

## NEURAL NETWORK MODEL FOR OVERCOMING TIME GAP OF SENTIMENT CLASSIFICATION

Y. V. Rubtsova

A. P. Ershov Institute of Informatics Systems, Novosibirsk State University,  
630090, Novosibirsk, Russia

---

This paper presents a neural network model to improve sentiment classification in dynamically updated text collections in natural language. As social networks are constantly updated by users there is essential to take into account new jargons, crucial discussed topics while solving classification task. Therefore neural network model for solution this problem is suggested along with supervised machine learning method and unsupervised machine learning method all of them were used for sentiment analysis. It was shown in the paper that the quality of text classification by sentiment is reduced up to 15 % according to F-measure over one and a half year. Therefore the aim of the approach is to minimise the decrease according to F-measure while classifying text collections that are spaced over time. Experiments were made on sufficiently representative text collections, which were briefly described in the paper.

Automatic sentiment classification is rather a topical subject. The great amount of information contained in social networks is represented as text in natural language. Therefore it is requires computational linguistics methods to proceed all this information. Over the about past ten years, a lot of researcher worldwide were involved in the task of automatically extracting and analysing the texts of social media. Moreover as one of the main tasks was considered the problem of sentiment classification of texts in natural language.

Researches and experiments on automatic text classification show that the final results of classification highly depend on the training text set and also the subject are that the training collection corresponds to. Great amount of projects centred on feature engineering and the involvement of additional data, such as external text collections (that do not overlap with the training collection) or sentiment vocabulary. Additional information can reduce the reliance on the training collection and improve classification results. In order to successfully classify texts by sentiment, it is necessary to have tagged by sentiment text collections. Moreover, in order to improve sentiment classification in dynamically updated text collections, it is necessary to have several collections identical by their properties, compiled in different periods of time.

The prepared text collections formed the basis for training and test collections of Twitter posts used to assess the sentiments of tweets towards a given subject at classifier competition at SentiRuEval in 2015 and 2016. It was shown that the collections are complete and sufficiently representative

Previously author shows quite good results of the models builded on feature space for training the classifier based on the training collection and is therefore highly dependent on the quality and completeness of this collection. Described above, there are no semantic relationships between the terms, and the addition of new terms leads to an increase in the dimension of feature vectorspace. Another way to overcome the obsolescence of a lexicon is the use of the distributed word representations as features to train the classifier. So this paper was focused on distributed word representations.

In the basis of this approach is the concept of a distributed word representation and the Skip-gram neural language model. External resources were used here. The distributed word representationspace was built on an untagged collection of tweets gathered in 2013 that was many times larger than the automatically tagged training collection. It is important to mention that the length of the vector space

was only 300 this is the first advantage of the approach. A second advantage of this approach is the classification results: the difference between

Collection of 2013 and of 2015 years is 0.26 % according to F-measure.

In summary, proposed approach can reduce the deterioration of sentiment classification results for collections staggered over time.

