RuSentiment: An Enriched Sentiment Analysis Dataset for Social Media in Russian

Anna Rogers[†], **Alexey Romanov**[†], **Anna Rumshisky**[†],
Svitlana Volkova[‡], **Mikhail Gronas**[§], **Alex Gribov**[†]
[†]Dept. of Computer Science, University of Massachusetts Lowell, Lowell, MA, USA
{arogers,aromanov,arum}@cs.uml.edu,alexander_gribov@student.uml.edu
[‡]Pacific Northwest National Laboratory, Richland, WA, USA
svitlana.volkova@pnnl.gov
[§]Dept. of Russian, Dartmouth College, Hanover, NH, USA
mikhail.gronas@dartmouth.edu

Abstract

This paper presents RuSentiment, a new dataset for sentiment analysis of social media posts in Russian, and a new set of comprehensive annotation guidelines that are extensible to other languages. RuSentiment is currently the largest in its class for Russian, with 31,185 posts annotated with Fleiss' kappa of 0.58 (3 annotations per post). To diversify the dataset, 6,950 posts were pre-selected with an active learning-style strategy. We report baseline classification results, and we also release the best-performing word embeddings trained on 3.2B corpus of Russian social media posts.

1 Introduction

Over the past several years sentiment analysis has been increasingly important in political science (Ceron et al., 2015) and journalism (Jiang et al., 2017). Such applications necessitate resources for languages spoken in the conflict zones. Our study focuses on Russian, which to date has little annotated data (Loukachevitch and Rubtsova, 2016; Koltsova et al., 2016), and no openly available sentiment detection systems beyond the dictionary-based ones. However, lexical features have time and again shown inferior performance compared to the supervised learning approaches using annotated data (Gombar et al., 2017), and lack of such a resource severely limits sentiment analysis applications for Russian.

We present RuSentiment, a dataset of public posts on VKontakte (VK), the largest Russian social network that currently boasts about 100M monthly active users.[1] RuSentiment was developed with new comprehensive guidelines that enabled light and speedy annotation while maintaining consistent coverage of a wide range of explicitly and implicitly expressed sentiment. The overall inter-annotator agreement in terms of Fleiss' kappa stands at 0.58. In total, 31,185 posts were annotated, 21,268 of which were selected randomly (including 2,967 for the test set). 6,950 posts were pre-selected with an active learning-style strategy in order to diversify the data. This makes RuSentiment the largest openly available sentiment dataset for social media, and the largest general domain sentiment dataset for this relatively low-resource language.

We also present baseline classification results on the new dataset. The best results were achieved with a neural network model that made use of word embeddings trained on the VKontakte corpus, which we also release to enable a fair comparison with our baselines in future work. This model achieved an F1 score of 0.728 in a 5-class classification setup. The dataset, the source code for the baseline classifiers, and the in-domain word embeddings are available at the project website[2].

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

[2] http://text-machine.cs.uml.edu/projects/rusentiment/

2 Related Work

Russian is generally not as well resourced as English, and that includes the sentiment analysis data. RuSentiLex, the largest sentiment lexicon for Russian[3] (Loukachevitch and Levchik, 2016), currently contains 16,057 words, which exceeds the size of such manually constructed English resources as, for example, SentiStrength (Thelwall and Buckley, 2013). However, there is nothing like SentiWordNet (Baccianella et al., 2010), SentiWords, (Gatti et al., 2016), or SenticNet (Cambria et al., 2018) for Russian.

There are also few annotated datasets. The datasets from the SentiRuEval 2015 and 2016 competitions are the largest resource that has been available to date (Loukachevitch and Rubtsova, 2016). The SentiRuEval 2016 dataset is comprised by 10,890 tweets from the telecom domain and 12,705 from the banking domain. The Linis project (Koltsova et al., 2016) reports to have crowdsourced annotation for 19,831 blog excerpts, but only 3,327 are currently available on the project website.

