Estimating Aggressiveness of Russian Texts by Means of Machine Learning

Dmitriy Levonevskiy(✉), Dmitrii Malov, and Irina Vatamaniuk

St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), 14th Line, 39, 199178 St. Petersburg, Russia
DLewonewski.8781@gmail.com

Abstract. This paper considers emotional assessment of texts in Russian using machine learning on the example of aggression detection. It summarizes the related work, methods, models and datasets, describes actual problems, proposes a text processing pipeline and a software system for training neural networks on heterogeneous datasets. The experiments show that neural networks trained on the annotated corpora both in Russian and English, allow to determine whether a text item in Russian contains an aggressive message. Authors thoroughly compare different assessment methods, particularly corpus-based approaches, machine learning solutions and hybrid variants. Results, obtained here, can be used to estimate the aggressiveness probability, for example, to rank messages for subsequent manual verification. They also enable feasibility studies on the possibilities of detecting a particular type of emotion in a text using corpora in other languages. The paper highlights further research directions, where different Python toolkits (NLTK, Keras) could be used for better model performance.

Keywords: Emotion detection · Sentiment Analysis · Natural language processing · Text analysis · Aggressive text detection · Neural networks · Machine learning

1 Introduction

The problem of emotion detection in text is of current interest, as it can be applied in various domains: network discussion moderation, analysis of public opinion on companies, goods, events; text classification [1, 2]. At the same time this problem causes a lot of difficulties. The problems associated with the task of automating of emotion detection in text content are related to ambiguity and subjectivity of the natural language. It should be considered that the methods of identifying emotions are practically limited and, as a rule, are suitable primarily for detection of explicit emotions [2]. A more difficult task consists, for example, in identification of implicit aggression and, more generally, in correct processing of the content that can be either aggressive or neutral when taken out of context.

Moreover, it is necessary to pay attention to the peculiarities of the environment. In particular, discussions in social media and forums may contain heterogeneous textual and audiovisual content in different languages [21]. Depending on the analyzed media,

A. A. Salah et al. (Eds.): SPECOM 2019, LNAI 11658, pp. 270–279, 2019.
https://doi.org/10.1007/978-3-030-26061-3_28

the common terms, jargon, memes, lexicon and cultural canons of social groups may differ significantly. The techniques used by intruders to bypass auto-moderation in social media complicate technical text processing. The content is also characterized by the presence of messages with spelling errors, typos, punctuation quirks, emoticons. Poor grammatical correctness and vague syntactic structure of social media posts complicates the usage of natural language processing tools [8]. The task turns out to be challenging even for human annotators, although they could refer to context of each message [9].

Another feature of the social media content is a large number of short messages: such messages can be classified well only provided, they contain explicitly expressed emotions. Another problem consists in detecting sarcasm and irony in text messages as there is no agreement on formal description of these concepts. The results in [17] are satisfactory but have a limited practical applicability.

Large amount and heterogeneous structure of the content require its preprocessing, before the methods described here could be applied. The preprocessing is performed by reducing the text dimension for further consumption by neural networks and other classifiers. The diversity of the social network content complicates the research: it should be noticed that working on domain-specific corpus gives better results than working on the domain-independent corpus [5].

2 Related Work

2.1 Methods and Systems

Considering the aggression as a kind of sentiment expressed in text, we can use Sentiment Analysis (SA) as a method of data mining [13] for its detection. SA identifies the sentiment expressed in a text and then analyzes it. The datasets used in SA are of high importance in this field. The social network sites and micro-blogging sites are considered a very good data source because people share and discuss their opinions about a certain topic freely there [5]. Fields in SA include emotion detection (ED) that aims to extract and analyze emotions, both explicit and implicit, present in the sentences. It was argued in [15] that there are eight basic and prototypical emotions, specifically: joy, sadness, anger, fear, trust, disgust, surprise, and anticipation; there are also more approaches as well [27]. The problem is either handled as a binary classification case, where only positive and negative sentiments are considered, or as a multiclass classification problem when a fine-grained list of sentiments is used (e.g., anger, disgust, fear, guilt, interest, joy, sadness, shame, surprise) [4].

The difference between SA and ED consists in following: SA is concerned mainly in specifying positive or negative opinions, whereas ED is concerned with detecting various emotions from text. As a SA task, ED can be implemented using ML approach or Lexicon-based approach, but Lexicon-based approach is more common one [5].

