PolSentiLex: Sentiment Detection in Socio-political Discussions on Russian Social Media*

Olessia Koltsova[1][0000−0002−2669−3154], Svetlana Alexeeva[2][0000−0002−8540−6178], Sergei Pashakhin[1][0000−0003−0361−2064], and Sergei Koltsov[1][0000−0002−2932−2746]

[1] Laboratory for Social & Cognitive Informatics,
National Research University Higher School of Economics,
room 216, 55/2 Sedova Str., 190008, St.Petersburg, Russia
{ekoltsova,spashahin,skoltsov}@hse.ru
[2] Laboratory for Cognitive Studies,
St. Petersburg State University
mail@s-alexeeva.ru

Abstract. We present a freely available Russian language sentiment lexicon PolSentiLex designed to detect sentiment in user-generated content related to social and political issues. The lexicon was generated from a database of posts and comments of the top 2,000 LiveJournal bloggers posted during one year (~1.5 million posts and 20 million comments). Following a topic modeling approach, we extracted 85,898 documents that were used to retrieve domain-specific terms. This term list was then merged with several external sources. Together, they formed a lexicon (16,399 units) marked-up using a crowdsourcing strategy. A sample of Russian native speakers (n = 105) was asked to assess words' sentiment given the context of their use (randomly paired) as well as the prevailing sentiment of the respective texts. In total, we received 59,208 complete annotations for both texts and words. Several versions of the marked-up lexicon were experimented with, and the final version was tested for quality against the only other freely available Russian language lexicon and against three machine learning algorithms. All experiments were run on two different collections. They have shown that, in terms of F_{macro}, lexicon-based approaches outperform machine learning by 11%, and our lexicon outperforms the alternative one by 11% on the first collection, and by 7% on the negative scale of the second collection while showing similar quality on the positive scale and being three times smaller. Our lexicon also outperforms or is similar to the best existing sentiment analysis results for other types of Russian-language texts.

Keywords: social media · socio-political domain · sentiment analysis · Russian language · lexicon-based approach

* This work is an output of a research project implemented as part of the Basic Research Program at the National Research University Higher School of Economics (HSE University).

1 Introduction

Tools and resources for sentiment analysis (SA) in the Russian language are often created with a focus on consumer reviews and remain mostly underdeveloped for other types of communication taking place on social media. Increasingly crucial in public life, social media have become valuable for social scientists studying sentiment in online political discussions or interested in predicting public reaction to events with online data. However, such studies face a lack of SA resources, as they are usually language- and domain-specific, or demand expertise in machine learning and feature engineering from their users. Additionally, for the Russian language, the ever obtained quality in various SA tasks is modest even when the most advanced machine learning methods are applied. Our work seeks to overcome these obstacles.

In this article, we make the following contributions:

- We propose *PolSentiLex*, a freely available (https://linis-crowd.org/) and easy to use lexicon for sentiment analysis of social media texts on social and political issues in the Russian language;
- We employ several heuristics, including topic modeling, that allow obtaining the maximum number of domain-specific sentiment words;
- The proposed lexicon—*PolSentiLex*—outperforms the only other freely available Russian-language lexicon that was released after the start of our project as a general-purpose lexicon;
- We demonstrate that, for our case, lexicon-based approaches significantly outperform machine learning, especially on the collection that was not used for lexicon generation.

The paper is organized as follows. The next section presents a review of the literature on sentiment analysis in the Russian language, social media, and socio-political domain. In sections 3 and 4, we describe the process of PolSentiLex development and the procedure of its quality assessment, respectively. Section 5 contains the results. We conclude with suggestions for sentiment analysis of socio-political messages from Russian language social media in section 6.

2 Related Work

Lay texts of the socio-political domain present larger difficulties for sentiment classification than consumer product reviews, which is explained by a more subtle expression of sentiment in the former, including the presence of sarcasm and ironies [33]. Usually, socio-political issues are described with larger vocabularies than consumer reviews, and contain a complex mixture of factual reporting and subjective opinions [19]. Moreover, as pointed out by Eisenstein [11], the language on social media where such texts are created significantly differs from the "norm," containing more non-standard spelling and punctuation, vocabulary, and syntax. Additionally, language use varies both within and across social media platforms. Following in line with Androutsopoulos [1] and Darling et al.

[10] Eisenstein argues that, for instance, Twitter language varies a lot in "the degree of standardness." Such domain- and medium-specific features have to be accounted for when applying both lexicon or machine learning (ML) approaches to SA [26,23].

2.1 Sentiment Analysis in the Russian langauge

The state of sentiment analysis (SA) in the Russian language is reviewed in [30], and the review of existing Russian-language lexicons is available in Kotelnikov et al [17]. The latter work tests the performance of 10 versions of 8 lexicons on the task of consumer reviews using SVM algorithm. It is indicative that not only the task is related to consumer reviews, but also 6 out of 8 lexicons, including the authors' lexicon, were created specifically for review classification [7,2,24,5,32,16]. The two that were not, are RuSentiLex [8] and the early version of PolSentiLex described in the paper by the title of the website where it was published (Linis-Crowd). All lexicons lose the race to the full dictionary of the corpus, the second place being taken by the united lexicon, and the third by ProductSentiRus [7].