The choice of VK social network makes RuSentiment qualitatively different from the above resources. Unlike Linis, it contains standalone mini-texts, and unlike SentiRuEval, the postings vary by length and were not pre-selected by topic. We also found the VK data used for our RuSentiment to be more noisy than SentiRuEval tweets.[4] RuSentiment thus fills an existing gap, providing a large annotated dataset of general-domain posts from the largest Russian social network.

3 Annotation

Our VK data was originally collected for research on political bias, and contained the posts from the personal "walls" (i.e., posts on personal pages) of the users that were members of anti-Maidan and pro-Maidan communities during the 2014 Maidan conflict in Ukraine. RuSentiment only includes the posts that were posted outside these communities, and do not contain political keywords.[5] No pre-selection by topic makes RuSentiment currently the largest manually annotated general domain sentiment dataset for Russian, exceeded in size only by automatically annotated silver dataset by Rubtsova (2015).

To remove noisy posts, we used the following selection criteria. The posts included in the dataset were 10-800 characters in length, at least 50% of which were alphabetical, and at least 30% used the Russian Cyrillic alphabet. URLs and VK postcards were excluded. To ensure the meaningfulness of the posts, we also excluded any posts with over 4 hashtags or less than 2 comments. RuSentiment is distributed without VK post ids, and only includes posts that were posted publicly.[6] The annotation was performed by six native speakers with backgrounds in linguistics over the course of 5 months. The average annotation speed was 250-350 posts per hour. A screenshot of our custom web-interface is shown in Figure 1.

3.1 Annotation Policy

Despite the popularity of the sentiment task, the problem of developing comprehensive, yet easy to follow and "light" guidelines that would ensure high enough agreement is far from being solved. Sentiment is an extremely multi-faceted phenomenon, and each research team in the end has to make its own choices about how it would prefer to treat implicit vs. explicit sentiment, subjective feeling and emotion vs. evaluation, and also irony, sarcasm, and other types of mixed sentiment.

We aimed to develop comprehensive guidelines that would cover the most frequent potentially ambiguous cases and would be easy to apply consistently. Most of the categories we have used have been defined and used before (Liu, 2015; Toprak et al., 2010; Wiebe et al., 2005; Thelwall et al., 2010), and our contribution is mainly their combination that enabled the right balance between coverage and ease of application.

[3]There are at least two more projects that attempt to crowdsource sentiment lexicons: SentiBase (`http://web-corpora.net/wsgi/senti\_game.wsgi/rules`), and Sentimeter (`http://sentimeter.ru/assess/instruction/`). At the moment, both appear to be unfinished.

[4]Baldwin et al. (2013) did not find significant grammatical or spelling differences between Twitter, Youtube comments, or blogs, but domain (telecom in SentiRuEval) could impact the ratio of professionally edited commercial or news-like texts.

[5]The list comprised 169 keywords, including political entities (such as *Moscow* or *Putin*) and words coined and used during the Maidan conflict (such as *ukrop* "dill", a Russian derogatory term for Ukrainians).

[6]This work is covered by an IRB protocol at the authors' institution.