In order to implement SA or ED, feature selection (FS) should be carried out first of all. FS may be performed by lexicon-based methods that require human annotation, and statistical methods which are automatic methods that are more frequently used; statistical methods may ignore or retain the information on the word sequence [5].

Key features mostly used for ED are terms presence and frequency [16], parts of speech (POS), opinion words and phrases, negations.

As an example of such features we can consider activity markers, psycholinguistic, lexical and semantic markers described in [14]. Natural language markers allow evaluating possibly aggressive or other harmful text aspects (presence of manipulative techniques, negative emotional background), reveal "hot" news characteristic of tabloid press, fake news, etc. Psycholinguistic markers (number of personal pronouns, POS frequency ratios, etc.), lexical markers (injective lexicon, destructive semantics) can be measured and used for text analysis.

Various methods for emotional text classification and, in particular, for aggressive text detection, are discussed in review articles [5, 12] and in the article [10]. Some web services for solving SA tasks are analyzed in [11]. At the same time, a lot of sources deal with a binary classification problem of single messages, without analyzing entire threads; they often employ a very similar text preprocessing pipeline comprising stop-word removal, tokenization, POS tagging, emoticon detection, stemming, etc., and a typical text feature extraction step which resulted in bag-of-words, or, bag-of-stems representations [4]. Some methods that deal with the problems specified in the previous section are summarized in Table 1.

Table 1. Methods for emotion detection.

Method	Classes	Features	Advantages	Application
Hybrid (lexicon-based + super-vised machine learning) [1]	Anger, disgust, fear, happiness, sadness, surprise, trust, neutral; bullying and neutral posts and threads	Sentiment uni- and bi-grams (occurrences of sentiment changes in consecutive posts); personal pronouns; bullying bi- and tri-grams (using BullyTracer lexicon)	Performs sentiment analysis at message level, but considers the whole threads as the context; building "sentiment n-grams" for threads	Web forums, discussions (tested on MySpace)
Machine learning (random forest) [3]	Bully/aggressive/spam/normal messages	User-based: post frequency, account existence time, etc.; text-based: number of hashtags, emoticons, upper cases, emotional scores, etc.; network-based: follower and friend lists, etc.	Deals with short and imperfect messages, takes the context (chains of tweets) into account, tries to detect sarcasm and trolling	Social media (Twitter)

(*continued*)

Table 1. (*continued*)

Method	Classes	Features	Advantages	Application
Corpus-based approach [4]	Types of bullying (threat/blackmail; insult; curse/exclusion; defamation; sexual talk), victim defense, encouragement to the harasser, other	Word and character n-grams in bag of words; term lists (for example, "allness" indicators, intensifiers, etc.); subjectivity lexicon features	Deals with short and distorted messages. Robustness to spelling variations	Social media (ASK.fm, etc.)
Support vector machine + recurrent neural network [5]	Openly aggressive/covertly aggressive/not aggressive	GloVe features; sentiment scores according to SentiWordNet features; N-gram TF-IDF features	Detecting messages with covert aggression	Social media (Facebook, Twitter, etc.)
Traditional and deep machine learning [6]	Openly aggressive/covertly aggressive/not aggressive	Bad words; POS tags; text length; capitalization; numerical tokens; named entities; sentiment polarity	Deals with short messages and their context. Detecting messages with covert aggression	Social media (Facebook, Twitter)
Hybrid classifier (Naïve Bayes, random forest, support vector machine, logistic regression) [7]	Positive/negative polarity	Vector representation of the "Bag of words"	Classifies short messages	Social media (Twitter)
Profile-based representations (TF-IDF, NN) [8]	Aggressive texts, including sexual aggression	Word and character n-grams	Early recognition for sexual predator detection and aggressive text identification. Possible application for irony/sarcasm detection, opinion mining, etc.	Social media
NB, SVM, and DT [14]	Ironic and sarcastic texts	N-grams, POS n-grams, funny, positive/negative, affective, pleasantness profiling	Irony and sarcasm detection	Amazon reviews

Among the considered approaches, neural networks show the most robust and high performance [9, 10]. While applying the methods described above, some problems still remain. In particular, the overwhelming majority of methods require that corpora of labeled texts exist. Beside the tasks of constructing such a corpus for the Russian language, the problem is that the social media lexicon is volatile, so the corpus becomes obsolete.

