, and the characteristics of merchants.

1) CHARACTERISTICS OF REVIEWERS
The research group from Saint Petersburg University analysed topics and sentiments of online job reviews for 989 organisations operating across 12 different knowledge-intensive industries in Russia [70]. As a primary source of data, the authors used one of the largest Russian portals of reviews of employers Otrude,[17] operating in Russia. After the data filtering and data removal process, a sample of 6,145 observations was created. The preprocessing stage consisted of several steps: lemmatisation using MyStem,[18] removing punctuation, and stop words. The authors automatically classified texts into two classes based on the review rating, where reviews with at least three stars were assumed to be positive, and all other texts were assumed to be negative. Then the research group employed Latent Dirichlet Allocation [102] or topic modelling and an unspecified classification model for sentiment analysis. Based on the analysed data, Sokolov *et al.* identified the following six main factors of job satisfaction: working arrangements and schedule, working conditions, job content,

---

[16]http://www.opendata.university/

[17]https://otrude.net/
[18]https://yandex.ru/dev/mystem/

salary/wage, career development, psychological climate, and interpersonal relations with coworkers. The two latter factors, mental environment and interpersonal relationships were the most widely discussed by employees online concerning job satisfaction. Thus, the authors suggested that when deciding to leave the company, employees tend to tolerate economic satisfaction factors (for instance, working career development perspectives, and salary) rather than socioemotional factors (for example, poor content of work, poor relationships with their coworkers). One key issue that needs to be raised is the validity of the usage of sentiment analysis in this research. In most cases, each review from Otrude contained the rating provided by the author, so technically the evaluation of job reviews can be performed without sentiment classification. An additional problem is that the authors did not provide a description of sentiment analysis approach and sentiment classification scores on the test set.

2) CHARACTERISTICS OF PRODUCTS AND SERVICES
Based on the user reviews from transport portal Autostrada,[19] Seliverstov *et al.* evaluated road pavement assessment of the Northwestern Federal District of Russia [71]. As a source of training data, the authors used RuTweetCorp [141], that is, the largest automatically annotated corpus with minor manual data filtering, which was collected automatically from the Russian part of Twitter. As a classification model, they trained the regularised linear model with stochastic gradient descent and the Bag of Words model with TF-IDF vectorisation. The trained model achieved binary classification accuracy up to 72%. Analysing user reviews from the 1st of March 2009 to the 1st of November 2018, the authors found that the length of positively assessed roads was 9,874 km (or 75% of the whole length of roads), and the length of negatively evaluated roads was 3,385 km (or 25%). However, this research suffers from a series of pitfalls. First of all, the authors did not describe the preprocessing step, which is crucial for training on RuTweetCorp. The nuance lies in the fact that RuTweetCorp was initially designed for the creation of a sentiment lexicon, not for direct sentiment classification. The dataset was collected automatically based on the [144] strategy, i.e. each text was associated with the sentiment class based on the emoticons it contains. As a consequence, even a simple rule-based approach is able to demonstrate outstanding results. For example, if a model classifies text as a positive if it contains '(' character and as a negative otherwise, it achieves $F_1 = 97.39\%$ in a binary classification task. In order to deal with the automatic sentiment analysis task, the author of the dataset recommended removing emoticons during the preprocessing stage. According to the paper [145], in this case, it is possible to achieve the macro-average F_1 score up to 75.95% using Support Vector Machine [33]. As a consequence, without knowing the preprocessing technique, it is hard to evaluate the validity of the research. Secondly, this raises many questions about the effectiveness of using

training data from one domain and applying to another domain. A solution to this uncertainty may be to manually annotate a small dataset in the target domain (user reviews from the transport portal) and test the trained model on it. Thirdly, RuTweetCorp consists of three classes, but the authors did not consider the neutral class in their study. Positive and negative tweets were published on 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 [146]–[150] utilised only positive and negative tweets, performing binary classification. It can be supposed that the neutral class may change the overall distribution of negative and positive reviews about roads. Lastly, most reviews from Autostrada contained the rating provided by the author, so technically the evaluation of road pavement assessment could be performed without sentiment classification. In this case, it can be interesting to compare the assessment based on the review ratings and sentiment classification labels.

3) CHARACTERISTICS OF MERCHANTS
Li and Chen from the University of Arizona developed a machine learning framework for identifying sellers' product quality based on customer feedback [72]. This framework was constructed from three key modules: snowball sampling using keywords and related users, thread classification using Maximum Entropy classifier, and sentiment analysis using deep-learning methods. One distinctive feature of the sentiment analysis module is that it firstly translated the Russian-language text into English using Google Translate, and only then performed sentiment classification using the Recursive Neural Tensor Network with Sentiment Treebank word representations [121]. The proposed framework was tested on a Russian carding forum, and as a result, top malware and carding sellers from this forum were identified. Based on a more detailed analysis, it was observed that carding sellers generally tend to have lower ratings in comparison with malware sellers. The authors supposed that this tendency was caused by the fact that the quality of malware is easier to determine in comparison with the stealing of carding information. The authors mentioned that the sentiment classifier was trained on an online review dataset, which is similar to their problem, however, they did not provide any details about used dataset and classification quality metrics. The translation step can affect the meaning and the polarity of the texts significantly, so without a test set of texts in Russian, it is virtually impossible to measure the real sentiment analysis quality.

C. NEWS FROM MASS MEDIA
UGC content in social networks, as well as reviews, usually represents subjective texts, because authors express their opinions freely. However, the situation is different in the context of news analysis. The news agencies try to avoid making judgments and overt partiality and steer clear of doubt and ambiguity. Objectivity, or at least widely acceptable

---

[19]https://autostrada.info/

neutrality, is their philosophical basis [73]. As a consequence, the journalist often abstains from using negative or positive vocabulary, but resort to other ways to express their opinion [74]. For example, journalists may highlight some facts and omit others, embed statements in a complex discourse structure, quote persons who share their point of view. Widespread human interest in the news was repeatedly observed over the centuries [151], [152]. Globally, the news as a data source for sentiment analysis is used in a variety of directions, e.g. evaluation sentiment in the news [153], [154], stock price prediction [155], [156], election result prediction [157], [158], the price prediction of e-commerce products [159], and futures buying behaviour [154]. As for the Russian-language news, we identified two categories of studies: measuring the sentiment aspect of the news and making economic and business forecasts.

1) CONTENT OF NEWS

A series of Belyakov's articles [94], [95] was devoted to the sentiment analysis of news messages from the online site of the Russian Ministry of Foreign Affairs. The author utilised articles in the ''News'' section published from the 1st of February to the 28th of February 2015. As a unit of analysis, a piece of text was adopted corresponding to one of the following categories: ''the Ukrainian question'', ''Cooperation between Russia and China'', ''Relations between Russia and Ukraine'', ''The conflict in Syria'', ''Cooperation with Turkmenistan'', ''Relations between Russia and Greece'', ''Sanctions against Russia'', ''Diplomacy today''. The author constructed a basic rule-based classifier, which sums the polarity of emotional words in the processed text and predicts the final binary label. A dictionary of 300 positive stems and 300 negative stems was additionally compiled. According to the results, the categories ''Russian Cooperation with China'', ''Russian Cooperation with Turkmenistan'', ''Russian-Greek Relations'' and ''Diplomacy today'' were positively coloured. In contrast, categories ''The Ukrainian Question'', ''Relations of Russia with Ukraine'' and ''Sanctions against Russia'' were negatively coloured. One of the critical nuances lies in the fact that this study takes into account only the content of articles written by journalists, thereby expressing the official intention of the Ministry of Foreign Affairs about these topics. Further work can be focused on incorporating user reactions and comments to the news article published on the site. From the sentiment analysis perspective, the main shortfall of Belyakov's research is in the lack of the model evaluation stage. Without knowing performance metrics on the test data, it is impossible to validate the model and, as a consequence, to validate the outcomes of the sentiment analysis process.

A research group from the Russian Academy of Sciences examined sentiment coverage of technologies and innovations mentioned in the mass media [96]. Using Exactus Expert [160], the authors selected more than 240,000 articles dedicated to innovation and technology from 16 media

sources, which were published from 2005 to 2015. Then they categorised the articles into 11 technological trends from ''The list of critical technologies of the Russian Federation'' based on manually defined keywords. Next, the authors selected 120 articles and manually annotated each object of tonality mentioned in the article with the corresponding sentiment: positive or negative. Based on the created training corpus of annotated 346 pairs, they constructed emotive vocabulary and designed a rule-based classification algorithm. It was found that generally, mass media sources tend to write neutrally about technologies, which can be explained by the consistency of the style of news materials. The comparatively low share of negative mentions about information technology and biotechnology and the positive coverage of these articles, in general, demonstrated that society does not care about the potentially negative consequences of these technologies. At the same time, the share of negative feedback on military technologies is higher than in other trends. However, the authors did not provide classification metrics of the developed algorithm. Moreover, as mentioned, these articles were written by journalists, which may express not only society point of view, but also the official intention of their media. To evaluate social opinions on different analysed topics, it is necessary to examine their reactions to the news articles additionally.

Kazun and Kazun [75] performed analysis of Russian media coverage of Trump's activities during and after the election campaign. The authors used data from the Integrum database for network analysis and the Medialogia database for discussion sentiment analysis. For the study, they selected three time intervals: a month before the election, a month after the election, and seven months after the election. Medialogia's sentiment analysis approach was utilised for the classification of the sentiment of texts into three classes: positive, negative and neutral. It was carried out that media coverage of Trump before the votes was more negative than positive. However, in some months media coverage of Clinton's campaign was even more positive than Trump's, although in the four months before the election the articles related to Clinton were predominantly critical. One of the drawbacks of this study lies in the fact that the authors did not report the classification quality on the target domain data. As a consequence, it has become challenging to validate the accuracy of the outcomes.

These studies were dedicated to the analysis of political or governance news. In contrast with content from social networks, there were no challenges regarding the access to the historical data, because mass media commonly holds no restrictions to the access to all published data. However, some studies based on the content of the news attempted to identify public attitudes to specific topics, which, to our mind, require further elaboration. Mass media, of course, are supposed to be a proxy of public opinion. However, in some cases, the publisher policy may affect the presentation form, so the news does not always represent society's opinion.

2) ECONOMIC AND BUSINESS FORECASTS

Yakovleva proposed an approach for constructing a high-frequency indicator of economic activity in Russia based on news articles from internet resources, incorporating sentiment aspects of the texts [76]. There were two components constructed within the study: the first one aims at reflecting the quantitative component of the topics and the second one aims at identifying the sentiment of news. The preprocessing stage consisted of several steps: stemming using MyStem, removing punctuation, stop words, and unnecessary spaces. As a sentiment classification model, Yakovleva trained Support Vector Machine (SVM) [33] on the manually annotated dataset, which consists of 3,438 positive and negative news articles. The accuracy score on the test subset achieved 64%. The author mentioned that if the SVM model defined the text tonality with a probability of less than 60%, its tone was determined to be neutral and excluded from the analysis. All topics obtained from the first component were incorporated with sentiment information from the second model and based on this joint information the regression model was developed, which targets forecasting Purchasing Managers' Indexes[20] (PMI). The testing phase was performed on test data covering the time period from February 2017 to August 2018. The model demonstrated relatively strong predictive power, approximating the actual PMI index accurately for the new period. Thus, the results showed the ability of the model to monitor economic performance carefully, allowing faster responses to the current financial situation and prompt decision-making. However, this research suffers from a number of uncertainties. To begin with, it remains unclear what kind of probabilities were used by the author, since a basic SVM implementation does not directly provide probability estimates. Moreover, the methodology of selecting the threshold value was not specified. Next, Yakovleva's training data was annotated by only one annotator, i.e. without aggregating annotations by different annotators according to best practices [142], [161], [162]. Finally, the authors provided a graphical representation of comparison predicted and actual PMI values, but did mention any of the regression performance metrics.