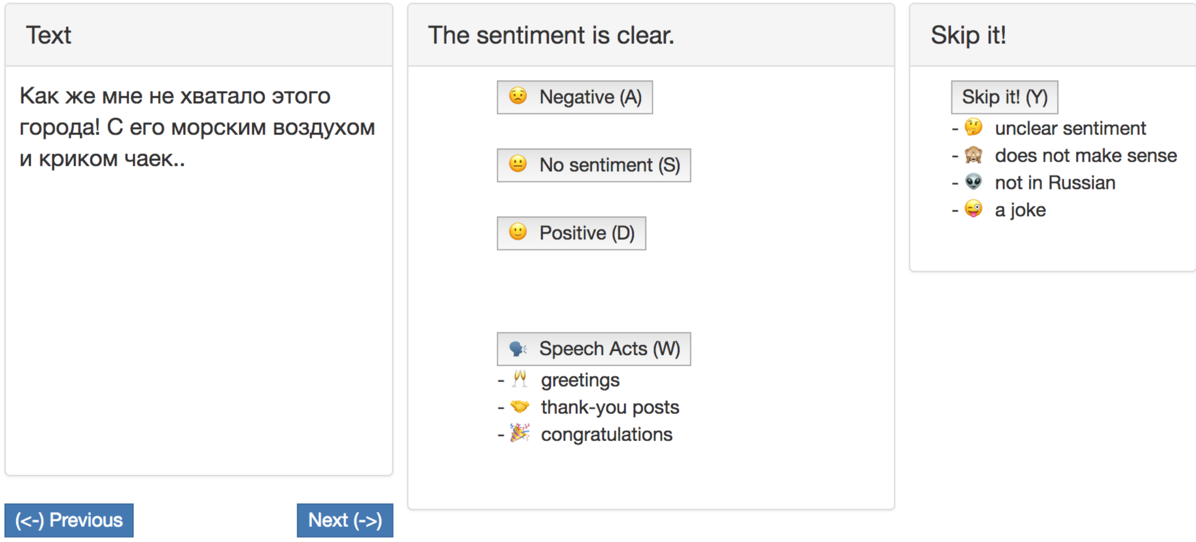


Figure 1: Annotation web-interface.

Finding this point of balance required multiple pilots and extensive linguistic analysis. This difficulty is likely the reason why many sentiment annotation projects, including the Russian crowdsourced initiatives mentioned in Section 2, provide only minimal annotation instructions. Such instructions often yield very inconsistent data. Our guidelines are available at the project website in English, with Russian examples.[7] We hope that they would be useful for subsequent work in other languages and domains.

We prioritized the speed of annotation over detail, opting for a 3-point scale rather than e.g., the 5-point scale in SemEval Twitter datasets (Rosenthal et al., 2017). Thus, the task was to rate the prevailing sentiment in complete posts from VK on a three-point scale (“negative", “neutral”, and “positive”). We also defined the “skip” class for excluding the posts that were too noisy, unclear, or not in Russian (e.g., in Ukrainian). We also made the decision to exclude jokes, poems, song lyrics, and other such content that was not generated by the users themselves. It could be argued that posting jokes on social media should be interpreted as an expression of positive mood, but such data is easy to import from existing web collections.

Although the only sentiment classes we annotated are “positive” and “negative”, RuSentiment guidelines address both explicit and implicit forms of expressions for the speaker’s *internal emotional state* (mood) and *external attitude* (evaluation), as shown in Table 1. These distinctions are often not accounted for in sentiment data, including many of the English datasets (Volkova et al., 2013; Abdul-Mageed and Ungar, 2017). The guidelines cover such cases of implicit sentiment as rhetorical questions, (non-)desirability, recommendations, and descriptions or mentions of the experiences that most people would consider positive or negative.

Additionally, we defined a subcategory of positive posts that covers frequent speech acts, such as expressions of gratitude, greetings, and congratulations. They are very frequent in VK data, and the sentiment they express is overtly positive, but they are also very formulaic. The separate subcategory enables excluding them from the category of positive posts, depending on the practical goals of the analysis. In our binary classification experiments, we chose to exclude this category.

The neutral posts were defined as those that describe something in a matter-of-fact way, without clear sentiment (e.g., *That’s a girl I know.*) They also included factual questions, commercial information, plot summaries, descriptions, etc..