The language problem is also significant: the majority of methods are optimized for English language; some other languages under research are Germanic and Latin languages, some languages of South-Eastern Asia and the Near East.

The text analysis services mentioned in [11] are shown in Table 2. It should be noted that some services described there are not available now, though they are said to be able to provide a wide range of possibilities, including evaluating not only the message polarity, but also the separate emotional constituents like fear, gratitude, shame (Lymbix).

Table 2. Services for text analysis.

Service	Possibilities	Methods
SentiStrength http://sentistrength.wlv.ac.uk/	Estimating the polarity of short messages	Text Mining – detecting "good" and "bad" words, their relations
OpinionCrawl http://www.opinioncrawl.com/	Estimating the polarity of relation to a certain subject in the web (news, analytics) Languages: EN, FR, DE, SP	Text mining and multidocument Summarization
OpenDover http://demo.opendover.nl/	Extracts semantic features from the text, calculates the text rating	Ontologies on applied areas (law, education, etc.)
Semantria http://semantria.com/	Languages: EN, FR, DE, SP, PT Good accuracy (about 74% в [9])	Connotative lexicon, calculating frequency of such words in text and their proximity to the object in question
Sentiment140 http://www.sentiment140.com/	Tweets classification (positive/negative/neutral)	Naive Bayes, Maximum Entropy, Support Vector Machines (SVM)
uClassify https://uclassify.com/browse https://uclassify.com/browse/uclassify/sentiment	Provides a set of classifiers for language, sentiment detection, text gender and age recognition Languages: EN, SP, FR, SE Good accuracy (about 76% in [9])	ML: trained on 2.8 million documents with data from Twitter, Amazon product reviews and movie reviews

2.2 Datasets

The data problem arises most pronounced when analyzing non-English texts. For example, there is an annotated corpus of messages from more than 200 000 units [19] in Russian, but those messages are classified just as negative and positive, without any detailed description of the emotions expressed. Datasets in English are much more diverse. Some of them are analyzed in [18]. These datasets are characterized by a large variety in emotion handling: classification by Ekman [20], Plutchik [15], and also some

other approaches are present. Datasets of tweets in Russian [19], "The Emotion in Text, published by CrowdFlower" (39 740 tweets, Ekman) [22], TEC (Twitter Emotion Corpus, Ekman) [23], Emobank (Valence - Arousal - Dominance) [24] were used as well as some smaller corpora. In this work they were processed separately to determine which corpora provide the most accurate results.

One of the options for the use of English-language datasets for the classification of Russian-language text is the use of machine translation. Currently, machine translation systems show quite good results when using English as source or target language. Translation causes accuracy loss, but it can be assumed that the features discussed in Sect. 2.1 are preserved to a large extent.

3 Processing Scheme

To handle various datasets in uniform manner, they were supplied by JSON metadata files containing descriptions of the dataset format and structure. Such file pairs were used as the input data. Firstly, a cleanup operation is performed on the datasets, particularly, removal of irrelevant and special characters, hyperlinks, identifiers. Then comes standardization of whitespace characters, converting all characters to uniform case. In addition, the converted versions (translated and normalized) are created for the datasets.

Emotion estimates were converted into a numerical form. For datasets providing binary classification [19], the estimate was normalized. For the datasets annotated with a variety of emotions, transformed datasets were created with score values in the range [0; 1] for each considered emotion. In the context of identifying aggression, the classes "hate", "anger", "aggression", etc. were assigned the value 1.0; all classes that do not carry any negative constituent ("happiness", "fun", "trust") were characterized by the value 0.0; neutral classes with 0.5; classes with negative properties that do not characterize aggression explicitly ("fear", "worry", "boredom") were described with values from the range (0.5; 1).

For the datasets, n-gram dictionaries are built. In this paper, n-grams of characters and words with different values of n were used. The approach with n = 1 for words is identical to the "bag of words" concept. The n-gram occurrence is used to build vectors for neural network training.

Summarizing the aforementioned concerns, the pipeline of data preprocessing can be represented in Fig. 1.

