Neural Network Doc2vec in Automated Sentiment Analysis for Short Informal Texts

Natalia Maslova[1(✉)] and Vsevolod Potapov[2]

[1] Institute of Applied and Mathematical Linguistics,
Moscow State Linguistic University, Ostozhenka 38, Moscow 119034, Russia
natalia.maslova277@gmail.com
[2] Faculty of Philology, Lomonosov Moscow State University,
GSP-1, Leninskie Gori, Moscow 119991, Russia
volikpotapov@gmail.com

Abstract. The article covers approaches to automated sentiment analysis task. Under the supervised learning method a new program was created with the help of Doc2vec – a module of Gensim that is one of Python's libraries. The program specialization is short informal texts of ecology domain which are parts of macropolylogues in social network discourse.

Keywords: Automated sentiment analysis · Social network discourse · Deprivation · Supervised learning · Word embeddings · Russian

1 Introduction

The automated sentiment classification task is very relevant for opinion mining. Sentiment analysis is rapidly developing nowadays. Researchers need new instruments to put automated text processing on a large scale. There are such methods developed by now as: (a) rules- and lexicon-based analysis, (b) supervised learning, (c) unsupervised learning.

Lexicon-based sentiment classification is very popular at present. It forms the basis of many massmedia monitoring systems which work upon positive and negative lists of words – tonality dictionaries. The machine calculates representatives of which group prevail. Lexicon-based analysis is usually rules-based: n-grams are processed in a different way, syntagmatic boundaries marked with punctuation characters are taken into account as well as diminutives, augmentatives and negation. Beyond that a variation of this method was developed according to which words play different role in text sentiment formation (graph-theoretical models for Norwegian [2], for Russian [19]). Such algorithms construct text graphs, rank nodes and compute word weights based on the sentiment dictionary and the word rank. The mentioned method found its practical application in online Russian media monitoring systems such as Integrum[1], Mediaology[2], IQBuzz[3], SemanticForce[4], PalitrumLab[5] (for a review of the latter system see [12]).

[1] www.integrum.ru/.
[2] www.mlg.ru/.
[3] www.iqbuzz.pro/.
[4] www.semanticforce.net/.
[5] www.palitrumlab.ru/.

A. Karpov et al. (Eds.): SPECOM 2017, LNAI 10458, pp. 546–554, 2017.
DOI: 10.1007/978-3-319-66429-3_54

Nevertheless, lexicon-based sentiment analysis has its own weaknesses. First, a word gets its expressive meaning only when it becomes a part of an utterance [1, 20]. Before this moment, while the word is only a part of the language system it cannot have an expressive meaning even if it belongs to the emotive lexicon. Second, this method gets stuck with polysemy, homonymy and idioms. Third, it ignores nonce words which occur quite often in unofficial texts. Then, such systems are very vulnerable to errors and misprints (though Thelwall et co. [18] made a smart effort to work it around). Last, sentiment dictionary creation involves a lot of human resources, deep linguistic work. That is why other algorithms of media monitoring should be developed.

POS-labelling and word recognition were confronted with a lot of pitfalls such as polysemy, homonymy, idioms and their modifications, word coinage. A new step was made in big text data processing with the creation of neural network modules (also known as word embeddings or neural language models) Word2vec and Doc2vec by T. Mikolov [9]. These modules were developed on the basis of R. Rehurek's library Gensim [15]. No labels, minimum preprocessing, a big scale of analyzed data – these are the advantages of neural networks. Word2vec models work fine on micro-texts such as posts on Tweeter [10, 17]. However, the results are not so encouraging on longer texts due to the word order ignorance. Word2vec is an example of a popular bag-of-words method [16].

Baroni et al. compare the new method with the classic distribution semantic models, or context-predicting with context-counting models, and gets empirical demonstration of the former's excellence [3]. Levy and Goldberg used dependency-based contexts "as an alternative to the linear bag-of-words approach". With the help of parsing technologies they derived contexts based on the syntactic relations the word participates in [8]. Because of the sentiment analysis popularity a lot of forums are created to discuss this problem, for example [5, 21]. The program code described there got development in our software solution.