We opted to not define a separate “mixed sentiment” class, as this would not be particularly useful, and is also difficult for models to capture (Liu, 2015, p. 77). All cases of mixed sentiment were annotated as either negative or positive. To improve consistency, the guidelines covered 7 frequent cases of mixed sentiment. For example, irony was annotated with the dominant (usually negative) sentiment (e.g. *I*

[7] http://text-machine.cs.uml.edu/projects/rusentiment/

Mode \ Target	Emotional state	Attitude towards some entity
Explicit	• direct statement of one's mood, emotions, feelings - either positive or negative e.g., УРРРРААААААА!!!!!!)))) *Hooray!!!!!!))))*	• direct statement of one's positive or negative attitude towards the target e.g., годный дабстеп, очень годный. *That's fairly decent dubstep.*
Implicit	• descriptions of experience that most people would consider positive or negative e.g., 15 дней к окончанию сессии... *15 more days of exams...* • rhetorical questions e.g., Вот как я теперь её найду?(((*So how am I supposed to find it now?*	• expressing (non-)desirability of the target, (non-)recommending it e.g., Обязательно попробуй шампусик. *Do taste the champagne.* • rhetorical questions e.g., НАТАШ ТЕБЕ НЕ КАЖЕТСЯ ШО ТЕБЕ ПОРА ЗАНИМАТЬСЯ УЧЕБОЙ? *Natasha, don't you think you should be studying?*

Table 1: Types of positive and negative sentiment covered by RuSentiment annotation guidelines.

broke my heels and got drenched on the way back. What a great day.) Generally neutral posts with some positive formulaic language were annotated as neutral (e.g. *Looking to buy a used electric guitar. Please share this post. Thank you, and have a great day!*) In cases of conflict between a speaker's emotional state and their attitude towards something, we annotated the mood of the speaker. For example, *I miss you* expresses the sadness of the speaker while also implying their high opinion of the addressee, and is annotated as negative.

Hashtags such as *#epicentr* (company name) were considered neutral, but those that could agree or disagree with the general sentiment of the post were treated as sentiment markers. This concerned the hashtags that expressed the sentiment explicitly (e.g. #ihate, #sad) or implicitly, via experiences that most people would consider positive or negative (e.g. #beach, #party).

The annotators were instructed to *not* treat the emoticons as automatic sentiment labels, as done by Go et al. (2009), Davidov et al. (2010), Sahni et al. (2017), and others. Some emoticons do indeed strengthen the message (Derks et al., 2008), but others serve to soften its illocutionary force without changing its content (Ernst and Huschens, 2018).[8] A user may end a post with a "hedging" emoticon just to express friendliness or politeness, and we found that often to be the case for VK data. Therefore, emoticons were not taken into account when no sentiment was expressed verbally or when they aligned with the content of the message. Emoticons were considered relevant only when they contradicted the verbal clues of sentiment. In that case, the annotators were instructed to annotate the dominant overall sentiment of the post, including the emoticons.

In total, five categories were annotated: "Neutral", "Negative", "Positive", "Speech Act", and "Skip".

3.2 Annotating Randomly Selected Posts

In the first stage, 18,453 randomly selected posts were annotated. Fleiss' kappa for three annotators in this sample constituted 0.654. A post was deemed to belong to a class if at least 2 of 3 annotators attributed it to that class. The class distribution is as follows: 41.3% neutral posts, 10.3% negative and 20.5% positive posts, and 9.4% skipped, with an additional 13.9% posts in the speech acts category. In 4.6% cases, all three annotators disagreed, and we included them in the skipped class as unclear. Furthermore, we annotated 2,967 random posts to create a test set (Fleiss' kappa is 0.604).

4 Experiments

4.1 Baseline model selection

We experimented with several classifiers of different types, including logistic regression, linear SVM, and gradient boosting classifier (Pedregosa et al., 2011). We also implemented a simple neural network classifier (NNC) consisting of four fully-connected layers with non-linear activation functions between

[8]This also concurs with classifications drawn in pragmatics and discource analysis, among others, by Kavanagh (2010; Yus and Yus (2014).