Fig. 1. Dataset preprocessing pipeline

To organize the full processing pipeline, the following class model was developed (Fig. 2):

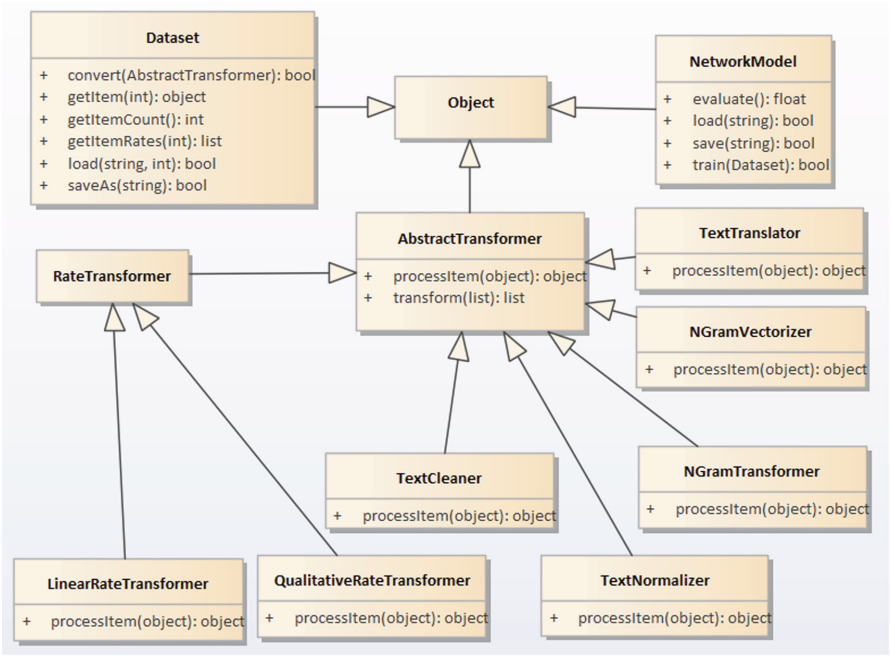

Fig. 2. Classes for data processing

For data processing, as well for creating and training neural networks Python 3.6 was used. The train and test datasets comprise 67% and 33% of the original datasets, respectively. Libraries NLTK and Keras were applied to process text data and train neural networks, respectively, to predict the text aggressiveness using a regression predictive model.

4 Experiments

The modelling results are shown in Figs. 3 and 4. Experiments show that the highest accuracy is achieved for binary classification using the original Russian corpus. Text normalization does not positively influence the result, which can be explained by the semantic loss caused by converting word forms. The considered neural network architectures contained 1 or 2 hidden layers and up to k neurons, where k is the vector size. The maximum accuracy 83% was achieved with the configuration of a neural network with 2 hidden layers consisting of 50 neurons each. The achieved accuracy is lower than in the work [26], but it deals with domain-specific texts (film, customer reviews) which simplifies the classification task.