**Key words:** natural language processing, sentiment analysis, sentiment classification, machine learning.

## References

1. Loukachevitch N. et al. SentiRuEval: testing object-oriented sentiment analysis systems in Russian // Proceedings of International Conference Dialog. 2015. P. 3–9.
2. Loukachevitch N., Rubtsova Y. Entity-Oriented Sentiment Analysis of Tweets: Results and Problems // Text, Speech, and Dialogue. Springer International Publishing, 2015. P. 551–559.
3. Chetviorkin I., Braslavskiy P., Loukachevitch N. Sentiment Analysis Track at ROMIP 2011 // Computational Linguistics and Intellectual Technologies: Annual International Conf. „Dialogue“, CoLing&InTel. N 11 (18). 2012. P. 739–746.
4. Chetviorkin I., Loukachevich N. 2013. Sentiment analysis track at romip 2012 // In Proceedings of International Conference Dialog. V. 2. 2012. P. 40–50.
5. Amigó E. et al. Overview of RepLab 2012: Evaluating Online Reputation Management Systems // CLEF (Online Working Notes/Labs/Workshop). 2012.
6. Amigó E. et al. Overview of replab 2013: Evaluating online reputation monitoring systems // International Conference of the Cross-Language Evaluation Forum for European Languages. Springer Berlin Heidelberg. 2013. P. 333–352.
7. Loukachevitch, N., Rubtsova, Y. SentiRuEval-2016: Overcoming Time Gap and Data Sparsity in Tweet Sentiment Analysis // In Proceedings of International Conference on Computational Linguistics and Intellectual Technologies Dialog-2016. 2016. P. 375–384.
8. Rosenthal, S., Farra, N., & Nakov, P. SemEval-2017 task 4: Sentiment analysis in Twitter // In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). 2017. P. 502–518.
9. Pang B., Lee L., Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques // Proceedings of the ACL-02 conference on Empirical methods in natural language processing. V. 10. Association for Computational Linguistics, 2002. P. 79–86.
10. Turney P. D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews // Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002. P. 417–424.
11. Wilson T., Wiebe J., Hoffmann P. Recognizing contextual polarity in phrase-level sentiment analysis // Proceedings of the conference on human language technology and empirical methods in natural language processing. Association for Computational Linguistics, 2005. P. 347–354.
12. Jiang L. et al. Target-dependent twitter sentiment classification // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. V. 1. Association for Computational Linguistics, 2011. P. 151–160.
13. Lukashovich N., Rubtsova Yu. Ob“ektno-orientirovannyj analiz tvitov po tonal’nosti: rezul’taty i problemy // Trudy Mezhdunarodnoj konferencii DAMDID/RCDL-2015. Obninsk, 2015. S. 499–507.
14. Klekovkina M. V., Kotel’nikov E. V. Metod avtomaticheskoy klassifikacii tekstov po tonal’nosti, osnovannyj na slovare ehmocional’noj leksiki // Trudy konferencii RCDL. 2012. S. 118–123.
15. Read J. Using emoticons to reduce dependency in machine learning techniques for sentiment classification // In Proceedings of ACL-05, 43rd Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2005.

16. Rubtsova Yu. V. Metod postroeniya i analiza korpusa korotkih tekstov dlya zadachi klassifikaci i otzyvov // *Elektronnyye biblioteki: perspektivnye metody i tekhnologii, ehlektronnyye kollekcii: Trudy XV Vserossijskoj nauchnoj konferencii RCDL'2013*, YArosavl', Rossiya, 14–17 oktyabrya 2013 g. YArosavl': YArGU, 2013. S. 269–275.
17. Rubtsova Yu. V. Razrabotka i issledovanie predmetnonezavisimogo klassifikatora tekstov po tonal'nosti // *Trudy SPIIRAN*. 2014. T. 5. N 36. S. 59–77.
18. Rubtsova Yu. V. Avtomaticheskoe postroenie i analiz korpusa korotkih tekstov (postov mikroblov) dlya zadachi razrabotki i trenirovki tonovogo klassifikatora // *Inzheneriyaznanij i tekhnologii semanticheskogo veba*. 2012. T. 1. S. 109–116.
19. Rubtsova Y. Reducing the Degradation of Sentiment Analysis for Text Collections Spread over a Period of Time // *International Conference on Knowledge Engineering and the Semantic Web*. Springer, Cham, 2017. P. 3–13.
20. Rubtsova Y. Preodolenie degradacii rezul'tatov klassifikacii tekstov po tonal'nosti v kollekcijah, raznesennyh vo vremeni // *Sistemnaya informatika*. 2016. S. 45–68.
21. Titov, I. Modeling Online Reviews with Multi-grain Topic Models // *Proceedings of the 17th International Conference on World Wide Web (WWW'08)*. 2008. P. 111–120.
22. Levy, O. Improving Distributional Similarity with Lessons Learned from Word Em-beddings // *Transactions of the Association for Computational Linguistics*. 2015. P. 211–225.
23. Mikolov, T., Chen, K., Corrado, G., & Dean, J. Efficient estimation of word representations in vector space. arXiv preprint arXiv: 1301.3781. 2013.
24. Mikolov T., Sutskever I., Chen K., Corrado G., and Dean J. Distributed Representations of Words and Phrases and their Compositionality // *In Proceedings of NIPS, 2013*. P. 3111–3119.

# МОДЕЛЬ НЕЙРОННОЙ СЕТИ ДЛЯ ПРЕОДОЛЕНИЯ ДЕГРАДАЦИИ РЕЗУЛЬТАТОВ КЛАССИФИКАЦИИ ТЕКСТОВ ПО ТОНАЛЬНОСТИ

Ю. В. Рубцова

Институт систем информатики СО РАН,  
Новосибирский государственный университет,  
630090, Новосибирск, Россия

УДК 004.912

В данной работе описан алгоритм построения классификатора текстов по тональности, использующий пространство распределенных представлений слов и нейронную языковую модель Skip-gram. Экспериментально показано, что построенная модель классификатора текстов может быть перенесена на коллекции, собранные в другой временной промежуток без потери качества классификации.