Similar focus on reviews can be seen in sentiment analysis tracks hosted by Russian Information Retrieval Seminar (ROMIP) in 2012–2016. Apart from consumer reviews, these tracks have also included tasks on social media texts (blogs and Twitter) and news items [6,9,20,22]. However, except for the latter, during all events, participants have been offered to detect sentiment in opinions about services, products, and organizations. As expected, most winners follow machine learning approaches, with SVM on average being the most accurate algorithm. Nonetheless, most of those solutions rely on manually created dictionaries of sentiment-bearing words or features engineered based on such resources. Lexicon-based approaches, on the other hand, show worst and rarely comparable results with only one exception (see below).

ROMIP participants and organizers note that the performance of solutions is highly dependent on the number of sentiment classes, data sources, and task domains. Blogs and tweets as well as texts from the socio-political domain in general are considered the most difficult. While general social media texts, compared to consumer reviews, are reported to be more diverse in sentiment expression, less structured and less grammaticaly sound, the difficulty of socio-political texts is attributed to the greater variety of subtopics, compared to other domains. Interestingly, the only occasion where a lexicon-based approach outperformed machine learning was during the task of detecting sentiment in news items [18]. The winner solution used an extensive dictionary of opinion words and expressions obtained manually as well as with text mining techniques. The system estimated sentiment following several rules regarding sentiment intensifiers, negation, and following opinion words. Although significantly better than the baseline, this solution showed $F_{macro} = 0.62$. All this points at the need for the development of SA resources for Russian-language general interest social media texts and for socio-political texts, including professional news and lay messages.

This lack of generalist SA resources for the Russian language was addressed by Loukachevitch and Levchik [21] to create a general sentiment lexicon named *RuSentiLex*. It is the successor of the very first publicly available Russian sentiment lexicon, which had no polarity scores and was developed by the same researchers [8]. RuSentiLex was constructed in a semi-automatic way from several sources: (1) words from domain-specific lexicons matched with Russian language thesaurus; (2) words and phrases extracted following special rules from a large corpus of news articles; (3) slang and curse words extracted from Twitter with a supervised model of sentiment word extraction. The lexicon consists of 16,057 n-grams, with 63.7% of them being negative, 23.6% – positive, 10.9% – neutral, and with 1.8% having context-depended sentiment polarity.

To the best of our knowledge, the only work that tested the quality of RuSentiLex is the one by Kotelnikov et al. [17]; who, as it was noted, focused on consumer reviews. Also, some ROMIP participants successfully used RuSenitLex for feature engineering which allowed them to beat either the baseline or some other competitors [21]. As RuSentiLex is, to the best of our knowledge, the only freely available Russian-language lexicon not aimed at review classification, in this work we use it as our baseline. We test both RuSentiLex and PolSentiLex using a rule-based algorithm designed for them and further compare them to several algorithms of machine learning on two different datasets.

3 PolSentiLex

In this section, we briefly review the process of PolSentiLex construction. It closely follows the procedure adopted for the early version of our lexicon [14] that had produced only a very modest quality; all differences of this version are accounted for further below.

3.1 LiveJournal collection of social and political posts

Given our primary goal to develop a lexicon for SA of messages from politicized social media, we began with extracting domain-specific words. We constructed two text collections that contain posts and comments from top 2,000 accounts in the Russian language section of LiveJournal for the period of one year (March 2013–March 2014). At the time of data collection, LiveJournal was still an active and a highly politicized blog platform in Russia. However, our previous research based on this data indicates that only about one-third of these documents can be classified as political or social, which posed a problem of retrieving relevant texts out of approximately 1.5 million posts and 0.9 million comments.

To solve this problem, we performed Latent Dirichlet Allocation (LDA) topic modeling with Gibbs sampling [12] on the collections of posts and comments separately. The early version of our lexicon [14] used only posts as presumably more informative in terms of topic and sentiment vocabulary. In this work we introduce the merged lexicon which also uses the collection of comments that were added due to the obviously insufficient quality of the first version. To overcome

poor performance of topic modeling on short texts, all comments related to the same post were merged. Based on experiments performed with similar data and described in [4,25], we modeled both collections with 300 topics.

Next, each topic was manually labelled by three annotators from our team based on reading of a maximum of 200 top terms and 100 top texts (ranked by their probability in a topic). They identified 104 and 88 topics from social or political domains in each of the two collections of posts and comments, respectively. Additionally, nine and 20 topics composed mostly from general emotional words were identified in the collections of posts and comments, respectively. We considered a topic to be relevant to our chosen domains if at least two annotators agreed on that.