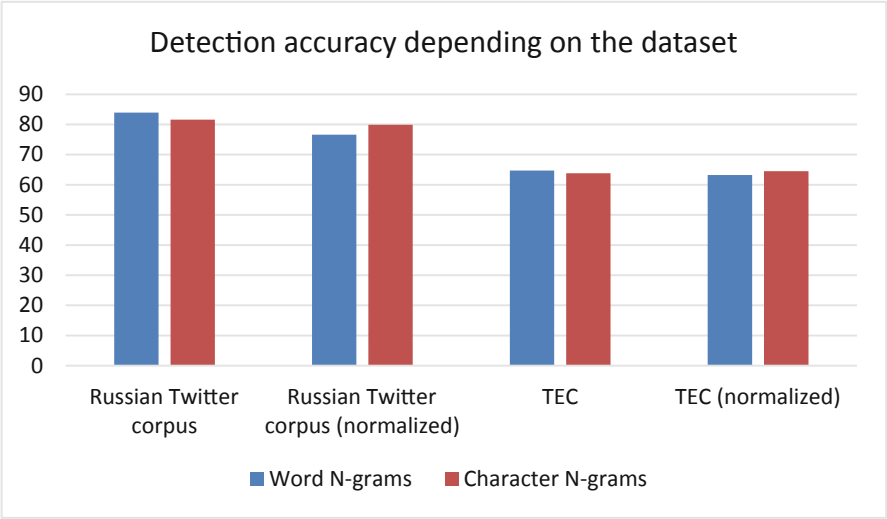

Fig. 3. Accuracy of aggression detection depending on the input data

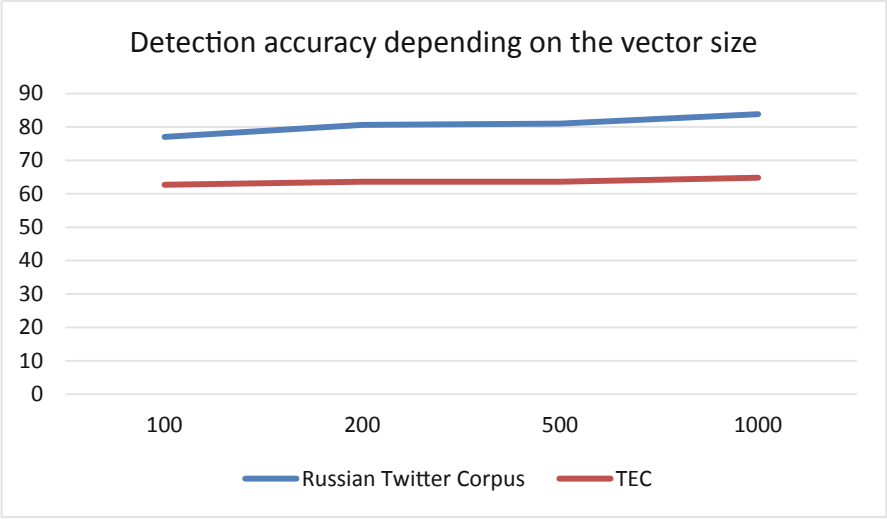

Fig. 4. Accuracy of aggression detection depending on the vector size

The use of machine translation enables distinguishing particular emotions, but the accuracy is much lower in this case. The best results were achieved when using the TEC dataset (65%) for recognizing aggressive messages. Approximately the same accuracy was obtained in [25], and the authors are also able to distinguish overtly and covertly aggressive messages, but that work deals with English texts, so the authors could use English corpora directly.

Using datasets smaller than of 10 000 items did not result in a sufficient accuracy level. Despite the high error rate, such an approach can be used to estimate the aggressiveness probability, for example, to rank messages for subsequent manual verification.

5 Conclusions

The conducted experiments show that, using neural networks trained on the annotated corpora both in Russian and English it is possible to determine with a certain accuracy whether a text item in Russian contains an aggressive message. Such results can be used to estimate the aggressiveness probability, for example, to rank social network messages for subsequent manual verification or to adjust the chatbot behavior models. These results also enable feasibility studies on the possibility of detecting particular emotion types, i.e. fear, interest, in a text using corpora in another languages.

Further research directions include comparison of different approaches to build dictionaries and reduce vector dimensions, comparative analysis and feasibility studies of detecting particular types of emotions, complex analysis of multimodal content on the basis of the technique proposed in [21].

Acknowledgment. This research is supported by the Russian Foundation for Basic Research (project No. 18-29-22061_MK).

References

1. Kocharov, D.A., Menshikova, A.P.: Detection of prominent words in Russian texts using linguistic features. SPIIRAS Proc. **6**, 216–236 (2017)
2. Glazkova, A.V.: An approach to text classification based on age groups of addressees. SPIIRAS Proc. **3**, 51–69 (2017)
3. Vorobiev, V.I., Evnevich, E.L., Levonevskiy, D.K., Fatkieva, R.R., Fedorchenko, L.N.: A study and selection of cryptographic standards on the basis of text mining. SPIIRAS Proc. **5**, 69–87 (2016)
4. Ventirozos, F.K., Varlamis, I., Tsatsaronis, G.: Detecting aggressive behavior in discussion threads using text mining. In: Gelbukh, A. (ed.) CICLing 2017. LNCS, vol. 10762, pp. 420–431. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-77116-8_31
5. Medhat, W., Hassan, A., Korashy, H.: Sentiment analysis algorithms and applications: a survey. Ain Shams Eng. J. **5**(4), 1093–1113 (2014)
6. Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., Vakali, A..: Mean birds: detecting aggression and bullying on twitter. In Proceedings of the 2017 ACM on Web Science Conference, pp. 13–22. ACM (2017)
7. Van Hee, C., et al.: Automatic detection of cyberbullying in social media text. PLoS One **13** (10), e0203794 (2018)
8. Tommasel, A., Rodriguez, J.M., Godoy, D.: Textual aggression detection through deep learning. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, TRAC-2018, pp. 177–187 (2018)