**Ключевые слова:** анализ тональности, классификация текстов, машинное обучение.

**Введение.** В сети Интернет содержится огромное количество текстовой информации. Большая часть информации представлена в неструктурированном виде на естественном языке, что усложняет ее обработку. Попытки структурировать информацию и извлечь из нее пользу интересны как коммерческому сектору, так и исследователям. Поэтому за последние несколько лет увеличилось количество исследований, посвященных лингвистическим задачам. Также возросло число программных систем, которые извлекают факты из неструктурированных массивов текстовой информации, классифицируют и кластеризируют информацию. Разрабатываются системы, нацеленные как на анализ самих сообщений в сети, так и на выявление источников распространяемой информации и лидеров мнений. В настоящее время задачей автоматического извлечения и анализа отзывов и мнений из социальных медиа занимается достаточно большое количество ученых и исследователей по всему миру. Одной из главных решаемых задач рассматривается задача классификации текстов по тональности.

Тема автоматической классификации текстов по тональности актуальна в России и за рубежом. Ежегодно проводятся соревнования систем и программных комплексов по автоматической классификации текстов по тональности [1–8].

Классификация текстовых документов по тональности на разных уровнях исследуется в трудах российских и зарубежных ученых. Классификация на уровне всего документа целиком описывается в работах [9, 10]. Чуть позже классификацию на уровне коротких фраз и выражений, а не на уровне абзацев или целых документов, проводили Wilson, Wiebe и Hoffmann [11]. Еще один вид классификации текстов — это классификация относительно заданного объекта. В одном тексте может быть упомянуто несколько сущностей, и автор сообщения может высказывать различные мнения относительно каждой из упомянутых сущностей, поэтому стала актуальной задача анализа тональности по отношению к заданным объектам, упомянутым в тексте [2, 5, 7, 12–14].

Однако, практически все исследования сводятся к построению и оценке классификаторов на текстовых коллекциях, собранных в один временной промежуток, не рассматриваются на текстовых коллекциях, собранных в разные временные интервалы. В этой работе приводятся алгоритм и результаты работы классификатора, который был обучен на текстовой коллекции, собранной в 2013 году, и протестированный на коллекциях, собранных в 2014 и 2015 годах. Предложен метод выделения признаков, основанный на нейронной сети, который показал стабильные результаты на всех текстовых коллекциях, разнесенных во времени.

**1. Текстовые коллекции.** Исследования по автоматической классификации текстов показывают, что результаты классификации, как правило, зависят от обучающей текстовой выборки и предметной области, к которой относится обучающая коллекция. Классификатор может показывать отличные результаты на одной коллекции текстов и совершенно не справиться с такой же задачей на другой коллекции.

Для качественного решения задачи классификации текстов по тональности необходимо иметь размеченные коллекции текстов. Более того, для решения задачи улучшения классификации по тональности в динамически обновляемых коллекциях, необходимо иметь несколько текстовых коллекций, которые были собраны в разные временные промежутки.

Сбор первого корпуса текстов проходил в декабре 2013 года — феврале 2014 года, для краткости будем называть ее коллекцией 2013 года. В соответствии с письменным обозначением эмоций был произведен поиск позитивно и негативно окрашенных сообщений. Таким образом, из коллекции 2013 года сформировано две коллекции: коллекция положительных твитов и коллекция негативных твитов. Нейтральная коллекция была сформирована из сообщений новостных и официальных аккаунтов twitter. С помощью метода [15] и предложенной автором фильтрации [16] из текстов 2013 года была сформирована обучающая коллекция.

Сбор второго корпуса, который состоит из около 10 миллионов коротких сообщений, проходил в июле-августе 2014 года. Третий корпус, состоящий из около 20 млн. сообщений, был собран в июле и ноябре 2015 года.

Из текстов 2014 и 2015 гг. сформированы две тестовые коллекции. Тексты 2014 и 2015 годов подверглись идентичной фильтрации, что и обучающая коллекция 2013 года. Формирование тестовых коллекций по классам тональности происходило аналогично обучающей коллекции. Распределение количества сообщений по классам тональности в коллекциях представлено в табл. 1. Все три коллекции являются предметно независимыми, то есть не относятся ни к какой заранее определенной предметной области.