(a) Negative posts vs. the rest

(b) Positive posts vs. the rest

Figure 2: Distribution of true labels in probability bins of NNC binary classifier.

them. We used the PyTorch library to implement it.[9] The selected models cover most types of commonly used models, spanning from simple linear models to ensembles and neural networks. The source code for these baselines is available on the project page[10].

The posts were represented either with a sparse TF-IDF representation (Manning et al., 2008) or as an average of word vectors for the constituent tokens. Averaging is by no means the only way to combine word vectors to obtain a representation of a sentence (Mitchell and Lapata, 2010; Baroni et al., 2014; Socher et al., 2012; Hill et al., 2016), but it is one of the computationally cheapest options that still encodes a non-trivial amount of information about the sentences (Adi et al., 2016). FastText (Bojanowski et al., 2017) was chosen for its capacity to represent subword information. This is beneficial for a morphologically rich language such as Russian, especially with a corpus that is as noisy and full of misspellings as our VK corpus.

Model	**Feat.**	**F1**	**Prec.**	**Rec.**
Logistic Regression	CC	0.526	0.619	0.513
	VK	0.622	0.691	0.611
	Wiki	0.574	0.652	0.559
	TF-IDF	0.626	0.654	0.615
Linear SVM	CC	0.586	0.589	0.610
	VK	0.687	0.690	0.691
	Wiki	0.632	0.628	0.646
	TF-IDF	0.664	0.660	0.670
Gradient Boosting	CC	0.527	0.619	0.509
	VK	0.624	0.692	0.611
	Wiki	0.577	0.646	0.557
	TF-IDF	0.587	0.605	0.588
Neural Net Classifier	CC	0.604	0.603	0.623
	VK	**0.717**	**0.718**	**0.717**
	Wiki	0.661	0.658	0.666
	TF-IDF	0.593	0.599	0.589

Table 2: Classifier performance with different post representations on 5-class classification (20,896 random posts).

We performed experiments using the published FastText embeddings trained on CommonCrawl (CC)[11] and on Russian Wikipedia[12] (Wiki). We also trained our own embeddings on 3.2B tokens of VK data. The training corpus included the posts that are part of RuSentiment, but they constitute less than 0.001% of the data. We trained these embeddings with the default FastText parameters, with vector size 300 and minimum frequency 100. Table 2 shows the performance in 5-way classification, for the models trained using the 20,896 posts annotated without active learning. The best accuracy was achieved by the NNC classifier. The in-domain VK embeddings consistently improved performance of all models, as could be expected.

4.2 Active Learning Data Selection Strategy

Unbalanced datasets present difficulties in classification, especially when the classes of interest – in this case, positive and negative sentiment – are in the minority. With the setup described in Section 4.1

[9] `http://pytorch.org`
[10] `http://text-machine.cs.uml.edu/projects/rusentiment/`
[11] `https://github.com/facebookresearch/fastText/blob/master/docs/crawl-vectors.md`
[12] `https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md`

we conducted preliminary experiments with 13,764 posts for training and 3,441 for testing[13]. In these preliminary experiments, the best-performing model (NNC) reached an F1 of 0.637 for the positive class vs. the rest and F1 of 0.550 for negative vs. the rest in binary classification. The recall was very low, 0.477 and 0.587, respectively). Figure 2 shows that the distribution of correct labels in each probability bin was nearly even, which suggests that the classifier did not have a well-formed representation of the target class. For example, the number of examples correctly assigned to the negative polarity class with probability of 0.9 was nearly equal to the number of misclassified negative polarity examples which were assigned the probability of 0.1. Although there were twice as many training examples, the same trend was present for the positive polarity posts.

To provide the classifier with more examples that it was unsure about, we used NNC to pre-select additional 3,500 "negative" and 2,500 "positive" posts with certainty sampling (Koncz and Paralič, 2013; Fu et al., 2013). We drew an equal number of samples from the probability bins 0.3-0.7, annotating additional 6,950 posts.