D. BOOKS

During the last 60 years, the analysis of scholarly writing has come a long way, ranging from manual citation counts and word frequency analysis to modern automatic text mining methods [163]. One of the hot topics in this direction is the analysis of the sentiment of educational materials.

1) CONTENT OF BOOKS

Solovyev's research group conducted a sentiment analysis study on the corpus of textbooks on Social Studies and History used in Russian secondary and high schools [77]. For this study, researchers compiled The Russian Academic Corpus (RAC) based on 14 Russian textbooks on Social and

History for middle and high school, edited by Bogolyubov and by Nikitin. The preprocessing stage included sentence tokenization, words tokenization, and Part-of-Speech tagging using TreeTagger [164]. Using Russian-language lexicon RuSentiLex [25], the authors calculated the frequency of sentiment words occurring in each document and measured the number of occurrences per 1,000 words in a document. As a result of the RAC analysis, the authors found that the discourse within History textbooks for high school as well as Social Studies textbooks for middle and high school written by Nikitin incorporates predominantly negative sentiment by means of using negatively polarised words and presenting negative referents. Meanwhile, Social Studies textbooks written by Bogolubov contain mainly positive sentiments. Though, a significant source of unreliability is in the accuracy and relevance of the extracted sentiment words from the corpus, because RuSentiLex was originally created for other domain fields. Additionally, RuSentiLex provides a single context-independent representation of the sentiment polarity regardless of where the word occurs in a sentence and regardless of the different meanings it may have. As a consequence, this approach does not allow the capturing of different meanings of words based on the sentence context.

2) EDUCATIONAL PROCESS

Kolmogorova conducted an experiment with Chinese students learning Russian as a foreign language [78]. The author measured the correlation between the sentiment of educational texts, a subjective assessment by the international students of the attractiveness and effectiveness of the course, and the real success of instruction based on such texts. For sentiment classification, the authors applied a machine-learning-based emotion classifier developed by the Laboratory of Applied Linguistics and Cognitive Studies at the Siberian Federal University. The sentiment analysis model classified texts into nine classes with a macro-averaged F_1 score value of 50%, where eight of the classes correspond to basic L'ovheim's emotions [165], and the last grade is emotionally neutral texts. The texts for the training sample were selected from the public group ''Overheard'' of the social network Vkontakte. 231 native speakers of the Russian language labelled the texts, subjectively assessing the degree of expression of any emotion in it, where each text was assigned only one emotion. Each text was marked with at least three annotators. If two or three of them attributed the text to the same emotional class, then this emotion was attributed to the text. Otherwise, the text was removed from the training sample for this emotional class. For this study, Kolmogorova used texts for which the leading emotion was pleasure/joy and sadness/yearning. The experiment was attended by 30 students from China, who were divided into three equal groups. Each group learnt and was examined on the topic ''Punctuation'' on the material of joyful, sad and neutral texts, respectively. After conducting experimental education and examination phases, examinees filled a questionnaire, where they indicated their overall interest in

---

[20]Purchasing Managers' Indexes are economic indicators derived from monthly surveys of private sector companies.

the course, the effectiveness of the course, and their level of satisfaction with the educational process. Based on the analysis of the questionnaire and examination results, Kolmogorova found that the sentiment of the educational text has a significant impact on the subjective assessment of the educational process and its objective success. On average, students commonly made fewer mistakes in sad texts than in joyful and neutral texts, but working with them causes the lowest level of satisfaction. Work with joyful educational texts is of the most considerable interest but demonstrates less educational effectiveness. A significant source of uncertainty is in the method used to classify sentiment polarity of the texts, which lies in the fact that texts classification model was trained on texts from one domain and applied to texts in another domain field without additional examination of the classification performance. The authors did not provide any details about the machine learning model used for the classification or more information about preprocessing and training stages.

In the case of textbook analysis, the critical challenge is the absence of the sentiment lexicons and training datasets within the target domain. When scholars analysed texts at the word-level using sentiment lexicons, for each word they commonly extracted a single context-independent representation of the sentiment polarity, regardless of where the term occurs in a sentence and regardless of the different meanings it may have. For the analysis texts on the document-level, it is becoming hard to associate texts with the corresponding sentiment class, since texts in textbooks are long, and during the whole texts authors may express different emotions.

E. MIXED DATA SOURCES

To cover a broader range of materials, some studies used texts from different data sources. For instance, if the authors utilise news and UGC from social network sites, they are able not only to measure the polarity of media coverage of certain events by news agencies and government organisations but also to measure attitudes on differen discussed topics.

In the paper published by Berkman Center for Internet & Society [97], Etling examined social media sentiment in an online conversation about the Ukrainian Euromaidan protests across a range of Russian and English online and traditional media sources. The research used software provided by Crimson Hexagon [166], which was based on a content analysis approach developed by Hopkins and King [167]. With respect to the protests, texts were classified into four classes: positive, neutral, negative, or irrelevant. As a source of data, the author used Russian-language and English-language content from Twitter, Facebook, blogs, forums, and news sites, which were published from 21st of November 2013 to 26th of February 2014. Due to limitations of Crimson Hexagon, Ukrainian-language sources were not taken into account. According to the results, Russian-language sources and users demonstrated more support for the Euromaidan protests that it was originally expected. Sentiment aspect in English-language sources of the UK and the US tends to be

more negative than anticipated based on the ideological support among western governments. At the same time, the content of social media in the UK, the US, and Ukraine tend to be more positive in comparison with the traditional media sources in those countries. The main limitation pitfall of this research lies in the sentiment classification model. Firstly, it was trained on a minimal amount of training data, i.e. approximately 120-140 labelled posts in total. Secondly, this training data was annotated by only one annotator, i.e. without aggregating annotations by different annotators according to best practices [142], [161], [162]. Last but not least, reliability and performance tests were not conducted, which is contrary to the basic principles of creating supervised machine learning models [168]. Moreover, the full list of analysed sources was not provided, so it is hard to validate the reliability of those selected. Since Ukrainian-language sources were not taken into account, it can be assumed Ukrainian-language content may be more supportive of the protests in comparison with Russian-language content.

Kazun analysed the intensity and sentiment of the media coverage of Alexei Navalny[21] based on data from Russian mass and social media from 2014 to 2016 [80]. Using Medialogia, the authors accessed more than 145,000 news items about Navalny from national newspapers, online media sources, and the three largest federal TV channels. For sentiment classification, Kazun used sentiment analysis algorithms developed by Medialogia (classification into positive, negative or neutral classes), after a prior verification of the adequacy of the sentiment analysis algorithms on the manually-labelled material of 200 articles. It was found that traditional media sources tend to ignore Navalny except for the occasional release of documentaries or news stories to smear the Russian opposition as a whole or Navalny personally. Blogs generally demonstrated relatively more positive coverage of Navalny than the other analysed media types. However, the discussion in these articles was still primarily critical. Additionally, the authors described the nuances of each media type, thereby clarifying their publishing strategies and sentiment coverage patterns. Despite the overall negative bias, the news coverage about Navalny tends to become more positive from year to year. This trend is caused by both the decrease of the share of critical articles and the increase of the positive articles. As in all identified cases of the usage of Medialpgia's sentiment analysis approach, the authors did not report classification metrics on the target domain.

Within the study [79], Brantly analysed the Euromaidan revolution in Ukraine during 2013-2014 using public content from social networks such as Twitter, Facebook, YouTube as well as other blogs, forums, and news sites. The collection of 2,809,476 unique pieces of content in Russian, Ukrainian and English languages was collected from various online sources via Crimson Hexagon platform. Only data which were published from Ukraine in the period from the 21st of November 2013 to the 1st of March 2014 was

---

[21]Alexey Navalny is a Russian lawyer and political activist

considered in the study. Two persons, which were fluent in Ukrainian, Russian and English, annotated a training dataset for BrightView, that is, a machine learning feature within Crimson Hexagon, which implements non-parametric content analysis algorithm described in [166]. The texts were categorised into three classes: positive, neutral, and negative. Control tests of the Crimson Hexagon platform reported 92% fidelity with the human coding. Jointly with the collected data, the author utilised information from Tone Dataset Global Knowledge Graph and Events Dataset and Global Events Language [169]. The results of the analysis outlined that in Ukraine, there was a clearly defined divergence in political association and a preference associated with linguistic characteristics. This fact was additionally confirmed by voting returns in the past election cycles, whereas Ukrainian speakers traditionally expressed a higher level of support for opposition parties. By comparing offline and online engagement directionality, Brantly outlined that social media had a significant impact on physical protest turnout, i.e. social media leads to an increase of the number of protestors in the streets.

The main downside of using different data sources lies in the fact that as an addition to a broad range of expressed opinions, the authors also received source-specific challenges and limitations. For example, access to representative data, a comprehensive description of limitations, lack of training datasets within target domain. Some of the identified studies perform sentiment analysis and sentiment indices aggregation based on the whole range of texts without differentiation to the different sources. For instance, during sentiment aggregation, they consider posts from social networks and news articles as equal units. We suggest that in this case, a more complicated weighted model is needed in order to examine texts from different sources carefully.

V. CURRENT CHALLENGES

Based on the analysis of the identified literature, the following ten common challenges were derived. In general, researchers usually experience numerous challenges, including access to the representative historical data, and to the training data, guidelines for sentiment annotations, comprehensive description of limitations of the study, and topics extraction from texts.

1) **Access to the representative historical data of analysed sources**. Historical data, for instance, posts and reviews, collected from API of analysed sources or data aggregation platforms, are commonly analysed and utilised in sentiment monitoring studies. In the case of access to data via API, sometimes API providers grant only partial access to all publicly available data. For example, the basic Twitter API follows this policy, while historical Twitter API[22] provides access to all publicly accessible tweets. In case of access to data via data aggregation platforms, even if they declare

full access to a particular data source, there is no way to validate this allegation. Thus, there are only two ways to make sure that the data for the research will be representative. The first one is to carefully examine the description of the API and select the API option, which provides full access to historical data. In case of the usage of data aggregation platforms, it is necessary to make sure that they use API options with full access to historical data. The second one is to request access to the historical data directly from the data source. For instance, access to the historical data from Odnoklassniki, the second largest social network in Russia [98], can be requested directly through OK Data Science Lab.[23]

2) **Access to the training data from the target domain field**. Even though Russian is one of the most common languages in the World Wide Web,[24] generally it is not as well-resourced as the English language, especially in the field of sentiment analysis. Even though many studies aimed at sentiment classification of Russian-language content, only a few of them made their datasets publicly available for the research community. In case if none of these datasets is applicable for the target domain of the study, researchers have to perform manually labelling of the training dataset. Based on literature analysis and the papers [142], [173], we identified and described 14[25] publicly available sentiment datasets in Russian (see Table 2), which can be used in further studies.

3) **Lack of the test data for calculating classification metrics in case of usage of third-party sentiment analysis systems**. In the case of usage third-party sentiment analysis systems such as SentiStrength [22], Medialogia's sentiment analysis algorithms or POLYARNIK [107], the authors commonly not reported the classification quality on the target domain data. As a consequence, it has become challenging to validate the accuracy of the outcomes. We assume that the use of third-party approaches is also related to the fact that researchers do not have an annotated test collection of texts for calculating classification metrics. However, we believe that the implementation of this step significantly increases the academic value of the work. Therefore, it is highly recommended for authors to manually annotate a sample of the target data for measuring sentiment classification metrics.