Ранее в работах [17, 18] автором было показано, что собранные коллекции являются полными и достаточно представительными.

**2. Использование распределенных представлений слов в качестве признаков.** В предыдущих работах [19, 20] было показано, что если пространство признаков для обучения классификатора строится на основе обучающей коллекции, то результаты работы классификатора сильно зависят от качества и полноты этой коллекции. Более того, использование в качестве признаков всех слов, входящих в обучающую коллекцию, приводит к тому, что пространство признаков исчисляется сотнями тысяч.

Одним способом преодоления устаревания лексикона является использование пространства распределенных представлений слов в качестве признаков для обучения классификатора текстов по тональности.

Таблица 1

Распределение сообщений в коллекциях по классам тональности

	Положительные сообщения	Отрицательные сообщения	Нейтральные сообщения
2013 год	114 911	111 922	107 990
2014 год	5 000	5 000	4 293
2015 год	10 000	10 000	9 595

**3. Пространство распределенных представлений слов.** *Распределенное представление слова* (англ. distributed word representation, word embedding) — это  $k$ -мерный вектор признаков  $\mathbf{w} = (w_1, \dots, w_k)$ , где  $w_i \in \mathbb{R}$  — это компоненты вектора [21]. Количество координат  $k$  такого вектора существенно меньше. Обычно это число не превосходит нескольких сотен, соответственно, пространство признаков имеет сравнительно небольшую размерность.

Основная идея векторного распределенного представления слов заключается в нахождении связей между контекстами слов. Идея заключается в том, что находящиеся в похожих контекстах слова, скорее всего, означают или описывают похожие предметы или явления, т. е. являются семантически схожими. Для этого каждый термин представляется в виде вектора из  $k$  координат, в которых закодированы полезные признаки, характеризующие этот термин и позволяющие определять сходство этого термина с похожими терминами в коллекции. Формально представление терминов является задачей максимизации косинусной близости между векторами слов, которые появляются рядом друг с другом в близких контекстах, и минимизация косинусной близости между векторами слов, которые не появляются в близких контекстах. Косинусная мера близости между векторами,  $\cos(\theta)$ , может быть представлена следующим образом (формула 1):

$$\cos(\theta) = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}}, \quad (1)$$

где  $A_i$  и  $B_i$  — координаты векторов  $\mathbf{A}$  и  $\mathbf{B}$  соответственно.

Помимо сокращения размерности вектора признаков, распределенное представление слов учитывает смысл слова в контексте. То есть распределенное представление слов позволяет обобщить, например, „быстрый автомобиль“ на отсутствующее в обучающей выборке „шустрая машина“, что позволяет снизить зависимость от обучающей выборки.

Результаты исследований показывают [22], что нейронная языковая модель Skip-gram превосходит другие модели по качеству получаемых векторных представлений. Поэтому в данной работе используется модель Skip-Gram.

**4. Модель Skip-Gram.** Модель Skip-Gram была предложена Томасом Миколовым с соавторами в 2013 году [23]. На вход модели подается неразмеченный корпус текстов, для каждого слова рассчитывается количество встречаемости этого слова в корпусе. Массив слов сортируется по частоте, редкие слова удаляются. Как правило, можно устанавливать порог встречаемости слова, при котором слово можно считать редким и до которого все редко встречающиеся слова будут удалены. Для того чтобы снизить вычислительную