4.3 Active Learning Effect

Figure 3 shows that the distribution of positive posts in the pre-selected sample turned out to be similar to the original one. However, the classifier was successful in reducing the number of skipped and speech act posts, and the ratio of negative posts increased.

Fleiss' kappa for the pre-selected sample was much lower (0.449), bringing the overall number down. The reason was the higher ratio of posts with agreement of two, rather than three annotators. We interpret this as success in bringing more borderline and diverse cases into the dataset, even at the cost of the overall agreement.

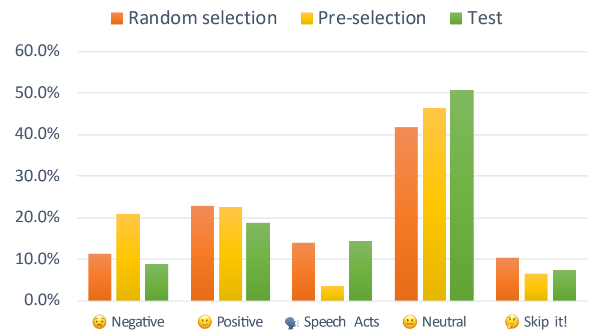


Figure 3: Sentiment class distribution in RuSentiment.

In order to investigate the effects of adding the pre-selected sample to the dataset, we conducted additional 5-way classification experiments using NNC and VK embeddings. Table 3 shows the comparison between the model trained on all the randomly selected examples and the model trained on the data in which 6,950 randomly selected examples were replaced with the ones pre-selected with the active learning strategy described above.[14] Replacing randomly selected examples with pre-selected ones yielded consistent, albeit small, improvement, with the recall being affected the most. Table 3 also shows that the 5-way classification results using the full training data, including the pre-selected sample, produces the best results. The performance is averaged across three runs.

Input data	**F1**	**Prec.**	**Rec.**
Randomly selected posts only (21,268)	0.717	0.718	0.717
6,950 out of 21,268 random posts replaced with pre-selected	0.720	0.722	0.730
21,268 random + 6,950 pre-selected posts	**0.728**	**0.729**	**0.736**

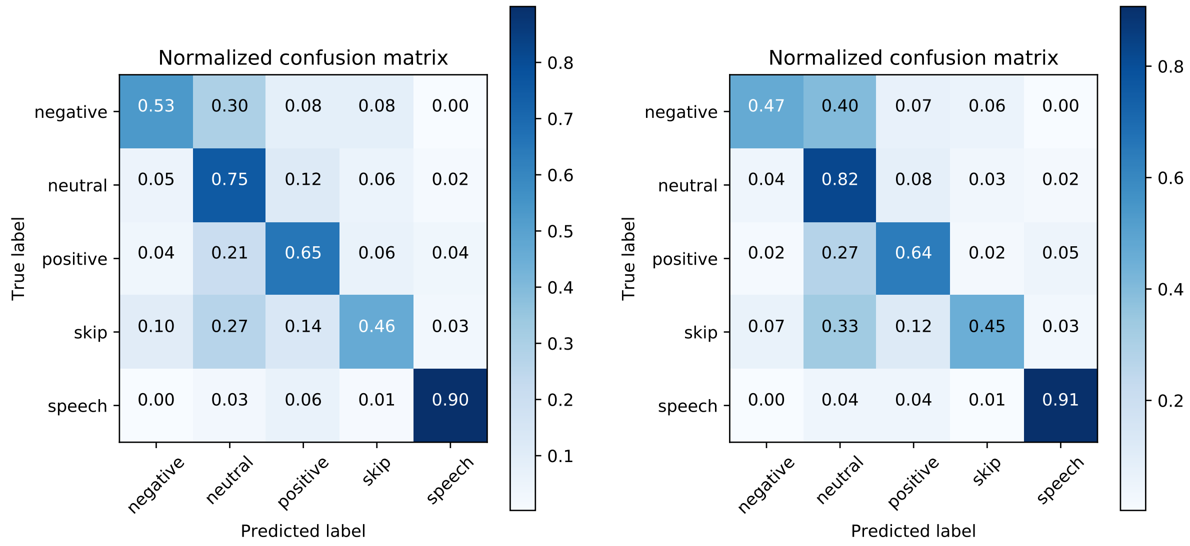
Table 3: 5-class classification with NNC and VK embeddings: the effect of pre-selected posts.