Finally, from each of the relevant topics we retrieved the most relevant texts based on the values from the topic-document matrix Φ. For post collection, the threshold of relevance was set to (>0.1) which produced a subsample of 70,710 blog posts, and for the comment collection it had to be set lower (>0.001), which yielded a smaller subsample of 15,188 merged comment texts.

3.2 Selection of potentially sentiment-bearing words

We created the core of our proto-lexicon using the list of the top 200 words from all social and political topics. Then, we extended our list of potentially sentiment-bearing terms with the words from several external sources described in detail in [14].

Next, we intersected the listed sources and retrieved only the terms that occurred in at least two of them. The dictionary that resulted from the work with post collection contained 9,539 units, and the one resulting from the comment collection consisted of 6,860 units.

3.3 Data mark up

One of our main ideas in the lexicon construction was that words, even those retrieved from general dictionaries, might have specific sentiment polarity or strength when occurring in social or political texts. Therefore, we chose to mark up the selected words providing the annotators with the context in which the words occurred; simultaneously, it allowed us to combine word mark up with text mark up.

As both negative and positive sentiment in texts can produce a social effect so far as it is perceived as such by society members, polarity scores for our lexicon and document collections were obtained from lay native speakers. Therefore, our assessors were not supposed to imitate experts; instead, we defined their contribution similar to that of respondents in an opinion poll: no “wrong” answers were possible. In total, 87 people took part in the assessment of posts, and 18 individuals from our team participated in the mark up of comments. Volunteers for post annotation were recruited via social media. All of them participated in instruction sessions and training annotation, where all of them were offered the same texts, after which some coders were discarded.

A special website (https://linis-crowd.org/) was designed for the mark-up which asked participants to assess words' sentiment as expressed in the texts where they occurred, as well as the prevailing sentiment of texts themselves, with a five-point scale, from −2 (strongly negative) to +2 (strongly positive). For each word, the system randomly selected three different texts relevant to politics or public affairs. Since some texts were not unique for each word, the texts, too, received multiple scores.

Each word was coded three times but not necessary in three different texts; some words from our proto-dictionary did not occur in our collections and were excluded. Also, since a word-text pair was randomly selected, several pairs were coded more than three times. As a result, at our first stage we received 32,437 annotation points both for posts and words: 7,546 unique words were annotated three times each, and 19,831 unique posts received one or more marks. At our second stage we repeated the entire procedure on the comment collection and obtained 26,851 annotation points for both merged comment texts and words, with 6,860 unique words receiving three marks each and with 15,188 unique comment texts received at least one mark. Intercoder agreement calculated among three random grades for each of the words is 0.553 in terms of Krippendorf's alpha. In the resulting lexicon, all grades of each word were averaged and rounded.

3.4 The Three Versions of PolSentiLex

In the course of all the experiments we tested three versions of our lexicon. The first version (further, *post version*) included 2,793 non-neutral words derived from the collection of social and political blog posts. The rest of 7,546 annotated words were found to carry no sentiment. This lexicon produced the quality of 0.44 in terms of F_{macro}, reported in [14] which is why further experiments were carried out. The next version (further, *comment version*) included 2,664 non-neutral words derived from the collection of merged comment texts. The experiments with it (not reported) produced no increase in quality. Eventually, we combined both lexicons into the final version (further, *combined version*); since some words occurred in both the post and the comment versions, their scores were averaged.

Table 1 shows the distributions of scores over words in the three versions of PolSentiLex. Obviously, although negatively assessed words prevail, positive words are also present. At the same time, very few highly emotional words are observed.

4 PolSentiLex quality assessment

In this section, we describe experiments in which we evaluate the quality of PolSentiLex against RuSentiLex [21]. These two lexicons are used as feature sets in three machine learning algorithms and in a dictionary-based technique, and tested on two datasets.

Table 1. Distribution of mean scores over words in the post, comment and combined versions of PolSentiLex

Mean score (rounded)	N words	Share of words, %
Post lexicon		
−2	225	3
−1	1,666	22
0	4,753	63.4
1	853	11
2	49	0.6
In total	7,546	100
Not neutral (in total)	2,793	37
Comment lexicon		
−2	173	2.5
−1	1,882	27.3
0	4,196	61
1	596	9
2	13	0.2
In total	6,860	100
Not neutral (in total)	2,664	39
Combined lexicon		
−2	252	2
−1	2,612	24
0	6,738	63.8
1	1,031	10
2	29	0.2
In total	10,662	100
Not neutral (in total)	3,924	37

4.1 Datasets

Both corpora used for quality assessment of our lexicon are comprised of socio-political texts from social media: one is a subsample of posts used to create our lexicon (see sections 3.1 and 3.3, further *LiveJournal posts*) and the other is an independent corpus (further *Ethnicity collection*).

The Ethnicity collection was sampled from all possible Russian social media and blogs for the period from January 2014 to December 2015 that contained texts with an ethnonym. Ethnonyms—linguistic constructions used to nominate ethnic groups—were derived from a predefined list of 4,063 words and bigrams covering 115 ethnic groups living in Russia. Based on this list, the sample was provided by a commercial company that collects the Russian language social media content. This collection was chosen for this project out of all the Russian-language collections known to us as the most relevant to the task of testing a socio-political lexicon (discussions of ethnicity are usually not only highly politicized, but also very heated). Furthermore, it was not used in the development of either PolSentiLex or RuSentiLex. The well-known RuSentiment dataset [29] that was made available a little later is not focused on political texts.