4) **Topics extraction from texts**. To extract topics from analysed texts, the majority of studies utilised topic modelling methods. However, in case of the share of

[22]https://developer.twitter.com/en/docs/tutorials/choosing-historical-api

[23]https://insideok.ru/dsl/about-data-lab

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

[25]We consider 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 [174], [175], because we were unable to get access to the datasets using their official website.

TABLE 2. Sentiment analysis datasets of Russian language texts.

Dataset	Description	Annotation	Classes	Access Link
RuReviews [143]	A sentiment dataset of online reviews from "Woman Clothes and Accessories" product category at the major e-commerce site in Russia.	Automatic	3	Project page
RuSentiment [142]	A public sentiment dataset of posts from the largest Russian social network Vkontakte.	Manual	5	GitHub page
Russian Hotel Reviews Dataset [171]	An aspect-based sentiment dataset of 50,329 Russian-language hotel reviews.	Automatic	5	Google Drive
RuSentRel [172]	A dataset of analytical articles from Internet-portal Inosmi, which incorporates both the author's opinion on the subject matter of the article and a large number of references mentioned between the participants of the described situations.	Manual	2	GitHub page
LINIS Crowd [26]	A public sentiment dataset, which was constructed based on social and political blog posts from social media sites.	Manual	5	Project page
Twitter Sentiment for 15 European Languages [173]	A sentiment dataset of over 1,600,000 tweets (tweet IDs) in 15 European languages, including Russian.	Manual	3	Project page
SemEval-2016 Task 5: Russian [49]	A public aspect-based sentiment dataset of the restaurants' domain, which is based on SentiRuEval-2015 [17].	Manual	3	Project page
SentuRuEval-2016 [18]	A public aspect-based sentiment dataset sentiment analysis of Russian-language tweets towards telecommunication companies and towards banks.	Manual	3	Project page
SentuRuEval-2015 [17]	A public aspect-based sentiment dataset aspect-oriented sentiment analysis of users' reviews about restaurants and automobiles.	Manual	4	Project page
RuTweetCorp [141]	A largest automatically annotated openly available corpus with minor manual data filtering, which was collected automatically from the Russian part of Twitter using [144] strategy.	Automatic	3	Project page
Kaggle Russian News Dataset	A public sentiment dataset of Russian news	n/s	3	Kaggle page
Kaggle Sentiment Analysis Dataset	A sentiment dataset of Russian news.	n/s	3	Kaggle page
Kaggle IS161AIDAY	A sentiment dataset published by Alem Research.	n/s	3	Kaggle page
Kaggle Russian_twitter_sentiment	A sentiment datasets of Russian tweets.	n/s	2	Kaggle page

texts related to target topics is well below than 1%, topic modelling is generally unable to deal with topics extraction [54]. Moreover, topic modelling demonstrates poor accuracy on analysing short texts, especially in case if texts represent everyday talk [54]. Thus, the more accurate and noise insensitive approaches are needed to be developed.

5) **Sentiment annotation guidelines for manual labelling**. Since relevant training data in the Russian language is not always available for the target domain field, it is common for researchers to perform manual annotation of texts. Without publishing annotation guidelines and other details of the annotation process, it is challenging to validate the annotation quality of the labelled dataset. Clear and simple step-by-step instructions are crucial for obtaining high-quality sentiment annotations from both certified linguists and assessors without linguistic background [176]. Some types of texts are especially challenging for sentiment annotation, for instance, a speaker's emotional state, neutral reporting of valenced information, sarcasm, ridicule and others [162]. As an example of sentiment annotation guidelines for Russian, further studies can utilise guidelines used during RuSentiment annotation [142]. In case if a group of certified linguists is not available for the annotation process, the researcher can perform annotation via assessors from Yandex.Toloka, that is a Russian crowdsourcing platform

for manual data annotation. It has already been used in several academic studies about Russian-language texts [177]–[180]. In addition to sentiment annotation guidelines, it is highly recommended to publish an inter-annotator agreement measure such as Fleiss' kappa [181] or Krippendorff's alpha [182] as well as other details of the annotation process.

6) **Comprehensive description of limitations.**. A majority of the analysed papers suffers from an incomprehensive list of limitations. To cover a broad range of study limitations, in addition to the technical and methodological limitations of the utilised approach, it highly recommended to specify the following limitations:

a) **Level of Internet penetration**. One of the critical limitations is at the level of internet penetration in a country since certain groups of people may not be considered in the study. According to the Omnibus GFK survey in December 2018 [9], internet penetration in Russia exceeds 75.4%, with 90 million Russian users aged 16 or more. The usage of the internet among young people (16-29) and middle-aged people (20-54) is close to saturation, reaching 99% and 88%, respectively. However, despite the significant growth, the older generation of 55+ years old is still underrepresented among Internet users reaching only 36%.

b) **Representativeness of a data source audience**. In the case of analysis of social networks content, the next source of uncertainty lies in the fact that the audience of a particular social network site may be generally not representative to the public [183]. Moreover, different social networks may have completely different audiences. However, the fact that a social network is not representative of the public does not actually mean it is impossible to infer insights about the public from it. For example, it is possible to up-weight the mood information gathered from the underrepresented age groups of the analysed social network, thus mathematically simulating a situation where the population of the analysed social network actually is more representative of the broader Russian population.

c) **Media freedom**. In Russia, as in many other countries, there are restrictive regulation policies on the dissemination of certain information. Since negative statements may contain identity-based attacks, as well as abuse and hate speech, they may be subject to censorship under the user agreement of the analysed social network site and the law. For example, propaganda instigating social, racial, national or religious hatred and strife; a rehabilitation of Nazism; blasphemy; slander, insult, drug propaganda, gay propaganda, using foul language in the media, dissemination of information about a person's private life without his or her consent, and others. Thus, these policies are supposed to affect the volume of strong negative statements in both online and offline discussions. As a consequence, this nuance should be clearly identified in the limitations section, especially in the context of studies regarding conflict situations.

d) **Internet censorship**. According to a rating compiled by Freedom House in 2018 [184], Russia ranked 53rd out of 65 in terms of Internet freedom. Currently, the government of Russia is waging a campaign to gain complete control over the country's access to, and activity on, the Internet [185]. Starting from 2012, the Federal Service for Supervision of Communications, Information Technology and Mass Media Roskomnadzor) maintained a centralised internet blacklist, which is used for the censorship of individual IP addresses, domain names, and URLs. In April 2019, the State Duma adopted the so-called law on sustainable Internet in Russia. Among other measures, the law provides that traffic exchange in the country will be carried out only through exchange points approved and entered in the relevant register. Thus, regulation policies may affect the number of online data sources for the study, which decreases the variety of points of view for analysis.

7) **Cross-domain sentiment analysis**. Since individuals can express their opinions in a vast number of domains, analysing all these opinions may become resource-intensive because of the training texts collections should be annotated for each domain field [186]. The lack of annotated text collections for training a sentiment analysis model for all domains results in restriction of the accuracy of sentiment analysis. According to the survey [187], there are three significant issues to be addressed in cross-domain sentiment analysis. The first one is that opinions expressed in the context of one domain may be reverted in the context of another domain field. The next one lies in the difference in sentiment vocabularies among different domain fields, that should be considered during sentiment analysis. And the last one is to reasonably assign a sentiment strength marker to each token in the sentiment vocabulary.

8) **Sarcasm and Irony Detection**. Online communication, especially relevant to politics, often contains sarcastic and ironic phrases [188], which can be difficult to recognise even for people, let alone natural language processing approaches. Currently, a minimal amount of studies [189] has been dedicated to irony and sarcasm detection in Russian. Thus, to correctly process a broader range of opinions, more automatic classification approaches are needed to be developed and further utilised in sentiment monitoring studies.

9) **Bots detection**. The presence of bots has a significant impact on a variety of aspects of social media, especially in the case of bots accounting for a large share of its users. Such bots can be used for different malicious tasks regarding public opinion, e.g. inflating the perceived popularity of celebrities or spreading false information regarding politics [190]. Thus, the methods for their identification are needed to be developed and integrated into the applied sentiment analysis studies.

10) **Effectiveness of the analysis results**. There is still considerable controversy surrounding the effectiveness of measuring reactions through automatic analysis of the online content. While some studies [191], [192] considered that social network-based approaches are less accurate than traditional surveys, other studies [193] stated that they demonstrate higher performance in comparison with traditional methods. Thus, it is highly recommended to confirm the results of the study by the results received by another methodological approach if it is possible.

VI. FUTURE RESEARCH OPPORTUNITIES

Based on the analysis of the identified literature, the following seven future research opportunities were identified. In general, future research needs to thoroughly investigate the sentiment monitoring approaches presented in this paper in

order to identify potential synergies between the individual approaches to allow for a more comprehensive analysis of sentiment expressed in a variety of textual sources.

1) **Transfer learning of language models for sentiment analysis**. The majority of analysed papers applied rule-based and basic machine learning approaches, and only two studies [69], [72] utilised neural networks. However, 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 [43], [194]–[198]. 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. Some initial research was performed in the paper [199], where the authors trained shallow-and-wide CNN with ELMo embeddings [42] and achieved new state-of-the-art classification metrics on RuSentiment [142], thereby outperforming all previous neural network-based approaches. As a first step towards this direction, researchers could train and publish transfer learning baselines for different sentiment analysis datasets in Russian.

2) **Sentiment analysis of multilingual content**. Russia is a multinational country, and therefore multilingual, so different individuals and groups of people tend to express their opinions in different languages. Linguistic scientists counted more than 150 languages, starting with Russian, which is spoken by 96.25% of the population in Russia, and ending with the language of the Negidal, a small people group living on the Amur River. Several analysed studies dealt with the analysis of more than one languages at once, which allows them to cover a broader set of informational sources and compare sentiment expressed in different languages about the same topics. To classify sentiment in several languages, some studies translated the content into one language and performed basic monolingual sentiment analysis (e.g. [72]), while others developed multilingual sentiment classification models (e.g. [79]). As an extension of the latter approach, scholars can utilise pre-trained multilingual language models such as Bidirectional Encoder Representations from Transformers [43] and Multilingual Universal Sentence Encoder [198] for sentiment classification.

3) **General domain topics extraction from texts**. In the majority of studies dedicated to topic modelling, the authors selected only several topics for the extraction and further analysis. However, this approach is unable to deal with extraction relevant topics from the large textual datasets, e.g. when the share of text related to target topics is well below than 1% [54]. Moreover, topic modelling demonstrates poor accuracy of analysing short texts, especially in case if texts represent everyday talk [54]. The task of topics extraction

can be narrowed down not only to topic modelling but also to the task of text classification in case if there is a comprehensive training dataset for general domain topics extraction is available. The construction of such a dataset seems to be a time consuming and resource-intensive process in case of a basic annotation approach via a group of linguists or crowdsource annotations. However, some social networks provide users with the ability to specify relevant tags for their posts, e.g. Reddit[26] and Pikabu.[27] It means that users of such social networks take over the annotation process, so we assume that this data with an additional validation can be used for the construction of training dataset for general domain topics extraction from user posts.

4) **Likes and other reactions to the content as an indirect way of expressing sentiment**. The majority of analysed studies measured expressed sentiment only via the content of posts. However, likes and other reactions to a post potentially can be considered as an additional source of expressed sentiment by the viewers of the post. Therefore, it may be taken into account in the results of sentiment monitoring. Some preliminary work in the field of exploring the connection between liking behaviour and sentiment of the post was conducted in the paper [200], where authors examine the role of content of posts, relationship with the poster and personality of the user. Based on the online survey, the authors claimed that posts with positive sentiment tend to be liked automatically, i.e. without a thorough examination. It also was noticed that positivity of posts correlated with relational and literal motives. As an extension of the simple Like button, some social networks introduced reaction functionality to allow users easily share their emotional reaction to a post. For instance, the set of reactions on Facebook consists of pre-existing Like, Love, Wow, Haha, Angry and Sad. In their investigation into emotional stimuli in reacting behaviour of Russian-language Facebook users, Smoliarova *et al.* [201] show that the Love reaction is commonly used straightforwardly, thereby becoming the alternative of the traditional Like. Conversely, the post that evokes the Wow reaction may also be marked with other emoji reactions with a certain level of probability. Such reactions as Love, Haha and Wow tend to hamper the willingness to additionally interact with reacted posts using shares or comments [202]. Thus, the connection between reacting behaviour, people's mood and sentiment of the post tend to be a potentially significant research direction, which can be further utilised in the sentiment monitoring approaches.