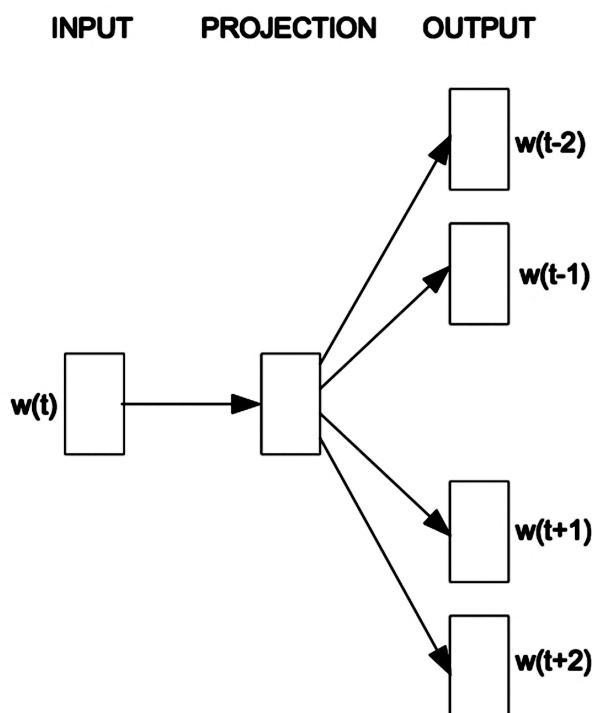


Рис. 1. Архитектура модели Skip-gram

сложность алгоритма, строится дерево Хаффмана (англ. Huffman Binary Tree). Далее алгоритм проходит заранее заданным размером окна по выбранному отрезку текста. Размер окна задается как параметр алгоритма. Под окном подразумевается максимальная дистанция между текущим и предсказываемым словом в предложении. То есть если окно равно трем, то для предложения „Я видел хороший фильм“ применение алгоритма Skip-gram будет проходить внутри блока, состоящего из трех слов: „Я видел хороший“, „видел хороший фильм“. Далее применяется нейронная сеть прямого распространения (англ. Feedforward Neural Network) с много переменной логистической функцией.

Схематически модель Skip-gramm представляется в виде нейронной сети (рис. 1):

Изображенная на рис. 1 нейронная сеть состоит из трех слоев: входной (англ. input), выходной (англ. output) и скрытый (англ. projection). Слово, подаваемое на вход, обозначено  $w(t)$ , в выходном слое слова  $w(t-2)$ ,  $w(t-1)$ ,  $w(t+1)$  и  $w(t+2)$  — слова контекста, которые пытается предсказать нейронная сеть. Иными словами, модель skip-gram предсказывает контекст при заданном слове.

4.1. *Формальная запись модели Skip-gram* представляется следующим: пусть задана текстовая коллекция, состоящая из слов  $w$  и их контекстов  $c$ . Задача модели состоит в том, чтобы подобрать вектор параметров  $\theta$  модели таким образом, чтобы максимизировать условную вероятность всей коллекции  $p(c/w)$  для всех возможных пар контекстов и слов (формула 2):

$$\arg \max_{\theta} \prod_{(w,c) \in D} p(c|w; \theta), \quad (2)$$

где  $D$  — множество всех возможных сочетаний слова и контекста.

Один из способов параметризации модели (формула 2) — это использование логистической функции (англ. Soft-max-function) для определения вероятности  $p(w/c, \theta)$  (формула 3):

$$p(c|w; \theta) = \frac{e^{v_c \cdot v_w}}{\sum_{c' \in C} e^{v_{c'} \cdot v_w}}, \quad (3)$$

где  $v_c$  и  $v_w \in \mathbb{R}^d$  — векторные распределенные представления контекста и слова.  $C$  — это множество всех контекстов. В числителе записана семантическая близость слов контекста ( $v_c$ ) и выбранного целевого слова ( $v_w$ ), в знаменателе — близость всех других контекстов коллекции ( $v_{c'}$ ) и выбранного целевого слова ( $v_w$ ).

Далее целевая функция логарифмируется (формула 2) и подставляется значение вероятностей (формула 3), в результате имеем:

$$\arg \max_{\theta} \sum_{(w,c) \in D} \log p(c|w) = \sum_{(w,c) \in D} (\log e^{v_c \cdot v_w} - \log \sum_{c'} e^{v_{c'} \cdot v_w}). \quad (4)$$

Целевая функция (формула 4) вычислима, однако она представляет собой вычислительно сложную задачу, так как для вычисления  $p(w/c, \theta)$  требуется суммирование  $\sum_{c' \in C} e^{v_{c'} \cdot v_w}$  по всем возможным контекстам термина, которых может быть огромное множество.

С целью оптимизации функции формулы 4 предлагается заменить обычную логистическую функцию (формула 3) на иерархическую софтмакс (англ. Hierarchical Softmax) или использовать негативное сэмплирование (англ. Negative Sampling).

Для каждого слова может существовать большое количество контекстов. Один из способов преодоления этой проблемы — это негативное сэмплирование. Его принцип заключается в том, что для выбранного термина считаются не все возможные контексты, но случайным образом выбирается несколько контекстов ( $v_{c'}$ ). Например, если слово „фильм“ появляется в контексте развлечений, то вектор слова „развлечения“ будет ближе к вектору слова „фильм“, чем векторы некоторых других случайно выбранных слов (таких как ствол, творог или ранец), и не нужно проверять все слова из обучающей коллекции. Такой подход существенно облегчает построение модели.