From our Ethnicity collection, 14,998 messages were selected so as to represent all ethnic groups and assessed by three independent coders each; of them 12,256 were left after filtering near-duplicates. The coders were asked to answer a number of questions about the texts, including two most important for this study: (a) how strongly a general negative sentiment is expressed in the text, if any? (no/weak/strong); (b) how strongly a general positive sentiment is expressed in the text, if any? (no/weak/strong). In this mark-up, we used two independent scales for positive and negative sentiment instead of the integral sentiment scale used for LiveJournal collection (see sections 3.1 and 3.3) as it corresponded better to the purpose for which Ethnicity collection was constructed.

While LJ collection was marked-up in parallel with word mark-up, as explained in 3.3, with the same set of annotators, marking-up of Ethnicity collection was a separate task, but it followed a similar procedure, with 27 student volunteers being specially trained for that. Intercoder agreement, as expressed with Krippendorf's alpha is, on LJ collection: 0.541 for a five-class task, on Ethnicity collection: 0.547 on the negative scale, and 0.404 on the positive scale, both being three-class tasks. This level of agreement is quite common in sentiment analysis [3]. Texts that received fewer than two marks were excluded.

Grades for the Ethnicity collection could vary from -2 to 0 for the negative scale and from 0 to $+2$ for the positive scale, and all these categories turned out to be reasonably populated. However, the LiveJournal collection which had been marked-up on a unified positive-negative scale, turned out to have very few texts with the extreme values $+2$ or -2 (about 6%). Therefore, we collapsed the five-point (-2, -1, 0, 1, 2) scale into a three-point scale (-1, 0, 1) where $-2 = -1$ and $2 = 1$. As a result, we formed three three-class classification tasks that thus became easier to compare. Final polarity scores for all texts were calculated as the mean values of individual human grades.

Table 2 shows the distribution of scores over texts. Most texts are marked as neutral or negative, with fewer positive marks. For the LiveJournal collection, the positive to negative class proportion is 1:5.8, and for the Ethnicity collection it is 1:3.2. The same unbalanced class structure in political blogs is also pointed at by Hsueh, Melville, and Sindhwani [13].

Table 2. Distribution of mean scores over text in the LiveJournal posts and Ethnicity collection

Mean score (rounded)	**N texts**	**Share of texts, %**
LiveJournal posts, integral score		
−1	2,104	33
0	3,940	61
1	360	6
In total	6,404	100
Not neutral (in total)	2,464	38
Ethnicity collection		
Negative scale		
−2	1,126	9
−1	4,181	33.2
0	7,272	57.8
In total	12,579	100
Not neutral (in total)	5,307	42
Positive scale		
0	10,882	86.6
1	1,436	11.4
2	261	2
In total	12,579	100
Not neutral (in total)	1,652	13

Before testing, both LiveJournal and Ethnicity collections were preprocessed: for the ML approach, we cleaned each collection from non-letter symbols and lemmatized each word with Pymorphy2 [15]; and for the lexicon approach, we used lemmatized documents with punctuation intact.

We performed multiple comparisons of PolSentiLex and RuSentiLex used as feature sets in one rule-based approach and in three ML algorithms. Based on preliminary experiments, we chose the version of our lexicon that performed better or not worse than the other versions. Predictably, it turned out to be the combined version (see 3.4) of PolSentiLex that comprised of 3,924 terms. As for RuSentiLex, we used all not-neutral context-independent unigrams that counted to 11,756 units in total because unigrams were reported to be the most useful features for sentiment classification [27].

In total, we performed 24 tests: three tasks (negative and positive sentiment prediction for the Ethnicity collection and overall polarity prediction for the LiveJournal collection) were performed with the two lexicons, each of which

was used with four approaches (thee ML algorithms and one rule-based). The total number of runs, including all parameter optimization and cross-validation iterations, was 390.

For the ML approach, we chose the three most popular algorithms for SA, namely support vector machine (SVM) with linear basis function kernel, Gaussian Naïve Bayes (NB), and k-nearest neighbors (KNN) classifier. For training, we used a document-term matrix with the presence of a term from a lexicon in documents. We used a random 75% data sample for training and validation, and the rest 25% for testing (held-out data). First, following the grid-search strategy with 2-fold cross-validation, we identified the best parameters for SVM and KNN on training data. For SVM, we explored hyper-parameter C in the range [0.0001, 0.001, 0.01, 0.1, 1, 100, 1000, 10000, 100000, 1000000] and identified that the algorithm performed best with C = 0.0001. For KNN, we varied the number of neighbors from 1 to 40, and in almost all tests, the best performance was achieved with k=1. The two exceptions were the Ethnicity collection with PolSentiLex as a feature set on a positive scale (k=3) and the Ethnicity collection with PolSentiLex as a feature set on a negative scale (k=4). Then, using the obtained parameters, we trained each classifier with 10-fold cross-validation. Finally, to obtain an unbiased evaluation, we applied a classifier with the highest F_{macro} on validation data to holdout datasets. To train classifiers, we used the Scikit-learn Python library [28].