5) **Contextual sentiment classification**. The context of ongoing dialogue can completely change the sentiment for a user response in comparison with the

---

[26]https://www.reddit.com/
[27]https://pikabu.ru/

sentiment when a response is examined as a standalone statement [203]. As a consequence, in the case of conversation sentiment analysis, e.g. analysis of comments responses, it is crucial to capture the context of the conversation in addition to the standalone sentiment of the response. Thus, researchers should draw their attention to the contextual sentiment classification in case they perform analysis of conversational data.

6) **Analysis of content from less explored data sources**. While a significant share of studies examines Vkontakte, Twitter, LiveJournal and YouTube, there are other widespread local social network sites which have high potential as data sources, e.g. Odnoklassniki, My World@Mail.Ru, and RuTube. In our opinion, researchers should draw their attention to the analysis of Odnoklassniki, that is, the second largest social network in Russia, which is used 42% of Russia's population [98]. In Russia, Odnoklassniki is widespread among the older audiences of 35-65+ years old so that it can be a useful platform for the analysis of the opinions of older generations. Moreover, the access to the representative historical data from Odnoklassniki can be requested directly through OK Data Science Lab, an exclusive platform developed by Odnoklassniki for the research purposes.

7) **Automatic content analysis as an alternative to traditional polls**. At the moment, the results of the analysis of online texts cannot be considered as a full-fledged alternative to the classical approaches for measuring opinions based on mass polls [204]. To overcome this uncertainty, a theoretical basis for generalizing data to more full groups of the population needed to be done [205]. A traditional mass survey involves associating opinions to socio-demographic groups, while in data from social media this reliable demographic information is commonly unavailable. To compare obtained results with the traditional opinion polls, researchers may utilise geolocation information, user profile information, and gender and age prediction systems [206]–[211] to compete with mass polls surveys.

8) **Monitoring of sentiment index of social media content in Russian**. In the cutting edge research of 2010 [212], Mislove *et al.* explore the pulse of mood throughout a day using over 300 million geolocated tweets from the United States of America using the lexicon-based approach. Some interesting trends were observed, for example, the level of happiness is the highest in the early morning and late evening. Through the week, weekends were much happier than weekdays. Identified patterns were confirmed by the study of Brazilians' Mood from Twitter [213], where Naive Bayes [30] was applied to classify sentiment. Dzogang also explored circadian patterns in mood changes [214]. While for many languages, such stud-

ies have already been conducted, the research of Russian-language content remains quite limited [93], [137]. It can be broadened and deepened in terms of analysed data volume, quality of sentiment classification model and methodology of social 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. For example, scholars from ITMO University described an approach to the emotional tonality assessment of public opinion [215]. In the paper [216], the authors discussed the general concept of social network monitoring through intelligent analysis of text messages. In the article [148], the author developed software for public mood monitoring through Twitter content in Russian. Sydorenko *et al.* developed a method for classifying time series of tonal ratings based on user posts from social networks [217].

VII. CONCLUSION

In this work, we surveyed existing applications of sentiment analysis for the Russian-language content. We identified five categories of studies based on utilised data source: *User-Generated Content from Social Network Sites*, *Product and Service Reviews*, *News from Mass Media*, *Books*, and *Mixed Data Sources*. We were able to further synthesise and systematically characterise existing identified studies by their purpose, employed sentiment analysis approach, and primary outcomes and limitations. Finally, this paper presents a research agenda to improve the quality of the applied sentiment analysis studies and to expand the existing research base to new directions.

The contributions of this survey to practice and research are fourfold. First, we outlined an existing knowledge base regarding applied studies on sentiment analysis of the Russian-language content. For each of the identified studies, we described the applied methods of sentiment analysis, major outcomes, and the most crucial drawbacks. Second, we analysed and summarised the most common and significant challenges that scholars were facing within their studies and proposed potential means to address. Third, to help scholars select an appropriate training dataset, we performed an additional literature review and identified publicly available sentiment datasets of Russian-language texts. Fourth, we contribute to research as we outlined potentials future research directions to improve applied studies on sentiment analysis of Russian texts.

REFERENCES

[1] A. E. O. Carosia, G. P. Coelho, and A. E. A. Silva, "Analyzing the Brazilian financial market through portuguese sentiment analysis in social media," *Appl. Artif. Intell.*, vol. 34, no. 1, pp. 1–19, Jan. 2020.

[2] S. Iacus, G. Porro, S. Salini, and E. Siletti, "An Italian composite subjective well-being index: The voice of Twitter users from 2012 to 2017," *Social Indicators Res.*, vol. 149, pp. 1–19, 2020.

[3] U. Sharma, R. K. Datta, and K. Pabreja, "Sentiment analysis and prediction of election results 2018," in *Social Networking and Computational Intelligence*, R. K. Shukla, J. Agrawal, S. Sharma, N. S. Chaudhari, and K. K. Shukla, Eds. Singapore: Springer, 2020, pp. 727–739.

[4] E. Georgiadou, S. Angelopoulos, and H. Drake, "Big data analytics and international negotiations: Sentiment analysis of brexit negotiating outcomes," *Int. J. Inf. Manage.*, vol. 51, Apr. 2020, Art. no. 102048. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0268401219309454

[5] L. Yue, W. Chen, X. Li, W. Zuo, and M. Yin, "A survey of sentiment analysis in social media," *Knowl. Inf. Syst.*, vol. 60, pp. 617–663, Jul. 2018.

[6] S. Tedmori and A. Awajan, "Sentiment analysis main tasks and applications: A survey," *J. Inf. Process. Syst.*, vol. 15, no. 3, pp. 500–519, 2019.

[7] M. Al-Ayyoub, A. A. Khamaiseh, Y. Jararweh, and M. N. Al-Kabi, "A comprehensive survey of arabic sentiment analysis," *Inf. Process. Manage.*, vol. 56, no. 2, pp. 320–342, Mar. 2019.

[8] D. M. Eberhard, G. F. Simons, and C. D. Fennig. (2020). *Ethnologue: Languages of the World*. Accessed: Apr. 22, 2020. [Online]. Available: http://www.ethnologue.com/

[9] GfK. (2018). *Penetration of Internet in Russia. The Results of 2017*. [Online]. Available: https://www.gfk.com/fileadmin/user_upload/dyna_content/RU/Documents/Press_Releases/2019/GfK_Rus_Internet_Audience_in_Russia_2018.pdf

[10] R. Vīksna and G. Jēkabsons, "Sentiment analysis in Latvian and Russian: A survey," *Appl. Comput. Syst.*, vol. 23, no. 1, pp. 45–51, May 2018.

[11] Q. Tul, M. Ali, A. Riaz, A. Noureen, M. Kamranz, B. Hayat, and A. Rehman, "Sentiment analysis using deep learning techniques: A review," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 6, pp. 424–433, 2017.

[12] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdiscip. Rev., Data Mining Knowl. Discovery*, vol. 8, no. 4, p. e1253, 2018. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1253

[13] H. H. Do, P. Prasad, A. Maag, and A. Alsadoon, "Deep learning for aspect-based sentiment analysis: A comparative review," *Expert Syst. Appl.*, vol. 118, pp. 272–299, Mar. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417418306456

[14] S. F. Maerz and C. Puschmann, "Text as data for conflict research: A literature survey," in *Computational Conflict Research*, E. Deutschmann, J. Lorenz, L. G. Nardin, D. Natalini, and A. F. X. Wilhelm, Eds. Cham, Switzerland: Springer, 2020, pp. 43–65, doi: 10.1007/978-3-030-29333-8_3.

[15] I. Chetviorkin and N. Loukachevitch, "Evaluating sentiment analysis systems in Russian," in *Proc. 4th Biennial Int. Workshop Balto-Slavic Natural Lang. Process*. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 12–17. [Online]. Available: https://www.aclweb.org/anthology/W13-2403

[16] N. V. Loukachevitch and I. I. Chetviorkin, "Open evaluation of sentiment-analysis systems based on the material of the Russian language," *Sci. Tech. Inf. Process.*, vol. 41, no. 6, pp. 370–376, Dec. 2014.

[17] N. Loukachevitch, P. Blinov, E. Kotelnikov, Y. Rubtsova, V. Ivanov, and E. Tutubalina, "SentiRuEval: Testing object-oriented sentiment analysis systems in Russian," in *Proc. Int. Conf. Dialogue Comput. Linguistics Intell. Technol.*, vol. 2, 2015, pp. 3–13.

[18] N. Lukashevich and Y. V. Rubtsova, "SentiRuEval-2016: Overcoming time gap and data sparsity in tweet sentiment analysis," in *Proc. Int. Conf. Dialogue Comput. Linguistics Intell. Technol.*, 2016, pp. 416–426.

[19] J. Trofimovich, "Comparison of neural network architectures for sentiment analysis of Russian tweets," in *Proc. Int. Conf. Dialogue Comput. Linguistics Intell. Technol.*, 2016, pp. 50–59.

[20] A. Zvonarev, "A comparison of machine learning methods of sentiment analysis based on Russian language Twitter data," in *Proc. 11th Majorov Int. Conf. Softw. Eng. Comput. Syst. (MICSECS)*, 2019, pp. 1–7.

[21] E. Kotelnikov, T. Peskisheva, A. Kotelnikova, and E. Razova, "A comparative study of publicly available Russian sentiment lexicons," in *Artificial Intelligence and Natural Language*, D. Ustalov, A. Filchenkov, L. Pivovarova, and J. Žižka, Eds. Cham, Switzerland: Springer, 2018, pp. 139–151.

[22] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment strength detection in short informal text," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 61, no. 12, pp. 2544–2558, Dec. 2010.

[23] C. J. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. 8th Int. AAAI Conf. Weblogs Social Media*, 2014, pp. 1–10.

[24] C. Gómez-Rodríguez, I. Alonso-Alonso, and D. Vilares, "How important is syntactic parsing accuracy? An empirical evaluation on rule-based sentiment analysis," *Artif. Intell. Rev.*, vol. 52, no. 3, pp. 2081–2097, Oct. 2019.

[25] N. Loukachevitch and A. Levchik, "Creating a general Russian sentiment lexicon," in *Proc. 10th Int. Conf. Lang. Resour. Eval. (REC)*. Portorož, Slovenia: European Language Resources Association, May 2016, pp. 1171–1176. [Online]. Available: https://www.aclweb.org/anthology/L16-1186

[26] O. Y. Koltsova, S. Alexeeva, and S. Kolcov, "An opinion word lexicon and a training dataset for Russian sentiment analysis of social media," in *Proc. Int. Conf. Dialogue Comput. Linguistics Intell. Technol.*, 2016, pp. 277–287.