9. Golem, V., Karan, M., Šnajder, J.: Combining shallow and deep learning for aggressive text detection. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, TRAC-2018, pp. 188–198 (2018)
10. Escalante, H.J., Villatoro-Tello, E., Garza, S.E., López-Monroy, A.P., Montes-y-Gómez, M., Villaseñor-Pineda, L.: Early detection of deception and aggressiveness using profile-based representations. Expert Syst. Appl. **89**, 99–111 (2017)
11. Serrano-Guerrero, J., Olivas, J.A., Romero, F.P., Herrera-Viedma, E.: Sentiment analysis: a review and comparative analysis of web services. Inf. Sci. **311**, 18–38 (2015)
12. Mäntylä, M.V., Graziotin, D., Kuutila, M.: The evolution of sentiment analysis—a review of research topics, venues, and top cited papers. Comput. Sci. Rev. **27**, 16–32 (2018)
13. Jo, H., Kim, S.M., Ryu, J.: What we really want to find by sentiment analysis: the relationship between computational models and psychological state. arXiv preprint arXiv: 1704.03407 (2017)
14. Smirnov, I.V., SHelmanov, A.O., Kuznecova, E.S., Hramoin, I.V.: Semantiko-sintaksicheskij analiz estestvennykh yazykov. CHast' II. Metod semantiko-sintaksicheskogo analiza tekstov (Semantic-syntactic analysis of natural languages. Part II. Method of semantic-syntactic analysis of texts). Iskusstvennyj intellekt i prinyatie reshenij, vol. 1, pp. 11–24. ISA RAS, Moscow (2014)
15. Plutchik, R.: A general psychoevolutionary theory of emotion. In: Theories of Emotion, pp. 3–33. Academic Press (1980)
16. Mejova, Y., Srinivasan, P.: Exploring feature definition and selection for sentiment classifiers. In: Fifth International AAAI Conference on Weblogs and Social Media (2011)
17. Reyes, A., Rosso, P.: Making objective decisions from subjective data: detecting irony in customer reviews. Decis. Support Syst. **53**(4), 754–760 (2012)
18. Bostan, L.A.M., Klinger, R.: An analysis of annotated corpora for emotion classification in text. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 2104–2119 (2018)
19. Rubtsova, Y.: Constricting a corpus for sentiment classification training. Softw. Syst. **1**(109), 72–79 (2015)
20. Ekman, P.: An argument for basic emotions. Cogn. Emot. **6**(3–4), 169–200 (1992)
21. Levonevskii, D., SHumskaya, O., Velichko, Uzdyaev, M., Malov, D.: Methods for determination of psychophysiological condition of user within smart environment based on complex analysis of heterogeneous data. Paper presented at the 14th International Conference on Electromechanics and Robotics "Zavalishin's Readings", ER(ZR)-2019 (2019)
22. Sentiment Analysis in Text. https://data.world/crowdflower/sentiment-analysis-in-text. Accessed 15 Feb 2019
23. Emotion, Sentiment, and Stance Labeled Data. http://saifmohammad.com/WebPages/SentimentEmotionLabeledData.html. Accessed 21 Jan 2019
24. Buechel, S., Hahn, U.: EMOBANK: studying the impact of annotation perspective and representation format on dimensional emotion analysis. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, vol. 2, pp. 578–585 (2017)
25. Risch, J., Krestel, R.: Aggression identification using deep learning and data augmentation. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (co-located with COLING), pp. 150–158 (2018)
26. Yussupova, N., Bogdanova, D., Boyko, M.: Applying of sentiment analysis for texts in Russian based on machine learning approach. In: IMMM 2012: The Second International Conference on Advances in Information Mining and Management, pp. 8–14 (2012)
27. Neidenthal, P.M., Kranth-Gruber, S., Ric, F.: Psychology of Emotions: Interpersonal, Experiential, and Cognitive Approach. Psychology Press, New York (2006)