Social network discourse (SND) has been under the spotlight of internet linguistics since the end of the XX-th century [4]. For example, [7] suggested the term "massive polylogue" under which "a multilingual and global comment thread following some video" was meant. We make use of Potapova's SND definition: "It is a special electronic macropolylogue, considering the relevant categories of its form, content and functional weight" [12]. The mentioned categories include the following:

(1) The passport of the utterance (URL, data and time, the trigger article and its data, the author of the utterance);
(2) The form type of SND ("distant, indirect, real-time (on-line) and put off-time (off-line), single-vector – polyvector, monochronic – polychronic");
(3) The function type of SND ("monothematic – polythematic, high contextual – low contextual, action- or polemic-provoking – not action- or polemic-provoking");
(4) The content type of SND ("informative with the sender's point of view, influencing, containing certain verbal means which can produce influence on recipient of the message, provoking with a certain aim to commit specific actions (particularly destructive, realized according to the "stimulus → pragmatic reaction in a form of specific destructive action" scheme), recipient's consciousness manipulation, aimed at a limited target group of users – aimed at an infinite number of users");

The SND parameter system has been applied to the written as well as spoken Russian language [13].

It is the linguistic manifestation of human deprivation that stands in the current research's spotlight. The problem of deprivation study is broader than aggression analysis as aggression is an instance (though a frequent one) of deprivation manifestation. The notion was introduced into the sociology by T. R. Gurr [6]: it "is the discrepancy between what people think they deserve, and what they actually think they can get". Meanwhile R. K. Potapova [12] was first to pronounce its influence on "speech production and speech perception of written and spoken language". Its role in speech behavior lays in the following. The subsystems of human beings (such as biological, physiological, sociopsychological, biomechanical, anthropophonical, cognitological and cogitological subsystems) get into deprivated condition under internal and external factors. They react to external stimuli and these reactions serve as stimuli for the system of verbal, extraverbal and paraverbal behavior. That is, they influence "speech production and speech perception of written and spoken language" [12].

2 Methods

Under our scrutiny was Russian ecology-focused social network discourse. Its main feature is informality which leads to high level of word coinage and colloquial grammar. This makes lexicon- and rules-based method of sentiment classification ineffective. Besides, the macropolylogue (Potapova's term, [11]) discourse structure makes users to write short utterances as if creating a big shared text. Due to this fact unsupervised learning cannot work either. Thus, the empiric material characteristics dictate conditions on our research instrument.

The empiric material was online discussions of Russian ecologic problems. Being a preliminary project, the sample is rather small (about 1,5 thousand utterances, each containing 1-5 sentences). Our goal is to test the innovative classification algorithm while the sample enlargement is the task of future research. Each utterance was described according to Potapova's SND parameters system [11]. Thus was formed the annotated database. The Doc2vec model constructs word- and utterance-vectors which are put into a SGD-classifier. Each word has the same word-vector in different texts of the database while the utterance-vectors are unique. The annotated database is divided into the training and testing subsets at the ratio 4:1. The model builds word-vectors for the training subset and then tries them on the testing subset. In such manner supervised learning is implemented. As one can see, neural networks are not attached to words' dictionary definitions that is why neither homonymy nor polysemy problems arise.

The main difference of the developed product is the ability to maintain not only binary classification (positive – negative) but also several levels of classification. This allows to investigate the influence of Potapova's SND parameters system [11] on the classification accuracy. For example, the deprivation type role is under examination (see further for the term elaboration). On the training stage each utterance gets a label of either positive or negative tonality according to the annotated database. Then the same utterance gets another label of either private or stratified deprivation type. The vectors are built and applied to the testing subset. After that the classification accuracy

is counted and the results are visualized with the help of a confusion matrix. The classification was run this way with every SND system parameter on the same empiric material and the resulting accuracies were compared.

In this research Potapova's SND parameter system is amplified with another content type of SND, namely, deprivation type. We declare that there are two types of deprivation: private and stratifying. **Private deprivation** arises when a person sees the cause of their problem in their own abilities/disabilities, action/inaction etc. **Stratifying deprivation** exists if a person believes that the cause of their problem lies in the society structure or some social circumstances – in other words, because of the place that person has in the society. An example of private deprivation happens when someone can be dissatisfied with their salary because their neighbor has a similar job and earns more. If a person is dissatisfied with their salary because (to the best of their knowledge) everyone of this profession has inadequate salary in this region – then we meet a case of stratifying deprivation. This SND parameter is significant for sentiment classification task because it has a pronounced impact on human verbal behavior. A person suffering stratifying deprivation is very likely not only to share their feelings with others but also to try to organize (or participate in) collective actions aimed at the settlement of the problem. This implies a distinguishing type of discourse with certain vocabulary and syntactic structures.