4.4 Error Analysis

Figure 4 presents the confusion matrices for RuSentiment test set with the NNC classifier trained in two settings: (a) using the base randomly selected posts only, and (b) using the additional posts pre-selected with active learning. There is a clear shift in the distribution of neutral and negative posts. Compared to

[13]These samples do not intersect with the final test set.

[14]An additional 463 randomly selected examples were included in this experiment and are released in the final dataset. Out of these, 372 examples were added to the training set and 91 to the test set. These data came in late from the annotation team, and additional experiments confirmed that they did not affect the overall pattern of model performance.



(a) Random posts only

(b) Random and pre-selected posts

Figure 4: Confusion matrix for the RuSentiment test set with different training sets (true labels in rows, predicted labels in columns).

the classifier trained only on randomly selected posts, the classifier trained on the entire dataset makes more errors misclassifying neutral and, to a lesser degree, skipped posts as negative. This suggests that the pre-selected sample makes the classifier more sensitive to borderline cases, since their amount is increased. Interestingly, the positive posts category was not affected in the same way, presumably due to the larger amount of positive posts in the base data.

5 Conclusion

We presented RuSentiment, a new general domain sentiment dataset for Russian social media, built with data from VK, Russia's largest social network. RuSentiment includes 31,185 posts, each carrying 3 annotations with Fleiss' kappa of 0.58, and is currently the largest openly available dataset of its class for Russian. RuSentiment was developed with new guidelines that enabled light, speedy and consistent 5-class annotation of explicit and implicit sentiment, and could be adapted for other languages.

We also presented 4 baseline models using FastText word embeddings as well as TF-IDF representation. The best performance was F1 of 0.728 in 5-class classification. It was achieved by a neural net classifier with in-domain FastText embeddings, which we release to enable fair comparison with future systems.

Acknowledgements

This work was supported in part by the U.S. Army Research Office under Grant No. W911NF-16-1-0174.

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. Emonet: Fine-grained emotion detection with gated recurrent neural networks. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, pages 718–728.

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. arXiv preprint arXiv:1608.04207.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Language Resources and Evaluation Conference, volume 10, pages 2200–2204.

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources? In Proceedings of the Sixth International Joint Conference on Natural Language Processing, pages 356–364.

Marco Baroni, Raffaela Bernardi, and Roberto Zamparelli. 2014. Frege in space: A program of compositional distributional semantics. Linguistic Issues in Language Technology, 9.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5(0):135–146, June.

Erik Cambria, Soujanya Poria, Devamanyu Hazarika, and Kenneth Kwok. 2018. SenticNet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. In AAAI.

Andrea Ceron, Luigi Curini, and Stefano M. Iacus. 2015. Using sentiment analysis to monitor electoral campaigns: Method matters—evidence from the United States and Italy. Social Science Computer Review, 33(1):3–20, February.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced sentiment learning using twitter hashtags and smileys. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 241–249. Association for Computational Linguistics.

Daantje Derks, Arjan ER Bos, and Jasper Von Grumbkow. 2008. Emoticons and online message interpretation. Social Science Computer Review, 26(3):379–388.

Claus-Peter Ernst and Martin Huschens. 2018. The effects of different emoticons on the perception of emails in the workplace. In Proceedings of the 51st Hawaii International Conference on System Sciences.

Yifan Fu, Xingquan Zhu, and Bin Li. 2013. A survey on instance selection for active learning. Knowledge and Information Systems, 35(2):249–283, May.