For lexicon approach, we used SentiStrength rule-based algorithm [31]. We chose SentiStrength because its implementation is freely available, and it was designed specifically for the social web texts. To classify a document, SentiStrength, firstly, searches the text for and scores terms from a sentiment dictionary defined by a user, correcting their scores for the presence of booster and negation words. It then applies one of several approaches to estimate sentence- and text-level sentiment on positive and negative scales separately. Based on the preliminary experiments, we chose the approach that showed the best results where the sentiment of a sentence equals to the sentiment of its strongest term, and the sentiment of a text equals the strongest sentence sentiment.

To accurately assess the quality of SentiStrength prediction, we had to transform its text sentiment scores so that they become comparable to the classes from the human mark-up. Because of booster words, SentiStrength sentiment score for a text could go beyond +/−2. Therefore, to align the scales of predicted sentiment scores and the true assessors' scores for Ethnicity collection (0,+1,+2 & 0, −1,−2), we considered all texts with the SentiStrength score above +2 to belong to the class "+2" (highly positive), and all texts with the SentiStrength scores below −2 to belong to class "−2" (highly negative).

As LiveJournal collection texts had been marked up using a single sentiment scale (from −2 to 2), we applied the following steps to transform two separate SentiStrength scores into a single score and to compare them to the respective human scores. (1) We calculated the mean of the two SentiStrength scores (positive and negative) and thus obtained the integral predicted score for each text, PS. (2) We calculated the difference between PS and the true score, TS, taken

as the non-rounded assessors' sentiment score for the same text. (3) As both TS, PS and their difference were likely to be non-integer, to determine whether the true classes were correctly predicted, we used the following logical expressions: (3a) If $|PS-TS| < 0.5$, then $PS = TS$, i.e. the true class is correctly predicted. (3b) If $0.5 \leqslant |PS-TS| < 1.5$ then $|PS-TS| = 1$, i.e, the classification error is $+/-1$ class. (3c) If $|PS-TS| \geqslant 1.5$ then $|PS-TS| = 2$, i.e. the classification error is $+/-2$ classes.

To evaluate the performance of all our sentiment analysis approaches, we used standard metrics for classifier performance: the F_{macro} measure (reported in Fig. 1), precision (reported in Fig. 2), recall (reported in Fig. 3) and accuracy (reported in Fig. 4).

5 Results

The most important results are presented in figures 1–4 and table 3. First, our best solutions (KNN with PolSentiLex for accuracy on the positive scale of Ethnicity collection and SentiStrength with PolSentiLex for all other tasks and quality metrics) significantly exceed the random baseline, accounting for the class imbalance. Thus, in terms of accuracy the gain over the random baseline is 14–51% which is very good for the Russian language. For instance, similar ROMIP tasks on three-class sentiment classification of news and on consumer blog posts demonstrated the gain of 5–49% on average [14]. Moreover, our best solution—PolSentiLex with SentiStrength—has also improved a lot as compared to our previous result [14]. On the LiveJournal collection, which was used for testing our lexicon last time, the improvement has been 14% and 13% in terms of precision and recall, respectively. Comparing performance of our new lexicon on the Ethnicity collection, from which it was not derived, and the performance of our old lexicon on LiveJournal collection, from which we did derive it, the new lexicon is still 3–12% better in both precision and recall on the negative scale and in recall on the positive scale, although it is slightly worse in precision on the negative scale.

Table 3. Advantage of PolSentiLex over RuSentiLex using SentiStrength

	F_{macro}	Precision	Recall	Accuracy
LiveJournal	11%	9%	13%	10%
Ethnicity - negative	8%	4%	11%	0%
Ethnicity - positive	0%	0%	0%	7%

Second, interestingly, rule-based approaches with any lexicons are visibly better than any ML approaches on all datasets and across all metrics, with one exception addressed further below. Thus, in terms of Fmacro measure, the lexicon approaches perform, on average, 11% better than the ML approaches,

which is a huge difference. It might be attributed to a non-optimal parameter choice in our ML solutions, however, the two best ML approaches (KNN and SVM) produce the gain over baseline comparable to that in ROMIP tracks [14].