3 Experiments

As was written above, neural networks can process colossal amounts of texts without misprints and error corrections usual for lexicon- and rules-based methods. No doubt, one need an annotated database to run supervised learning first. But as far as the word vectors are built unannotated data can be processed in disregard for size. The text volume growth stimulates classification accuracy though the dependency is not linear what is demonstrated on Fig. 1.

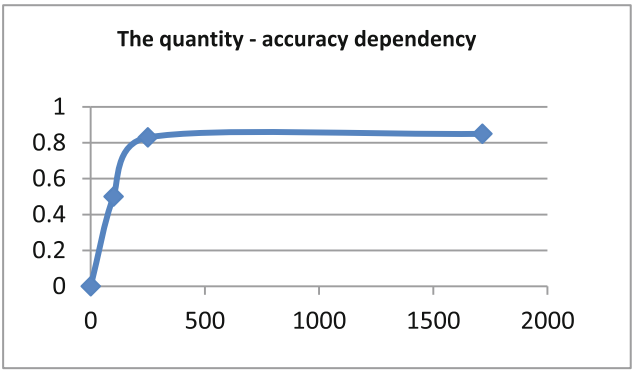

Fig. 1. The influence of utterances number upon classification quality

Two different classifiers were compared: logistic regression and k mean. The best result (98%) was achieved with logistic regression. On the other hand, k mean neighbors could not surpass 70%. Furthermore, the first classifier works better for negative utterances of stratifying deprivation (SD) while the other one – for utterances of private deprivation (PD). As we are more interested in stratifying deprivation the logistic regression classifier was chosen for this research. The confusion matrix was applied for visualization of classifiers efficiency (for illustration see Table 1) because its representation of rightly and wrongly classified cases is the most clear.

Table 1. The comparison of two classifiers

Logistic regression, %					k mean, %		
		Class 1	Class 2	Class 3	Class 1	Class 2	Class 3
Pos.		**25**	75	0	**25**	75	0
Neg.	PD	0	**0**	100	43	**43**	14
	SD	2	1	**98**	16	14	**70**

In the manner described in the previous section the classification was run with every SND system parameter on the same empiric material and the resulting accuracies were compared. As the parameters are grouped in a binary opposition, one half of them is called active parameters while the other – inactive. The analysis shows that the parameters have different influence on sentiment classification accuracy. Figure 2 shows cases of classification which were successful under active parameters. Correspondingly, Fig. 3 shows cases of classification which were effective under inactive parameters. There are two of them with most promising results: monochronic/polychronic (91%), private/stratifying deprivation (92%).

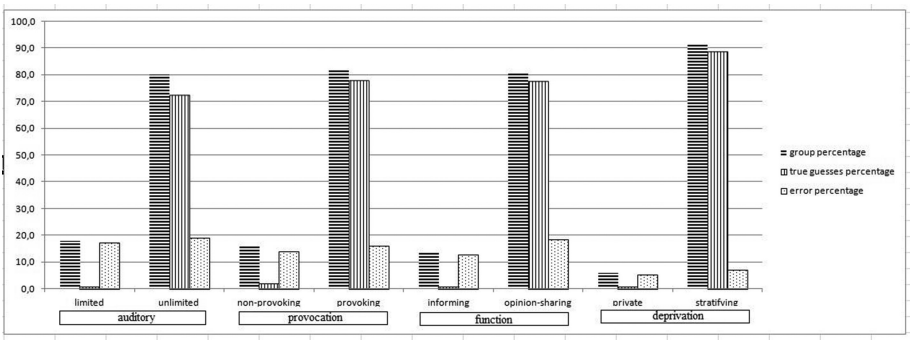

Fig. 2. Successful classification under active SND parameters

The best parameters for negative test recognition are monothematic/polythematic and private/stratifying deprivation. This fact serves as a validity evidence of deprivation type emphasizing. The confusion matrix for the deprivation type classification is shown below (Fig. 4). Table 2 gives precision, recall and F-measure for the best parameters (not the classifier in whole).