Lorenzo Gatti, Marco Guerini, and Marco Turchi. 2016. SentiWords: Deriving a high precision and high coverage lexicon for sentiment analysis. IEEE Transactions on Affective Computing, 7(4):409–421, October.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1(12).

Paula Gombar, Zoran Medić, Domagoj Alagić, and Jan Šnajder. 2017. Debunking sentiment lexicons: A case of domain-specific sentiment classification for Croatian. In Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing, pages 54–59.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. In Proceedings of NAACL-HLT 2016, pages 1367–1377, San Diego, California, June 12-17, 2016. Association for Computational Linguistics.

Ye Jiang, Xingyi Song, Jackie Harrison, Shaun Quegan, and Diana Maynard. 2017. Comparing attitudes to climate change in the media using sentiment analysis based on Latent Dirichlet Allocation. In Proceedings of the 2017 EMNLP Workshop: Natural Language Processing Meets Journalism, pages 25–30.

Barry Kavanagh. 2010. A cross-cultural analysis of Japanese and English non-verbal online communication: The use of emoticons in weblogs. Intercultural Communication Studies, 19(3):65–80.

O.Yu. Koltsova, S.V. Alexeeva, and S.N. Kolcov. 2016. An opinion word lexicon and a training dataset for Russian sentiment analysis of social media. In Proceedings of the International Conference on Computational Linguistics and Intellectual Technologies Dialog 2016, pages 277–287, Moscow, June 1–4, 2016.

Peter Koncz and Ján Paralič. 2013. Active learning enhanced document annotation for sentiment analysis. In Availability, Reliability, and Security in Information Systems and HCI, Lecture Notes in Computer Science, pages 345–353. Springer, Berlin, Heidelberg, September.

Bing Liu. 2015. Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge University Press, Cambridge.

Natalia V. Loukachevitch and Anatolii Levchik. 2016. Creating a general russian sentiment lexicon. In Proceedings of the Tenth International Conference on Language Resources and Evaluation, pages 1171–1176, Portorož, Slovenia. ELRA.

N.V. Loukachevitch and Yu.V. Rubtsova. 2016. SentiRuEval-2016: Overcoming time gap and data sparsity in Twitter sentiment analysis. In Proceedings of the International Conference on Computational Linguistics and Intellectual Technologies Dialog 2016, pages 375–384.

Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. 2008. Introduction to information retrieval, volume 1. Cambridge university press Cambridge.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. Cognitive science, 34(8):1388–1429.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12:2825–2830.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pages 502–518.

Yu.V. Rubtsova. 2015. Constructing a corpus for sentiment classification training. Mezhdunarodnyj zhurnal "Programmnye produkty i sistemy", 27:72–78, March.

Tapan Sahni, Chinmay Chandak, Naveen Reddy Chedeti, and Manish Singh. 2017. Efficient Twitter sentiment classification using subjective distant supervision. In Communication Systems and Networks (COMSNETS), 2017 9th International Conference On, pages 548–553. IEEE.

Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 1201–1211.

Mike Thelwall and Kevan Buckley. 2013. Topic-based sentiment analysis for the social web: The role of mood and issue-related words. Journal of the American Society for Information Science and Technology, 64(8):1608–1617, August.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. Journal of the Association for Information Science and Technology, 61(12):2544–2558.

Cigdem Toprak, Niklas Jakob, and Iryna Gurevych. 2010. Sentence and expression level annotation of opinions in user-generated discourse. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 575–584, Uppsala, Sweden 11-16 June 2010.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring sentiment in social media: Bootstrapping subjectivity clues from multilingual twitter streams. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), volume 2, pages 505–510.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. Language Resources and Evaluation, 39(2-3):165–210, May.

Francisco Yus and Francisco Yus. 2014. Not all emoticons are created equal. Linguagem em (Dis)curso, 14(3):511–529, December.