[27] E. Cambria, S. Poria, R. Bajpai, and B. Schuller, "SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives," in *Proc. 26th Int. Conf. Comput. Linguistics, Tech. Papers (COLING)*. Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 2666–2677. [Online]. Available: https://www.aclweb.org/anthology/C16-1251

[28] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining," in *Proc. 7th Conf. Int. Lang. Resour. Eval. (LREC)*. Valletta, Malta: European Languages Resources Association, May 2010, pp. 1–5. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf

[29] L. Gatti, M. Guerini, and M. Turchi, "SentiWords: Deriving a high precision and high coverage lexicon for sentiment analysis," *IEEE Trans. Affect. Comput.*, vol. 7, no. 4, pp. 409–421, Oct. 2016, doi: 10.1109/TAFFC.2015.2476456.

[30] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to Information Retrieval*, vol. 39. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[31] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.

[32] D. G. Kleinbaum, K. Dietz, M. Gail, M. Klein, and M. Klein, *Logistic Regression*. Atlanta, GA, USA: Springer, 2002.

[33] S. R. Gunn, "Support vector machines for classification and regression," ISIS Tech. Rep., 1998, pp. 5–16, vol. 14, no. 1. [Online]. Available: http://ce.sharif.ir/courses/85-86/2/ce725/resources/root/LECTURES/SVM.pdf

[34] D. Tang, B. Qin, and T. Liu, "Deep learning for sentiment analysis: Successful approaches and future challenges," *Wiley Interdiscip. Rev., Data Mining Knowl. Discovery*, vol. 5, no. 6, pp. 292–303, Nov. 2015.

[35] M. Cliche, "BB_twtr at SemEval-2017 task 4: Twitter sentiment analysis with CNNs and LSTMs," in *Proc. 11th Int. Workshop Semantic Eval. (SemEval)*. Vancouver, BC, Canada: Association for Computational Linguistics, Aug. 2017, pp. 573–580. [Online]. Available: https://www.aclweb.org/anthology/S17-2094

[36] C. Baziotis, N. Pelekis, and C. Doulkeridis, "DataStories at SemEval-2017 task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis," in *Proc. 11th Int. Workshop Semantic Eval. (SemEval)*. Vancouver, BC, Canada: Association for Computational Linguistics, Aug. 2017, pp. 747–754. [Online]. Available: https://www.aclweb.org/anthology/S17-2126

[37] C. Baziotis, A. Nikolaos, P. Papalampidi, A. Kolovou, G. Paraskevopoulos, N. Ellinas, and A. Potamianos, "NTUA-SLP at SemEval-2018 task 3: Tracking ironic tweets using ensembles of word and character level attentive RNNs," in *Proc. 12th Int. Workshop Semantic Eval.* New Orleans, LA, USA: Association for Computational Linguistics, Jun. 2018, pp. 613–621. [Online]. Available: https://www.aclweb.org/anthology/S18-1100

[38] V. Duppada, R. Jain, and S. Hiray, "SeerNet at SemEval-2018 task 1: Domain adaptation for affect in tweets," in *Proc. 12th Int. Workshop Semantic Eval.* New Orleans, LA, USA: Association for Computational Linguistics, Jun. 2018, pp. 18–23. [Online]. Available: https://www.aclweb.org/anthology/S18-1002

[39] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. 26th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, vol. 2. Red Hook, NY, USA: Curran Associates, 2013, pp. 3111–3119.

[40] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. [Online]. Available: https://www.aclweb.org/anthology/D14-1162

[41] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics*, vol. 2. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 427–431. [Online]. Available: https://www.aclweb.org/anthology/E17-2068

[42] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1. New Orleans, LA, USA: Association for Computational Linguistics, Jun. 2018, pp. 2227–2237. [Online]. Available: https://www.aclweb.org/anthology/N18-1202

[43] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1. Minneapolis, MN, USA: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423

[44] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu, "ERNIE: Enhanced language representation with informative entities," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 1441–1451. [Online]. Available: https://www.aclweb.org/anthology/P19-1139

[45] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 5754–5764.

[46] A. Kumar, K. Srinivasan, W.-H. Cheng, and A. Y. Zomaya, "Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data," *Inf. Process. Manage.*, vol. 57, no. 1, Jan. 2020, Art. no. 102141. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0306457319306934

[47] D. Meškelė and F. Frasincar, "ALDONAr: A hybrid solution for sentence-level aspect-based sentiment analysis using a lexicalized domain ontology and a regularized neural attention model," *Inf. Process. Manage.*, vol. 57, no. 3, May 2020, Art. no. 102211. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0306457319310222

[48] M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, and I. Androutsopoulos, "SemEval-2015 task 12: Aspect based sentiment analysis," in *Proc. 9th Int. Workshop Semantic Eval. (SemEval)*. Denver, CO, USA: Association for Computational Linguistics, Jun. 2015, pp. 486–495. [Online]. Available: https://www.aclweb.org/anthology/S15-2082

[49] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. Al-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq, V. Hoste, M. Apidianaki, X. Tannier, N. Loukachevitch, E. Kotelnikov, N. Bel, S. M. Jiménez-Zafra, and G. Eryiğit, "SemEval-2016 task 5: Aspect based sentiment analysis," in *Proc. 10th Int. Workshop Semantic Eval. (SemEval)*. San Diego, CA, USA: Association for Computational Linguistics, Jun. 2016, pp. 19–30. [Online]. Available: https://www.aclweb.org/anthology/S16-1002

[50] J. Hu, S. Shi, and H. Huang, "Combining external sentiment knowledge for emotion cause detection," in *Natural Language Processing and Chinese Computing*, J. Tang, M.-Y. Kan, D. Zhao, S. Li, and H. Zan, Eds. Cham, Switzerland: Springer, 2019, pp. 711–722.

[51] M. M. Trusca, D. Wassenberg, F. Frasincar, and R. Dekker, "A hybrid approach for aspect-based sentiment analysis using deep contextual word embeddings and hierarchical attention," 2020, *arXiv:2004.08673*. [Online]. Available: http://arxiv.org/abs/2004.08673

[52] L. Yang, Y. Li, J. Wang, and R. S. Sherratt, "Sentiment analysis for E-Commerce product reviews in Chinese based on sentiment lexicon and deep learning," *IEEE Access*, vol. 8, pp. 23522–23530, 2020.

[53] I. Sabatovych, "Do social media create revolutions? Using Twitter sentiment analysis for predicting the maidan revolution in Ukraine," *Global Media Commun.*, vol. 15, no. 3, pp. 275–283, Dec. 2019, doi: 10.1177/1742766519872780.

[54] O. Koltsova, "Methodological challenges for detecting interethnic hostility on social media," in *Internet Science*, S. S. Bodrunova, O. Koltsova, A. Følstad, H. Halpin, P. Kolozaridi, L. Yuldashev, A. Smoliarova, and H. Niedermayer, Eds. Cham, Switzerland: Springer, 2019, pp. 7–18.

[55] O. Borodkina and V. Sibirev, "Migration issues in Russian Twitter: Attitudes to migrants, social problems and online resources," in *Internet Science*, S. El Yacoubi, F. Bagnoli, and G. Pacini, Eds. Cham, Switzerland: Springer, 2019, pp. 32–46.

[56] A. Tokarev, "Ukrainian elites discourse in respect of the donbass territory and population of 2009–2018: Analysis of the national facebook segment," *MGIMO Rev. Int. Relations*, vol. 6, no. 63, pp. 194–211, Dec. 2018.

[57] O. Koltsova and O. Nagornyy, "Redefining media agendas: Topic problematization in online reader comments," *Media Commun.*, vol. 7, no. 3, pp. 145–156, 2019.

[58] N. Rulyova, "Russian new media users' reaction to a meteor explosion in Chelyabinsk: Twitter versus YouTube," in *Emerging Genres in New Media Environments*. Cham, Switzerland: Palgrave Macmillan, 2017, pp. 79–97.

[59] O. Koltsova and S. Koltcov, "Mapping the public agenda with topic modeling: The case of the Russian livejournal," *Policy Internet*, vol. 5, no. 2, pp. 207–227, Jun. 2013.

[60] A. P. Kirilenko and S. O. Stepchenkova, "Public microblogging on climate change: One year of Twitter worldwide," *Global Environ. Change*, vol. 26, pp. 171–182, May 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0959378014000375

[61] C. Boussalis, T. G. Coan, and M. Poberezhskaya, "Measuring and modeling Russian newspaper coverage of climate change," *Global Environ. Change*, vol. 41, pp. 99–110, Nov. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0959378016302151

[62] Y. Rykov, O. Nagornyy, and O. Koltsova, "Digital inequality in Russia through the use of a social network site: A cross-regional comparison," in *Digital Transformation and Global Society*, D. A. Alexandrov, A. V. Boukhanovsky, A. V. Chugunov, Y. Kabanov, and O. Koltsova, Eds. Cham, Switzerland: Springer, 2017, pp. 70–83.

[63] K. D. Mukhina, S. V. Rakitin, and A. A. Visheratin, "Detection of tourists attraction points using instagram profiles," *Procedia Comput. Sci.*, vol. 108, pp. 2378–2382, Jan. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050917306981

[64] A. Shirokanova and O. Silyutina, "Internet regulation media coverage in Russia: Topics and countries," in *Proc. 10th ACM Conf. Web Sci. (WebSci)*. New York, NY, USA: Association for Computing Machinery, 2018, pp. 359–363, doi: 10.1145/3201064.3201102.

[65] A. Filippov, V. Moshkin, and N. Yarushkina, "Development of a software for the semantic analysis of social media content," in *Recent Research in Control Engineering and Decision Making*, O. Dolinina, A. Brovko, V. Pechenkin, A. Lvov, V. Zhmud, and V. Kreinovich, Eds. Cham, Switzerland: Springer, 2019, pp. 421–432.

[66] A. Uteuov, "Topic model for online communities' interests prediction," *Procedia Comput. Sci.*, vol. 156, pp. 204–213, Jan. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050919311159

[67] L. Gokhberg, I. Kuzminov, E. Khabirova, and T. Thurner, "Advanced text-mining for trend analysis of Russia's extractive industries," *Futures*, vol. 115, Jan. 2020, Art. no. 102476. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0016328719303386

[68] E. Shchekotin, M. Myagkov, V. Goiko, V. Kashpur, and G. Kovarzh, "Subjective measurement of population ill-being/well-being in the Russian regions based on social media data," *Monit. Public Opinion, Econ. Social Changes*, vol. 155, no. 1, pp. 78–116, 2020.

[69] K. Svetlov and K. Platonov, "Sentiment analysis of posts and comments in the accounts of Russian politicians on the social network," in *Proc. 25th Conf. Open Innov. Assoc. (FRUCT)*, Nov. 2019, pp. 299–305.

[70] D. Sokolov, L. Selivanovskikh, E. Zavyalova, and M. Latukha, "Why employees leave Russian companies? Analyzing online job reviews using text mining," *Russian Manage. J.*, vol. 16, no. 4, pp. 499–512, 2018.

[71] Y. Seliverstov, K. Nikitin, N. Shatalova, and A. Kiselev, "Road pavement assessment of the north-west federal district using sentiment analysis of the Internet user reviews," *Russian Manage. J.*, vol. 13, no. 3, pp. 7–24, 2019.

[72] W. Li and H. Chen, "Identifying top sellers in underground economy using deep learning-based sentiment analysis," in *Proc. IEEE Joint Intell. Secur. Informat. Conf.*, Sep. 2014, pp. 64–67.

[73] R. L. Kaplan, *Politics and the American Press: The Rise of Objectivity 1865–1920*. Cambridge, U.K.: Cambridge Univ. Press, 2002.

[74] A. Balahur, R. Steinberger, M. Kabadjov, V. Zavarella, E. van der Goot, M. Halkia, B. Pouliquen, and J. Belyaeva, "Sentiment analysis in the news," in *Proc. 7th Int. Conf. Lang. Resour. Eval. (LREC)*. Valletta, Malta: European Language Resources Association (ELRA), May 2010, pp. 1–5.