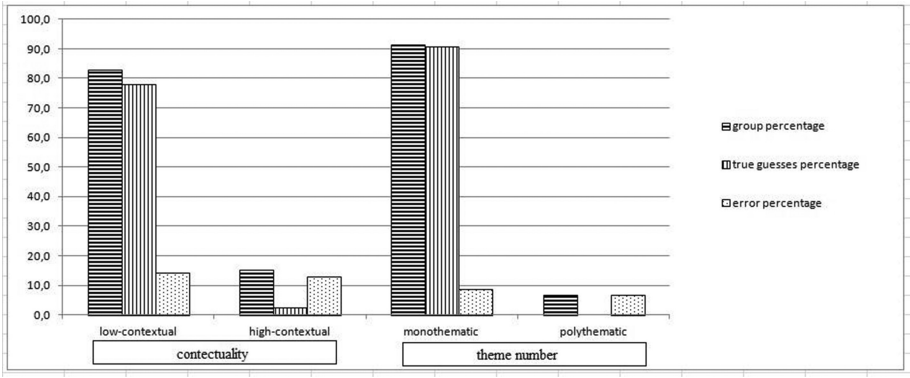

Fig. 3. Successful classification under inactive SND parameters

These results can be compared to the F-measure for negative sentences achieved by Thelwall [18] – 72,8%, in [10] 88,93, in [17] – 86,58. On the other hand, it should be admitted that there are groups that are poorly recognized by our classifier. This makes the classifier whole accuracy not so satisfying (54%).

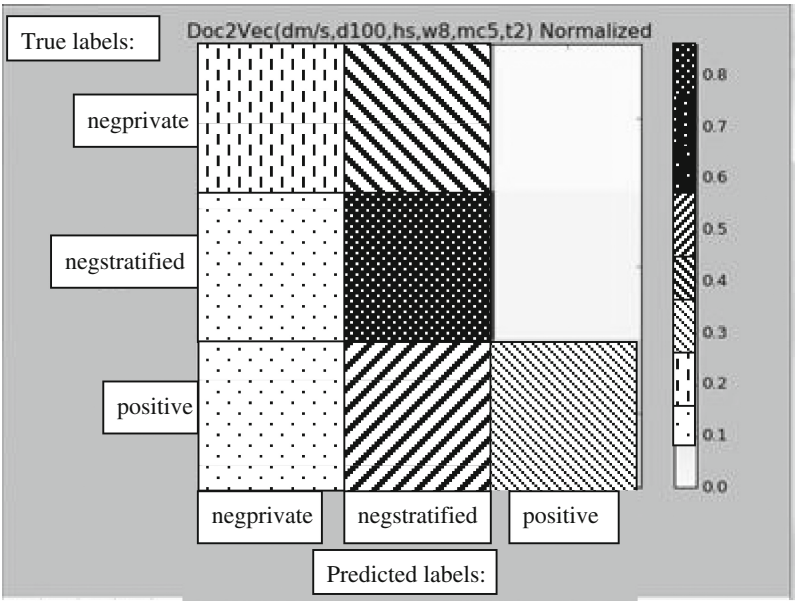

Fig. 4. Confusion matrix for deprivation type classification

Table 2. Precision, Recall and F-measure for optimal parameters

Parameter	Precision	Recall	F1
Stratified deprivation	96,85	92,67	95,96
Monothematic discourse	99,09	91,29	94,83

4 Conclusion

This research has confirmed that neural network models are quite advantageous for language processing in the framework of supervised learning because they do not demand POS-tagging and error correction as preprocessing.

There is no SND parameter which improves results for positive utterances as well as for negative. Some of them are better for positive utterances detection (such as auditory and provocative) and the others – for negative (contextuality, functions, monothematic, deprivation type). However, this experiment showed that word embeddings are suitable for sentiment classification of flexional languages, for example, Russian.

The SND parameter system elaborated by Potapova [11–14] makes a profound step in the social network discourse study. It reflects such key characteristics of SND as the contents (*what* is said), form (*how* it is said) and function (*why* it is said). This enables the SND formalization for automated processing and harnesses the achievements of semantics, stylistics and pragmatics.

The amplification of this system with the deprivation type parameter allows to filter a certain kind of discourse due to the fact that people try to find the collective solution to society problems. This kind of discourse is relevant for opinion mining systems. Thus, a new, more effective and flexible method is suggested for this task instead of lexicon- and rules-based monitoring systems.

5 Prospects of Investigation

As neural networks are not bound to dictionaries they are crossdomain (that was confirmed in [15–21]). Further investigation can be related with various domains, such as discussions of confessional, political, economic problems etc. Besides, the volume of empiric material will be increased. Methodological concepts of the project are described in [11–14].

Acknowledgement. This research is supported by Russian Science Foundation, Project № 14-18-01059. The head of the project – Potapova Rodmonga Kondratjevna.