A look at the exception—namely, the KNN method with PolSentiLex for the positive scale in terms of general accuracy—reveals a curious result: KNN solution does not exceed lexicon-based solutions either in precision or recall. A closer examination of the relevant confusion matrices shows that KNN ascribes almost all texts to the neutral class producing exceptionally low precision and recall for the classes +1 and +2, i.e., it often fails to detect positive sentiment. However, as the non-positive class in the Ethnicity collection is by far larger than the other two and constitutes 86%, a fair ability of KNN to detect this class contributes a lot to the overall accuracy (84%). This result is suboptimal for social scientists who usually aim to detect texts containing non-neutral emotions, which is what rule-based approaches perform better, albeit at the expense of neutral class detection. We can assume that lexicon-based approaches might be better for social science tasks including ours, which is consistent with Thelwall's conclusions [31] and with the ROMIP results reported in section 2. This means that, unlike consumer reviews, politicized social media texts are more diverse, less structured and are harder to divide into classes, which might make manually selected words more reliable class indicators than features engineered with ML.

Finally, the most important comparison is that between the two tested lexicons (see table 3). The fact that our lexicon outperforms RuSentiLex on the LiveJournal collection is predictable since this collection was used as a source for our lexicon, but not for its competitor. A more interesting observation is that on the two other tasks PolSentiLex is also either better or not worse than RuSentiLex in terms of all aggregated metrics. This deserves mentioning given that our lexicon is only 33% the size of RuSentiLex (3,924 words against 11,756). However, to give a fair treatment to RuSentiLex, we should look into respective confusion matrices and the distribution of quality metric values over classes.

On the positive scale, PolSentiLex, on average, has no advantage over RuSentiLex in precision, while its recall is higher for the neutral class and lower for class (+1). The overall recall of PolSentiLex on both sentiment-bearing classes is about 5% lower than that of RuSentiLex, even though the former has a slight advantage over the latter in class (+2). Since, as it has been mentioned, prediction of sentiment-bearing classes is a priority for social science tasks, PolSentiLex cannot be considered clearly better in predicting positive sentiment than RuSentiLex, despite the better overall accuracy of the former.

On the negative scale, with PolSentiLex having some advantage in precision, it yields a visibly smaller recall for class (−1). However, here the main confusion of PolSentiLex is not that severe, being between classes (−1) and (−2). It means that PolSentiLex is not prone for losing negative texts; instead, it tends to overestimate negative sentiment by classifying some moderately negative texts as highly negative. At the same time, PolSentiLex is much better in both precision and recall on class (−2), and its overall accuracy on the two sentiment-bearing classes is marginally higher than that of RuSentiLex. We can conclude that the

two lexicons are similar in quality, especially in precision, and that there is a trade-off between overall accuracy and the ability to detect sentiment-bearing classes.

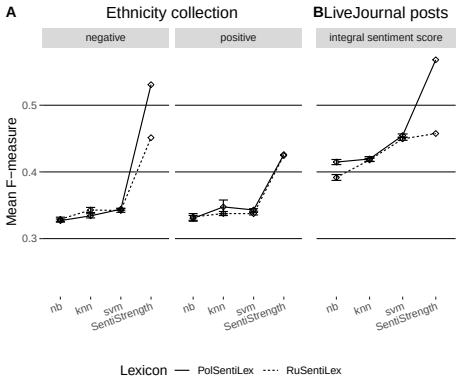

Fig. 1. F_{macro} results: a) the Ethnicity collection, b) the collection of LiveJournal posts.

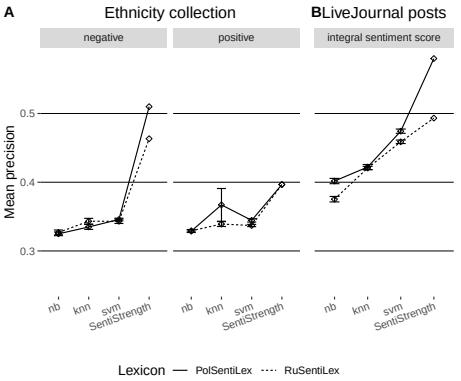

Fig. 2. Precision results: a) the Ethnicity collection, b) the collection of LiveJournal posts.

6 Conclusion

The results suggest that in sentiment analysis of socio-political messages from Russian-language social media, given the available resources, a social scientist will be better off with using a rule-based method, such as provided by SentiStrength package, with either PolSentiLex or RuSentiLex lexicons. With any of them, a user will not only get a visibly higher quality, but also lower computational complexity and a much more user-friendly and intuitive method. While PolSentiLex shows lower recall for moderate classes (moderately positive and moderately negative texts), it is either better or not worse than RuSentilex in detection of all other classes, according to all metrics, including those aggregated over all classes. Since PolSentiLex is also much smaller than RuSentiLex, it might be considered an optimal choice for the time being, although further improvements are needed.