[75] A. Kazun and A. Kazun, "A friend who was supposed to lose: How Donald trump was portrayed in the Russian media?" Higher School Econ., Moscow, Russia, Res. Paper WP BRP, 2017, vol. 51.

[76] K. Yakovleva, "Text mining-based economic activity estimation," *Russian J. Money Finance*, vol. 77, no. 4, pp. 26–41, Dec. 2018.

[77] V. Solovyev, M. Solnyshkina, E. Gafiyatova, D. McNamara, and V. Ivanov, "Sentiment in academic texts," in *Proc. 24th Conf. Open Innov. Assoc. (FRUCT)*, Apr. 2019, pp. 408–414.

[78] A. V. Kolmogorova, "Emotional tonality as a valuable subjective parameter of study text for Russian as foreign language learners," *Philol. Class*, vol. 57, no. 3, pp. 95–101, 2019.

[79] A. F. Brantly, "From cyberspace to independence square: Understanding the impact of social media on physical protest mobilization during Ukraine's Euromaidan revolution," *J. Inf. Technol. Politics*, vol. 16, no. 4, pp. 360–378, Oct. 2019.

[80] A. Kazun, "To cover or not to cover: Alexei Navalny in Russian media," *Int. Area Stud. Rev.*, vol. 22, no. 4, pp. 312–326, Dec. 2019.

[81] S. S. Bodrunova, O. Koltsova, S. Koltcov, and S. Nikolenko, "Who's bad? Attitudes toward resettlers from the post-soviet south versus other nations in the Russian blogosphere," *Int. J. Commun.*, vol. 11, pp. 3242–3264, Aug. 2017.

[82] O. Koltsova, S. Nikolenko, S. Alexeeva, O. Nagornyy, and S. Koltcov, "Detecting interethnic relations with the data from social media," in *Digital Transformation and Global Society*, D. A. Alexandrov, A. V. Boukhanovsky, A. V. Chugunov, Y. Kabanov, and O. Koltsova, Eds. Cham, Switzerland: Springer, 2017, pp. 16–30.

[83] O. Koltsova, S. Alexeeva, S. Nikolenko, and M. Koltsov, "Measuring prejudice and ethnic tensions in user-generated content," *Annu. Rev. Cybertherapy Telemed.*, vol. 2017, pp. 76–81, Jun. 2017.

[84] O. Nagornyy, "Topics of ethnic discussions in Russian social media," in *Digital Transformation and Global Society*, D. A. Alexandrov, A. V. Boukhanovsky, A. V. Chugunov, Y. Kabanov, and O. Koltsova, Eds. Cham, Switzerland: Springer, 2018, pp. 83–94.

[85] D. Duvanova, A. Nikolaev, A. Nikolsko-Rzhevskyy, and A. Semenov, "Violent conflict and online segregation: An analysis of social network communication across Ukraine's regions," *J. Comparative Econ.*, vol. 44, no. 1, pp. 163–181, Feb. 2016.

[86] S. Volkova, I. Chetviorkin, D. Arendt, and B. Van Durme, "Contrasting public opinion dynamics and emotional response during crisis," in *Social Informatics*, E. Spiro and Y.-Y. Ahn, Eds. Cham, Switzerland: Springer, 2016, pp. 312–329.

[87] A. Rumshisky, M. Gronas, P. Potash, M. Dubov, A. Romanov, S. Kulshreshtha, and A. Gribov, "Combining network and language indicators for tracking conflict intensity," in *Social Informatics*, G. L. Ciampaglia, A. Mashhadi, and T. Yasseri, Eds. Cham, Switzerland: Springer, 2017, pp. 391–404.

[88] A. Zaezjev, "Understanding political mobilization through social media content analysis: Facebook and vkontakte in the first days," in *Proc. Int. Conf. Dialogue Comput. Linguistics Intell. Technol.*, 2018, pp. 1–12.

[89] D. Donchenko, N. Ovchar, N. Sadovnikova, D. Parygin, O. Shabalina, and D. Ather, "Analysis of comments of users of social networks to assess the level of social tension," *Procedia Comput. Sci.*, vol. 119, pp. 359–367, Jan. 2017.

[90] A. P. Kirilenko and S. O. Stepchenkova, "Sochi 2014 olympics on Twitter: Perspectives of hosts and guests," *Tourism Manage.*, vol. 63, pp. 54–65, Dec. 2017.

[91] V. Spaiser, T. Chadefaux, K. Donnay, F. Russmann, and D. Helbing, "Communication power struggles on social media: A case study of the 2011–12 Russian protests," *J. Inf. Technol. Politics*, vol. 14, no. 2, pp. 132–153, Apr. 2017.

[92] A. Nenko and M. Petrova, "Comparing PPGIS and LBSN data to measure emotional perception of the city," in *Digital Transformation and Global Society*, D. A. Alexandrov, A. V. Boukhanovsky, A. V. Chugunov, Y. Kabanov, O. Koltsova, and I. Musabirov, Eds. Cham, Switzerland: Springer, 2019, pp. 223–234.

[93] A. Zaezjev, "Sentiment index of the Russian speaking Facebook," in *Proc. Int. Conf. Dialogue Comput. Linguistics Intell. Technol.*, vol. 13, 2014, pp. 506–517.

[94] M. Belyakov, "The analysis of news messages of the site of the Russian Federation Ministry of Foreign Affairs applying content-analysis (article 1)," *RUDN J. Lang. Stud., Semiotics Semantics*, vol. 7, no. 3, pp. 58–67, 2016.

[95] M. Belyakov, "The analysis of news messages of the RF Ministry of Foreign Affairs Web-site by the sentiment-analysis (article 2)," *RUDN J. Lang. Stud., Semiotics Semantics*, vol. 7, no. 4, pp. 115–124, 2016.

[96] I. V. Khramoin, M. A. Kamenskaya, I. A. Tikhomirov, and N. V. Toganova, "Sentiment analysis of innovations in Russian media," in *Proc. 2nd Russia Pacific Conf. Comput. Technol. Appl. (RPC)*, Sep. 2017, pp. 96–99.

[97] B. Etling, "Russia, Ukraine, and the west: Social media sentiment in the Euromaidan protests," Berkman Center Res. Publication, Cambridge, U.K., Tech. Rep., 2014. Accessed: Jun. 12, 2020. [Online]. Available: http://cyber.law.harvard.edu/publications/2014/euromaidan

[98] *Media Consumption in Russia 2018*, Deloitte CIS Res. Center, Russia, Moscow, 2018.

[99] S. Gibson, *Impact of Communication and the Media on Ethnic Conflict*. Hershey, PA, USA: IGI Global, 2015.

[100] J. Chan, A. Ghose, and R. Seamans, "The Internet and racial hate crime: Offline spillovers from online access," *MIS Quart.*, vol. 40, no. 2, pp. 381–403, 2016.

[101] S. Bodrunova, S. Koltsov, O. Koltsova, S. Nikolenko, and A. Shimorina, "Interval semi-supervised LDA: Classifying needles in a haystack," in *Advances in Artificial Intelligence and Its Applications*, F. Castro, A. Gelbukh, and M. González, Eds. Berlin, Germany: Springer, 2013, pp. 265–274.

[102] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[103] M. Apishev, S. Koltcov, O. Koltsova, S. Nikolenko, and K. Vorontsov, "Mining ethnic content online with additively regularized topic models," *Computación Sistemas*, vol. 20, no. 3, pp. 387–403, Sep. 2016.

[104] S. I. Nikolenko, S. Koltcov, and O. Koltsova, "Topic modelling for qualitative studies," *J. Inf. Sci.*, vol. 43, no. 1, pp. 88–102, Feb. 2017, doi: 10.1177/0165551515617393.

[105] A. Semenov and J. Veijalainen, "A modelling framework for social media monitoring," *Int. J. Web Eng. Technol.*, vol. 8, no. 3, pp. 217–249, Oct. 2013, doi: 10.1504/IJWET.2013.057226.

[106] A. Semenov, J. Veijalainen, and A. Boukhanovsky, "A generic architecture for a social network monitoring and analysis system," in *Proc. 14th Int. Conf. Netw.-Based Inf. Syst.*, Sep. 2011, pp. 178–185.

[107] E. S. Kuznetsova, N. V. Loukachevitch, and I. I. Chetviorkin, "Testing rules for a sentiment analysis system," in *Proc. Int. Conf. Dialogue Comput. Linguistics Intell. Technol.*, 2013, pp. 71–80.

[108] N. V. Loukachevitch and I. I. Chetviorkin, "Refinement of Russian sentiment lexicons using Ruthes thesaurus," in *Proc. 16th All-Russian Conf. Digit. Libraries, Adv. Methods Technol., Digit. Collections*, 2014, pp. 61–65.

[109] S. M. Mohammad and S. Kiritchenko, "Using hashtags to capture fine emotion categories from tweets," *Comput. Intell.*, vol. 31, no. 2, pp. 301–326, May 2015. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/coin.12024

[110] S. Volkova and Y. Bachrach, "On predicting sociodemographic traits and emotions from communications in social networks and their implications to online self-disclosure," *Cyberpsychol., Behav., Social Netw.*, vol. 18, no. 12, pp. 726–736, Dec. 2015.

[111] P. Ekman, "An argument for basic emotions," *Cognition Emotion*, vol. 6, nos. 3–4, pp. 169–200, May 1992.

[112] N. Godbole, M. Srinivasaiah, and S. Skiena, "Large-scale sentiment analysis for news and blogs," in *Proc. ICWSM*, 2007, vol. 7, no. 21, pp. 219–222.

[113] G. Artemov, A. Aleinikov, D. Abgadzhava, A. Pinkevich, and A. Abalian, "Social tension: The possibility of conflict diagnosis (on the example of St. Petersburg)," *Econ. Sociol.*, vol. 10, no. 1, p. 192, 2017.

[114] M. L. Williams, A. Edwards, W. Housley, P. Burnap, O. Rana, N. Avis, J. Morgan, and L. Sloan, "Policing cyber-neighbourhoods: Tension monitoring and social media networks," *Policing Soc.*, vol. 23, no. 4, pp. 461–481, Dec. 2013.

[115] P. Burnap, O. F. Rana, N. Avis, M. Williams, W. Housley, A. Edwards, J. Morgan, and L. Sloan, "Detecting tension in online communities with computational Twitter analysis," *Technol. Forecasting Social Change*, vol. 95, pp. 96–108, Jun. 2015.

[116] A. Rajaraman and J. D. Ullman, *Mining of Massive Datasets*. Cambridge, MA, USA: Cambridge Univ. Press, 2011.

[117] R. Rehurek and P. Sojka, "Software framework for topic modelling with large corpora," in *Proc. LREC Workshop New Challenges NLP Frameworks*, 2010, pp. 45–50.

[118] R. Arun, V. Suresh, C. E. V. Madhavan, and M. N. N. Murthy, "On finding the natural number of topics with latent Dirichlet allocation: Some observations," in *Advances in Knowledge Discovery and Data Mining*, M. J. Zaki, J. X. Yu, B. Ravindran, and V. Pudi, Eds. Berlin, Germany: Springer, 2010, pp. 391–402.

[119] M. M. Bakhtin, "Speech genres and other late essays, trans," in *Caryl Emerson and Michael Holquist*, vol. 103. Austin, TX, USA: Univ. Texas Press, 1986.

[120] A. P. Kirilenko, "Data for sochi 2014 olympics discussion on social media," *Data Brief*, vol. 13, pp. 605–608, Aug. 2017.

[121] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Seattle, WA, USA: Association for Computational Linguistics, Oct. 2013, pp. 1631–1642. [Online]. Available: https://www.aclweb.org/anthology/D13-1170

[122] J. Saldaña, *The Coding Manual for Qualitative Researchers*. Newbury Park, CA, USA: Sage, 2015.