References

1. Bahtin, M.M.: Aesthetics of Word Creation. Iskusstvo, Moscow (1979). (in Russian)
2. Bai, A., Engelstad, P., Hammer, H., Yazidi, A.: Building sentiment Lexicons applying graph theory on information from three Norwegian thesauruses. In: Amine, A., Bellatreche, L., Elberrichi, Z., Neuhold, E., Wrembel, R. (eds.) Computer Science and Its Application: 5th IFIP TC 5 International Conference, CIIA 2015, pp. 205–216, Saida, Algeria (2015), https://folk.uio.no/paalee/publications/2014-nik.pdf
3. Baroni, M., Dinu, G., Kruszewski, G.: Don't count, predict! A Systematic Comparison of context-counting vs. context-predicting semantic vectors. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pp. 238–247, Baltimore, Maryland, USA (2014)
4. Crystal, D.: Language and the Internet. University of Wales, Bangor (2004)
5. Czerny, M.: Modern methods for sentiment analysis, https://districtdatalabs.silvrback.com/modern-methods-for-sentiment-analysis#disqus_thread
6. Gurr, T.R.: Why Men Rebel. Princeton University Press, Princeton (1970)
7. Isosaevi, J., Lehti, L., Laippala, V., Luotolahti, M.: Linguistic analysis of online conflicts: a case study of flaming in the Smokahontas comment thread on YouTube (2016), http://widerscreen.fi/numerot/2016-1-2/linguistic-anaead-on-youtube/
8. Levy, O., Goldberg, Y.: Dependency-based word embeddings. In: Proceedings of th 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pp. 302–308, Baltimore, Maryland, USA (2014)
9. Mikolov, T., Le, Q.: Distributed representations of sentences and documents. In: Proceedings of the 31st International Conference on Machine Learning, Beijing, China (2014), https://cs.stanford.edu/∼quocle/paragraph_vector.pdf
10. Mohammad, S.M., Kiritchenko, S., Zhu, X.: NRC-Canada: building the state-of-the-art in sentiment analysis of tweets. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), vol. 2, Seventh International Workshop on Semantic Evaluation (SemEval 2013), pp. 321–327, Atlanta, Georgia (2013)
11. Potapova, R.K.: Social-network discourse in the spotlight of cross-disciplinary studies. In: Proceedings of the 2nd International Scientific Conference Discourse as a Social-Network Activity: Priorities and Perspectives, pp. 20–32, MSLU, Moscow (2014) (in Russian)
12. Potapova, R.K.: From deprivation to aggression: verbal and non-verbal social network communication. In: 6th International Scientific Conference on Global Science and Innovation, pp. 129–137. Accent Graphics Communications Publishing Office, Chicago (2015)
13. Potapova, R., Potapov, V.: On individual polyinformativity of speech and voice regarding speakers auditive attribution (forensic phonetic aspect). In: Ronzhin, A., Potapova, R., Németh, G. (eds.) SPECOM 2016. LNCS, vol. 9811, pp. 507–514. Springer, Cham (2016). doi:10.1007/978-3-319-43958-7_61
14. Potapova, R.K.: Deprivation as the basic algorithm of verbal and paraverbal human behavior (on the material of social-network communication). In: Verbal Communication in the Infospace. Lenand, Moscow (2017) (in Russian)
15. Rehurek, R., Sojka, P.: Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, ELRA, pp. 45–50 (2010). https://github.com/RaRe-Technologies/gensim#citing-gensim
16. Sadeghian, A., Sharafat, A.: Bag of Words Meets Bag of Popcorn (2015). https://www.kaggle.com/c/word2vec-nlp-tutorial

17. Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., Qin, B.: Learning sentiment-specific word embedding for twitter sentiment classification. In: Proceeding of the 52th Annual Meeting of the Association for Computational Linguistics, ACL, pp. 1155–1166 (2014). http://anthology.aclweb.org/P/P14/P14-1146.pdf
18. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A.: Sentiment strength detection in short informal text. J. Am. Soc. Inform. Sci. Technol. **61**(12), 2544–2558 (2010)
19. Ustalov, D.A.: Term extraction from Russian texts via graph models. In: Graphs Theory and Applications, pp. 62–69 (2012) (in Russian)
20. Volf, E.M.: The Functional Semantics of assessment. Editorial, Moscow (2002) (in Russian)
21. Word Embeddings for Fun and Profit: Document classification with Gensim, https://github.com/RaRe-Technologies/movie-plots-by-genre/blob/5a2d9157f9bf1bf908794051597b7851333dcfca/ipynb_with_output/Document%20classification%20with%20word%20embeddings%20tutorial%20-%20with%20output.ipynb#L1403