One of the directions for improvement is to merge the lexicons while giving priority to PolSentiLex mark-up and re-evaluating the polarity of the remaining RuSentiLex terms in a socio-political context. Another improvement might be gained by adding bigrams typical for socio-political texts, starting with those that contain sentiment words. Next, text mark-up may also be used as a source for lexicon enrichment: thus, assessors may be asked to mark text fragments that were most helpful to form their opinion on the text. Finally, both lexicon

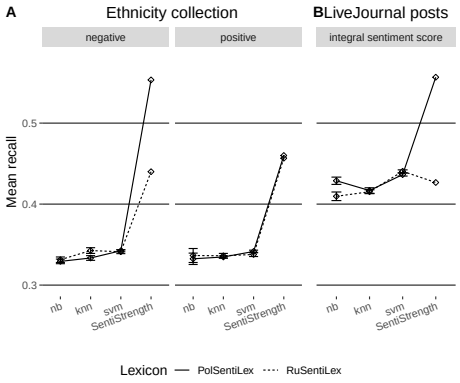

Fig. 3. Recall results: a) the Ethnicity collection, b) the collection of LiveJournal posts.

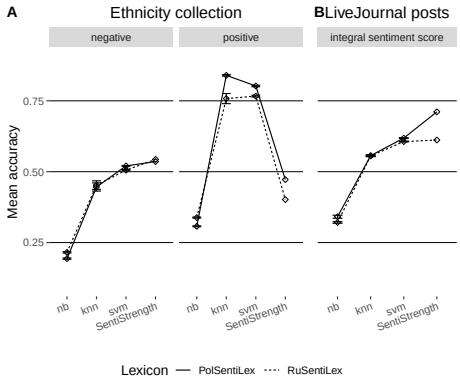

Fig. 4. Accuracy results: a) the Ethnicity collection, b) the collection of LiveJournal posts.

might be used as features in more advanced machine learning algorithms such as neural networks, along with distributed word representations.

References

1. Androutsopoulos, J.: Language change and digital media: a review of conceptions and evidence. In: Standard Languages and Language Standards in a Changing Europe, pp. 145–160. Novus, Oslo (2011)
2. Blinov, P.D., Klekovkina, M.V., Kotelnikov, E.V., Pestov, O.A.: Research of lexical approach and machine learning methods for sentiment analysis. In: Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue-2013", vol. 2, pp. 51–61. RGGU, Moscow (2013), http://www.dialog-21.ru/media/1226/blinovpd.pdf
3. Bobicev, V., Sokolova, M.: Inter-annotator agreement in sentiment analysis: Machine learning perspective. In: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017. pp. 97–102. INCOMA Ltd., Varna, Bulgaria (Sep 2017). https://doi.org/10.26615/978-954-452-049-6$_0$15
4. Bodrunova, S., Koltsov, S., Koltsova, O., Nikolenko, S., Shimorina, A.: Interval semi-supervised lda: Classifying needles in a haystack. In: Mexican International Conference on Artificial Intelligence. pp. 265–274. Springer (2013)
5. Chen, Y., Skiena, S.: Building Sentiment Lexicons for All Major Languages. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 383–389. Association for Computational Linguistics, Baltimore, Maryland (2014). https://doi.org/10.3115/v1/P14-2063, http://aclweb.org/anthology/P14-2063
6. Chetviorkin, I., Braslavski, P., Loukachevitch, N.: Sentiment Analysis Track at ROMIP 2011. In: Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2012). vol. 2, pp. 1–14 (2012)

7. Chetviorkin, I., Loukachevitch, N.: Extraction of Russian Sentiment Lexicon for Product Meta-Domain. In: Proceedings of COLING 2012: Technical Papers, pp. 593–610. The COLING 2012 Organizing Committee, Mumbai (2012), https://www.aclweb.org/anthology/C12-1037
8. Chetviorkin, I., Loukachevitch, N.: Extraction of Russian Sentiment Lexicon for Product Meta-Domain. In: Proceedings of COLING 2012: Technical Papers. pp. 593–610. Mumbai (Dec 2012)
9. Chetviorkin, I., Loukachevitch, N.: Sentiment Analysis Track at ROMIP 2012. In: Computational Linguistics and Intellectual Technologies (2013), http://www.dialog-21.ru/digests/dialog2013/materials/pdf/1_ChetverkinII.pdf
10. Darling, W., Paul, M., Song, F.: Unsupervised Part-of-Speech Tagging in Noisy and Esoteric Domains with a Syntactic-Semantic Bayesian HMM. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Avignon, France (2012)
11. Eisenstein, J.: What to do about bad language on the internet. In: Proceedings of the 2013 conference of the North American Chapter of the association for computational linguistics: Human language technologies. pp. 359–369 (2013)
12. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National academy of Sciences **101**(suppl 1), 5228–5235 (2004)
13. Hsueh, P.Y., Melville, P., Sindhwani, V.: Data quality from crowdsourcing: a study of annotation selection criteria. In: Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing. pp. 27–35. Association for Computational Linguistics (2009)
14. Koltsova, O., Alexeeva, S., Koltsov, S.: An Opinion Word Lexicon and a Training Dataset for Russian Sentiment Analysis of Social Media. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016". pp. 277–287. RSUH, Moscow (2016)
15. Korobov, M.: Morphological analyzer and generator for russian and ukrainian languages. In: International Conference on Analysis of Images, Social Networks and Texts. pp. 320–332. Springer (2015)
16. Kotelnikov, E., Bushmeleva, N., Razova, E., Peskisheva, T., Pletneva, M.: Manually Created Sentiment Lexicons: Research and Development. In: Computational Linguistics and Intellec- tual Technologies: Papers from the Annual International Conference "Dialogue-2016", vol. 15, pp. 300–314. RGGU, Moscow (2016), http://www.dialog-21.ru/media/3402/kotelnikovevetal.pdf
17. Kotelnikov, E., Peskisheva, T., Kotelnikova, A., Razova, E.: A Comparative Study of Publicly Available Russian Sentiment Lexicons. In: Ustalov, D., Filchenkov, A., Pivovarova, L., Žižka, J. (eds.) Artificial Intelligence and Natural Language, vol. 930, pp. 139–151. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-030-01204-5$_1$4, http://link.springer.com/10.1007/978-3-030-01204-5_14, series Title: Communications in Computer and Information Science
18. Kuznetsova, E., Loukachevitch, N., Chetviorkin, I.: Testing Rules for a Sentiment Analysis System. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2013". vol. 2, pp. 71–80 (2013), http://www.dialog-21.ru/digests/dialog2013/materials/pdf/KuznetsovaES.pdf
19. Liu, B.: Sentiment analysis and opinion mining. Morgan & Claypool Publishers (2012)
20. Loukachevitch, N., Blinov, P., Kotelnikov, E., Rubtsova, Y., Ivanov, V., Tutubalina, E.: SentiRuEval: Testing Object-Oriented Sentiment Analysis Sys-