[123] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[124] A. Van Hoorn, "A short introduction to subjective well-being: Its measurement, correlates and policy uses," in *Proc. Int. Conf., Happiness Measurable What Do Those Measures Mean Policy*, Rome, Italy, Apr. 2007, pp. 2–7.

[125] P. Killworth and H. Bernard, "Informant accuracy in social network data," *Human Org.*, vol. 35, no. 3, pp. 269–286, Sep. 1976.

[126] C. Martinelli and S. W. Parker, "Deception and misreporting in a social program," *J. Eur. Econ. Assoc.*, vol. 7, no. 4, pp. 886–908, Jun. 2009.

[127] P. S. Dodds and C. M. Danforth, "Measuring the happiness of large-scale written expression: Songs, blogs, and presidents," *J. Happiness Stud.*, vol. 11, no. 4, pp. 441–456, Aug. 2010.

[128] A. D. Kramer, "An unobtrusive behavioral model of gross national happiness," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst. (CHI)*. New York, NY, USA: Association for Computing Machinery, 2010, pp. 287–290, doi: 10.1145/1753326.1753369.

[129] P. S. Dodds, K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth, "Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter," *PLoS ONE*, vol. 6, no. 12, Dec. 2011, Art. no. e26752.

[130] J. Qi, X. Fu, and G. Zhu, "Subjective well-being measurement based on Chinese grassroots blog text sentiment analysis," *Inf. Manage.*, vol. 52, no. 7, pp. 859–869, Nov. 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378720615000609

[131] E. Sulis, C. Bosco, V. Patti, M. Lai, D. I. H. Farías, L. Mencarini, M. Mozzachiodi, and D. Vignoli, "Subjective well-being and social media: A semantically annotated Twitter corpus on fertility and parenthood," in *Proc. 3rd Italian Conf. Comput. Linguistics, CLiC 5th Eval. Campaign Natural Lang. Process. Speech Tools Italian, EVALITA*, vol. 1749, 2016, pp. 1–6.

[132] L. Chen, T. Gong, M. Kosinski, D. Stillwell, and R. L. Davidson, "Building a profile of subjective well-being for social media users," *PLoS ONE*, vol. 12, no. 11, Nov. 2017, Art. no. e0187278.

[133] J. Bollen and B. Gonçalves, "Network happiness: How online social interactions relate to our well being," in *Complex Spreading Phenomena in Social Systems: Influence and Contagion in Real-World Social Networks*, S. Lehmann and Y.-Y. Ahn, Eds. Cham, Switzerland: Springer, 2018, pp. 257–268, doi: 10.1007/978-3-319-77332-2_14.

[134] M. Lui and T. Baldwin, "Langid.py: An off-the-shelf language identification tool," in *Proc. ACL Syst. Demonstrations*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 25–30.

[135] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Retr.*, vol. 2, nos. 1–2, pp. 1–135, 2008.

[136] B. Liu, "Sentiment analysis and subjectivity," *Handbook Natural Lang. Process.*, vol. 2, pp. 627–666, Mar. 2010.

[137] T. V. Gavrilova, "Principles and methods of research quality of life," *Qual. Life Technol.*, vol. 4, no. 2, pp. 1–11, 2004.

[138] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS)*. Red Hook, NY, USA: Curran Associates, 2017, pp. 3149–3157.

[139] S. Waldron. (2010). *Measuring Subjective Wellbeing in the UK*. Newport. Office Nat. Statist. Accessed: Jun. 12, 2020. [Online]. Available: http://www.mas.org.uk/uploads/artlib/measuring-subjective-wellbeing-in-the-uk

[140] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," in *Proc. 8th Workshop Syntax, Semantics Struct. Stat. Transl. (SSST)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 103–111. [Online]. Available: https://www.aclweb.org/anthology/W14-4012

[141] Y. V. Rubtsova, "A method for development and analysis of short text corpus for the review classification task," in *Proc. Trudy XV Vserossiiskoy Naychnoy Konferencii RCDL*, 2013, pp. 269–275.

[142] A. Rogers, A. Romanov, A. Rumshisky, S. Volkova, M. Gronas, and A. Gribov, "RuSentiment: An enriched sentiment analysis dataset for social media in Russian," in *Proc. 27th Int. Conf. Comput. Linguistics*. Santa Fe, NM, USA: Association for Computational Linguistics, Aug. 2018, pp. 755–763. [Online]. Available: https://www.aclweb.org/anthology/C18-1064

[143] S. Smetanin and M. Komarov, "Sentiment analysis of product reviews in Russian using convolutional neural networks," in *Proc. IEEE 21st Conf. Bus. Informat. (CBI)*, vol. 1, Jul. 2019, pp. 482–486.

[144] J. Read, "Using emoticons to reduce dependency in machine learning techniques for sentiment classification," in *Proc. ACL Student Res. Workshop*. Ann Arbor, MI, USA: Association for Computational Linguistics, Jun. 2005, pp. 43–48. [Online]. Available: https://www.aclweb.org/anthology/P05-2008

[145] Y. Rubtsova, "Reducing the deterioration of sentiment analysis results due to the time impact," *Information*, vol. 9, no. 8, p. 184, Jul. 2018.

[146] O. S. Smirnova and V. V. Shishkov, "The choice of the topology of neural networks and their use for the classification of small texts," *Int. J. Open Inf. Technol.*, vol. 4, no. 8, pp. 50–54, 2016.

[147] V. Garshina, K. Kalabukhov, V. Stepantsov, and S. Smotrov, "Development of the system of sentiment analysis of the text," *Proc. Voronezh State Univ. Ser., Syst. Anal. Inf. Technol.*, vol. 3, pp. 185–194, Jul. 2017.

[148] S. I. Smetanin, "The program for public mood monitoring through Twitter content in Russia," *Proc. Inst. Syst. Program. RAS*, vol. 29, no. 4, pp. 315–324, 2017.

[149] A. Romanov, M. Vasilieva, A. Kurtukova, and R. Meshcheryakov, "Sentiment analysis of text using machine learning techniques," in *Proc. R. Piotrowski's Readings Lang. Eng. Appl. Linguistics*, 2018, pp. 86–95.

[150] K. Lagutina, V. Larionov, V. Petryakov, N. Lagutina, and I. Paramonov, "Sentiment classification of Russian texts using automatically generated thesaurus," in *Proc. 23rd Conf. Open Innov. Assoc. (FRUCT)*, Nov. 2018, pp. 217–222.

[151] M. Stephens, *A History of News*. Forth Worth, TX, USA: Harcourt Brace College Publishers, 1997.

[152] L. M. Salmon, *The Newspaper and the Historian*. London, U.K.: Oxford Univ. Press, 1923.

[153] R. P. Schumaker, Y. Zhang, C.-N. Huang, and H. Chen, "Evaluating sentiment in financial news articles," *Decis. Support Syst.*, vol. 53, no. 3, pp. 458–464, Jun. 2012.

[154] A. Yadav, C. K. Jha, A. Sharan, and V. Vaish, "Sentiment analysis of financial news using unsupervised and supervised approach," in *Pattern Recognition and Machine Intelligence*, B. Deka, P. Maji, S. Mitra, D. K. Bhattacharyya, P. K. Bora, and S. K. Pal, Eds. Cham, Switzerland: Springer, 2019, pp. 311–319.

[155] S. Deng, T. Mitsubuchi, K. Shioda, T. Shimada, and A. Sakurai, "Combining technical analysis with sentiment analysis for stock price prediction," in *Proc. IEEE 9th Int. Conf. Dependable, Auton. Secure Comput.*, Dec. 2011, pp. 800–807.

[156] A. E. Khedr and N. Yaseen, "Predicting stock market behavior using data mining technique and news sentiment analysis," *Int. J. Intell. Syst. Appl.*, vol. 9, no. 7, p. 22, 2017.

[157] S. Unankard, X. Li, M. Sharaf, J. Zhong, and X. Li, "Predicting elections from social networks based on sub-event detection and sentiment analysis," in *Web Information Systems Engineering*, B. Benatallah, A. Bestavros, Y. Manolopoulos, A. Vakali, and Y. Zhang, Eds. Cham, Switzerland: Springer, 2014, pp. 1–16.

[158] P. Sharma and T.-S. Moh, "Prediction of Indian election using sentiment analysis on hindi Twitter," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2016, pp. 1966–1971.

[159] K.-K. Tseng, R. F.-Y. Lin, H. Zhou, K. J. Kurniajaya, and Q. Li, "Price prediction of e-commerce products through Internet sentiment analysis," *Electron. Commerce Res.*, vol. 18, no. 1, pp. 65–88, Mar. 2018.

[160] G. Osipov, I. Smirnov, I. Tikhomirov, I. Sochenkov, and A. Shelmanov, "Exactus expert—Search and analytical engine for research and development support," in *Novel Applications of Intelligent Systems*. Cham, Switzerland: Springer, 2016, pp. 269–285, doi: 10.1007/978-3-319-14194-7_14.

[161] M. Sabou, K. Bontcheva, L. Derczynski, and A. Scharl, "Corpus annotation through crowdsourcing: Towards best practice guidelines," in *Proc. 9th Int. Conf. Lang. Resour. Eval. (LREC)*. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 859–866. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/497_Paper.pdf

[162] S. Mohammad, "A practical guide to sentiment annotation: Challenges and solutions," in *Proc. 7th Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal.* San Diego, CA, USA: Association for Computational Linguistics, Jun. 2016, pp. 174–179. [Online]. Available: https://www.aclweb.org/anthology/W16-0429

[163] J. Sell and I. G. Farreras, "LIWC-ing at a century of introductory college textbooks: Have the sentiments changed?" *Procedia Comput. Sci.*, vol. 118, pp. 108–112, Jan. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050917323542

[164] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *Proc. Conf. New Methods Lang. Process.*, Manchester, U.K., 1994, pp. 44–49.

[165] H. Lövheim, "A new three-dimensional model for emotions and monoamine neurotransmitters," *Med. Hypotheses*, vol. 78, no. 2, pp. 341–348, Feb. 2012.

[166] P. Hitlin, "Methodology: How crimson hexagon Wworks," Pew Res. Center, Washington, DC, USA, Tech. Rep., 2015. Accessed: Jun. 12, 2020. [Online]. Available: http://www.journalism.org/2015/04/01/methodology-crimson-hexagon/

[167] D. J. Hopkins and G. King, "A method of automated nonparametric content analysis for social science," *Amer. J. Political Sci.*, vol. 54, no. 1, pp. 229–247, Jan. 2010.

[168] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[169] K. Leetaru and P. A. Schrodt, "Gdelt: Global database of events, language, and tone," in *Proc. ISA Annu. Conv.*, San Francisco, CA, USA, Apr. 2013.

[170] V. Rybakov and A. Malafeev, "Aspect-based sentiment analysis of Russian hotel reviews," in *Proc. 7th Int. Conf. Anal. Images, Social Netw. Texts (AIST)*, 2018, pp. 75–84.

[171] N. Rusnachenko and N. V. Loukachevitch, "Extracting sentiment attitudes from analytical texts via piecewise convolutional neural network," in *Proc. 20th Int. Conf. Data Anal. Manage. Data Intensive Domains (DAMDID/RCDL)*, 2018, pp. 186–192.

[172] I. Mozetivc, M. Grčar, and J. Smailović, "Multilingual Twitter sentiment classification: The role of human annotators," *PLoS ONE*, vol. 11, no. 5, May 2016, Art. no. e0155036.

[173] V. Bobichev, O. Kanishcheva, and O. Cherednichenko, "Sentiment analysis in the Ukrainian and Russian news," in *Proc. IEEE 1st Ukraine Conf. Electr. Comput. Eng. (UKRCON)*, May 2017, pp. 1050–1055.