tems in Russian. In: Computational Linguistics and Intellectual Technologies. p. 13 (2015), http://www.dialog-21.ru/digests/dialog2015/materials/pdf/LoukachevitchNVetal.pdf
21. Loukachevitch, N., Levchik, A.: Creating a General Russian Sentiment Lexicon. In: Proceedings of Language Resources and Evaluation Conference LREC-2016. pp. 1171–6 (2016)
22. Loukachevitch, N., Rubcova, Y.: SentiRuEval-2016: overcoming the time differences and sparsity of data for the reputation analysis problem on Twitter messages [SentiRuEval-2016: preodoleniye vremennykh razlichiy i razrezhennosti dannykh dlya zadachi analiza reputatsii po soobshcheniyam tvittera]. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016". pp. 416–426 (2015)
23. Medhat, W., Hassan, A., Korashy, H.: Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal **5**(4), 1093–1113 (Dec 2014). https://doi.org/10.1016/j.asej.2014.04.011, http://linkinghub.elsevier.com/retrieve/pii/S2090447914000550
24. Mohammad, S.M., Turney, P.D.: Crowdsourcing a word-emotion association lexicon. Computational Intelligence **29**(3), 436–465 (Aug 2013). https://doi.org/10.1111/j.1467-8640.2012.00460.x, http://doi.wiley.com/10.1111/j.1467-8640.2012.00460.x
25. Nikolenko, S., Koltcov, S., Koltsova, O.: Topic modelling for qualitative studies. Journal of Information Science **43**(1), 88–102 (2017)
26. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval **2**(1-2), 1–135 (2008). https://doi.org/10.1561/1500000001, http://www.nowpublishers.com/article/Details/INR-001
27. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. pp. 79–86. Association for Computational Linguistics (2002)
28. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. Journal of Machine Learning Research **12**(Oct), 2825–2830 (2011)
29. Rogers, A., Romanov, A., Rumshisky, A., Volkova, S., Gronas, M., Gribov, A.: RuSentiment: An enriched sentiment analysis dataset for social media in Russian. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 755–763. Association for Computational Linguistics, Santa Fe, New Mexico, USA (Aug 2018), https://www.aclweb.org/anthology/C18-1064
30. Smetanin, S.: The Applications of Sentiment Analysis for Russian Language Texts: Current Challenges and Future Perspectives. IEEE Access **8**, 110693–110719 (2020). https://doi.org/10.1109/ACCESS.2020.3002215, https://ieeexplore.ieee.org/document/9117010/
31. Thelwall, M., Buckley, K., Paltoglou, G.: Sentiment strength detection for the social web. Journal of the American Society for Information Science and Technology **63**(1), 163–173 (Jan 2012). https://doi.org/10.1002/asi.21662, http://doi.wiley.com/10.1002/asi.21662
32. Tutubalina, E.: Metody izvlecheniya i rezyumirovaniya kriticheskih otzyvov pol'zovatelej o produkcii (Extraction and summarization methods for critical user reviews of a product). Ph.D. thesis, Kazan Federal University, Kazan (2016), https://www.ispras.ru/dcouncil/docs/diss/2016/tutubalina/dissertacija-tutubalina.pdf

33. Zhang, S., Zhang, X., Chan, J., Rosso, P.: Irony detection via sentiment-based transfer learning. Information Processing & Management **56**(5), 1633–1644 (Sep 2019). https://doi.org/10.1016/j.ipm.2019.04.006, https://linkinghub.elsevier.com/retrieve/pii/S0306457318307428