[174] I. Chetviorkin, P. Braslavskiy, and N. Loukachevich, "Sentiment analysis track at ROMIP 2011," in *Proc. Int. Conf. Dialogue Comput. Linguistics Intell. Technol.*, vol. 2, 2012, pp. 1–14.

[175] I. I. Chetvirokin and N. V. Loukachevitch, "Sentiment analysis track at ROMIP 2012," in *Proc. Int. Conf. Dialogue Comput. Linguistics Intell. Technol.*, vol. 2, 2013, pp. 1–11.

[176] S. M. Mohammad, "Challenges in sentiment analysis," in *A Practical Guide to Sentiment Analysis*, E. Cambria, D. Das, S. Bandyopadhyay, and A. Feraco, Eds. Cham, Switzerland: Springer, 2017, pp. 61–83, doi: 10.1007/978-3-319-55394-8_4.

[177] D. Ustalov and S. Igushkin, "Sense inventory alignment using lexical substitutions and crowdsourcing," in *Proc. Int. FRUCT Conf. Intell., Social Media Web (ISMW FRUCT)*, Sep. 2016, pp. 1–6.

[178] M. Ponomareva, K. Milintsevich, E. Chernyak, and A. Starostin, "Automated word stress detection in Russian," in *Proc. 1st Workshop Subword Character Level Models NLP*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 31–35. [Online]. Available: https://www.aclweb.org/anthology/W17-4104

[179] A. Panchenko, A. Lopukhina, D. Ustalov, N. Arefyev, A. Leontyev, and N. Loukachevitch, "Russe'2018: A shared task on word sense induction for the Russian language," in *Proc. Int. Conf. Dialogue Comput. Linguistics Intell. Technol.*, vol. 17, 2018, pp. 1–18.

[180] E. Chernyak, M. Ponomareva, and K. Milintsevich, "Char-RNN for word stress detection in east slavic languages," in *Proc. 6th Workshop NLP Similar Lang., Varieties Dialects*, 2019, pp. 35–41.

[181] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychol. Bull.*, vol. 76, no. 5, p. 378, 1971.

[182] K. Krippendorff. *Computing Krippendorff's Alpha Reliability*. Accessed: Jun. 12, 2020. [Online]. Available: https://repository.upenn.edu/asc_papers/43/

[183] G. Blank, "The digital divide among Twitter users and its implications for social research," *Social Sci. Comput. Rev.*, vol. 35, no. 6, pp. 679–697, Dec. 2017.

[184] A. Shahbaz. (2018). *Freedom on the Net 2018: The Rise of Digital Authoritarianism*. Freedom New. Freedom House. Accessed: Jun. 12, 2020. [Online]. Available: https://freedomhouse.org/report/freedom-net/2018/rise-digital-authoritarianism

[185] N. Duffy, "Internet freedom in Vladimir Putin's Russia: The noose tightens," *Amer. Enterprise Inst.*, vol. 12, pp. 1–12, Jan. 2015.

[186] T. Al-Moslmi, N. Omar, S. Abdullah, and M. Albared, "Approaches to cross-domain sentiment analysis: A systematic literature review," *IEEE Access*, vol. 5, pp. 16173–16192, 2017.

[187] K. Ravi and V. Ravi, "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications," *Knowl.-Based Syst.*, vol. 89, pp. 14–46, Nov. 2015.

[188] B. Liu, *Sentiment Analysis and Opinion Mining*. San Rafael, CA, USA: Morgan & Claypool, 2012.

[189] T. Zefirova and N. Loukachevitch, "Irony and sarcasm expression in Twitter," in *Proc. 3rd Workshop Comput. Linguistics Lang. Sci.*, in EPiC Series in Language and Linguistics, vol. 4, G. Wohlgenannt, R. von Waldenfels, S. Toldova, E. Rakhilina, D. Paperno, O. Lyashevskaya, N. Loukachevitch, S. O. Kuznetsov, O. Kultepina, D. Ilvovsky, B. Galitsky, E. Artemova, and E. Bolshakova, Eds. Manchester, U.K.: EasyChair, 2019, pp. 45–49.

[190] F. Morstatter, L. Wu, T. H. Nazer, K. M. Carley, and H. Liu, "A new approach to bot detection: Striking the balance between precision and recall," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2016, pp. 533–540.

[191] M. Ibrahim, O. Abdillah, A. F. Wicaksono, and M. Adriani, "Buzzer detection and sentiment analysis for predicting presidential election results in a Twitter nation," in *Proc. IEEE Int. Conf. Data Mining Workshop (ICDMW)*, Nov. 2015, pp. 1348–1353.

[192] A. Mitchell and P. Hitlin, "Twitter reaction to events often at odds with overall public opinion," Pew Res. Center, Washington, DC, USA, Tech. Rep., 2013. Accessed: Jun. 12, 2020. [Online]. Available: https://www.pewresearch.org/2013/03/04/twitter-reaction-to-events-often-at-odds-with-overall-public-opinion/

[193] D. J. S. Oliveira, P. H. D. S. Bermejo, and P. A. dos Santos, "Can social media reveal the preferences of voters? A comparison between sentiment analysis and traditional opinion polls," *J. Inf. Technol. Politics*, vol. 14, no. 1, pp. 34–45, Jan. 2017.

[194] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. (2018). *Improving Language Understanding by Generative Pre-Training*. [Online]. Available: https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/languageunderstandingpaper.pdf

[195] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2018, pp. 328–339.

[196] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune BERT for text classification?" in *Chinese Computational Linguistics*, M. Sun, X. Huang, H. Ji, Z. Liu, and Y. Liu, Eds. Cham, Switzerland: Springer, 2019, pp. 194–206.

[197] D. Cer, Y. Yang, S.-Y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y.-H. Sung, B. Strope, and R. Kurzweil, "Universal sentence encoder," in *Proc. EMNLP Demonstration*, Brussels, Belgium, 2018, pp. 1–7. [Online]. Available: https://arxiv.org/abs/1803.11175

[198] Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. H. Abrego, S. Yuan, C. Tar, Y.-H. Sung, B. Strope, and R. Kurzweil, "Multilingual universal sentence encoder for semantic retrieval," 2019, *arXiv:1907.04307*. [Online]. Available: http://arxiv.org/abs/1907.04307

[199] D. R. Baymurzina, D. P. Kuznetsov, and M. S. Burtsev, "Language model embeddings improve sentiment analysis in Russian," in *Proc. Int. Conf. Dialogue Comput. Linguistics Intell. Technol.*, vol. 18, 2019, pp. 53–63.

[200] A. Levordashka, S. Utz, and R. Ambros, "What's in a like? Motivations for pressing the like button," in *Proc. 10th Int. AAAI Conf. Web Social Media*, 2016, pp. 623–626.

[201] A. S. Smoliarova, T. M. Gromova, and N. A. Pavlushkina, "Emotional stimuli in social media user behavior: Emoji reactions on a news media Facebook page," in *Proc. Int. Conf. Internet Sci.* Cham, Switzerland: Springer, 2018, pp. 242–256.

[202] A. O. Larsson, "Diversifying likes: Relating reactions to commenting and sharing on newspaper Facebook pages," *Journalism Pract.*, vol. 12, no. 3, pp. 326–343, Mar. 2018.

[203] A. Chatterjee, K. N. Narahari, M. Joshi, and P. Agrawal, "SemEval-2019 task 3: EmoContext contextual emotion detection in text," in *Proc. 13th Int. Workshop Semantic Eval.* Minneapolis, MN, USA: Association for Computational Linguistics, Jun. 2019, pp. 39–48. [Online]. Available: https://www.aclweb.org/anthology/S19-2005

[204] V. Dudina and D. Judina, "Mining opinions on the Internet: Can the text analysis methods replace public opinion polls?" *Monitor. Public Opinion, Econ. Social Changes*, vol. 5, pp. 63–78, Nov. 2017.

[205] V. Dudina, "Digital data potentialities for development of sociological knowledge," *Sociol. Stud.*, vol. 9, pp. 21–30, Oct. 2016.

[206] P. Panicheva, A. Mirzagitova, and Y. Ledovaya, "Semantic feature aggregation for gender identification in Russian facebook," in *Artificial Intelligence and Natural Language*, A. Filchenkov, L. Pivovarova, and J. Žižka, Eds. Cham, Switzerland: Springer, 2018, pp. 3–15.

[207] R. Bhargava, G. Goel, A. Shah, and Y. Sharma, "Gender identification in Russian texts," in *Proc. FIRE Working Notes*, 2017, pp. 13–16.

[208] T. Litvinova, F. M. R. Pardo, P. Rosso, P. Seredin, and O. Litvinova, "Overview of the rusprofiling pan at fire track on cross-genre gender identification in Russian," in *Proc. FIRE Working Notes*, 2017, pp. 1–7.

[209] A. Sboev, I. Moloshnikov, D. Gudovskikh, and R. Rybka, "A comparison of data driven models of solving the task of gender identification of author in Russian language texts for cases without and with the gender deception," *J. Phys., Conf. Ser.*, vol. 937, Dec. 2017, Art. no. 012046.

[210] A. Sboev, I. Moloshnikov, D. Gudovskikh, A. Selivanov, R. Rybka, and T. Litvinova, "Automatic gender identification of author of Russian text by machine learning and neural net algorithms in case of gender deception," *Procedia Comput. Sci.*, vol. 123, pp. 417–423, Jan. 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050918300656

[211] A. Sboev, D. Gudovskikh, I. Moloshnikov, and R. Rybka, "A gender identification of text author in mixture of Russian multi-genre texts with distortions on base of data-driven approach using machine learning models," *AIP Conf.*, vol. 2116, no. 1, 2019, Art. no. 270006.

[212] A. Mislove. (2010). *Pulse of the Nation: Us Mood Throughout the Day Inferred From Twitter*. [Online]. Available: http://www.ccs.neu.edu/home/amislove/twittermood/

[213] D. N. Prata, K. P. Soares, M. A. Silva, D. Q. Trevisan, and P. Letouze, "Social data analysis of Brazilian's mood from Twitter," *Int. J. Social Sci. Humanity*, vol. 6, no. 3, pp. 179–183, Mar. 2016.

[214] F. Dzogang, S. Lightman, and N. Cristianini, "Circadian mood variations in Twitter content," *Brain Neurosci. Adv.*, vol. 1, pp. 1–14, Dec. 2017, doi: 10.1177/2398212817744501.

[215] I. Bessmertny and R. Posevkin, "Texts sentiment-analysis application for public opinion assessment," *Nauchno-Tekhnicheskii Vestnik Informatsionnykh Tekhnologii, Mekhaniki i Optiki*, vol. 15, no. 1, pp. 169–171, 2015.

[216] V. Averchenkov, D. Budylskii, A. Podvesovskii, A. Averchenkov, M. Rytov, and A. Yakimov, "Hierarchical deep learning: A promising technique for opinion monitoring and sentiment analysis in Russian-language social networks," in *Creativity in Intelligent Technologies and Data Science*. Cham, Switzerland: Springer, 2015, pp. 583–592.

[217] V. Sydorenko, S. Kravchenko, Y. Rychok, and K. Zeman, "Method of classification of tonal estimations time series in problems of intellectual analysis of text content," *Transp. Res. Procedia*, vol. 44, pp. 102–109, Jan. 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S235214652030065X

SERGEY SMETANIN received the B.S. degree in software engineering and the M.S. degree in business informatics from the National Research University Higher School of Economics, Moscow, Russia, in 2016 and 2018, respectively. His research interests include computational linguistics, sentiment analysis, and mobile applications development.

• • •