**5. Использование модели Skip-Gram для снижения зависимости от обучающей коллекции.** Для обучения модели Skip-Gram произвольным образом было выбрано 5 миллионов текстов из первоначальной, не разделенной по классам тональности коллекции 2013 года. Коллекции 2014 и 2015 годов в обучении не участвовали, так как делается предположение, что обученная модель должна быть переносима на более поздние коллекции.

В качестве программной реализации модели Skip-gram был использован программный инструмент Word2Vec [24].

Одной из особенностей Word2Vec является то, что алгоритм разделяет термины между собой, если между ними стоит пробел. Для задачи классификации текстов по тональности важны частицы не и ни, поэтому, чтобы „не + слово“ не было разделено на два различных термина, пробел между частицами не и ни был заменен нижним подчеркиванием (напр. „ни\_разу“, „не\_хотел“).

Каждый текст из обучающей и тестовых коллекций был представлен в виде усредненного вектора входящих в него слов (формула 5):

Таблица 2

Результаты классификации текстов по тональности с использованием векторов слов, полученных при использовании распределенных представлений слов в качестве признаков

	Acc.	Precision	Recall	F-мера
2013	0,7206	0,7250	0,7221	0,7226
2014	0,7756	0,7763	0,7836	0,7787
2015	0,7289	0,7250	0,7317	0,7252

$$d = \frac{\sum w_i}{n}, \quad (5)$$

где  $w_i$  — векторное представление  $i$ -го слова, входящего в исследуемый текст,  $i=(1,\dots,n)$ .  $n$  — число слов из словаря, входящих в исследуемый текст.

Классификатор был обучен на коллекции 2013 года, далее обученная модель классификатора применялась для тестирования на коллекциях 2014 и 2015 годов. Результаты работы классификатора представлены в табл. 2. В качестве метрик оценки качества классификации выбраны стандартные метрики: правильности — accuracy, полноты — recall, точности — precision, гармонического среднего — F-мера.

**Заключение.** При использовании распределенного представления слов в качестве признаков для классификатора текстов по тональности качество классификации на три класса не только не снижается на коллекциях, собранных с разницей в полгода–год, но и держится на уровне лучших значений, зафиксированных в исследованиях [3]. Важно также отметить, что число координат в векторе признаков — ровно 300 (задаваемый параметр), а не несколько сотен тысяч, как в булевой модели для этой же тестовой коллекции.

Данный метод хорошо подходит для применения при наличии внешней достаточно представительной коллекции текстов, схожей по лексике с обучающей и тестовой коллекциями. Однако, как для других нейронных сетей, в этом случае требуется большая обучающая выборка текстов. Метод позволяет получить устойчивые и стабильные результаты.

## Список литературы

1. Loukachevitch N. et al. SentiRuEval: testing object-oriented sentiment analysis systems in Russian // Proceedings of International Conference Dialog. 2015. P. 3–9.
2. Loukachevitch N., Rubtsova Y. Entity-Oriented Sentiment Analysis of Tweets: Results and Problems // Text, Speech, and Dialogue. Springer International Publishing, 2015. P. 551–559.
3. Chetviorkin I., Braslavskiy P., Loukachevitch N. Sentiment Analysis Track at ROMIP 2011 // Computational Linguistics and Intellectual Technologies: Annual International Conf. „Dialogue“, CoLing&InTel. N 11 (18). 2012. P. 739–746.
4. Chetviorkin I., Loukachevich N. 2013. Sentiment analysis track at romip 2012 // In Proceedings of International Conference Dialog. V. 2. 2012. P. 40–50.
5. Amigó E. et al. Overview of RepLab 2012: Evaluating Online Reputation Management Systems // CLEF (Online Working Notes/Labs/Workshop). 2012.
6. Amigó E. et al. Overview of replab 2013: Evaluating online reputation monitoring systems // International Conference of the Cross-Language Evaluation Forum for European Languages. Springer Berlin Heidelberg. 2013. P. 333–352.

7. Loukachevitch, N., Rubtsova, Y. SentiRuEval-2016: Overcoming Time Gap and Data Sparsity in Tweet Sentiment Analysis // In Proceedings of International Conference on Computational Linguistics and Intellectual Technologies Dialog-2016. 2016. P. 375–384.
8. Rosenthal, S., Farra, N., & Nakov, P. SemEval-2017 task 4: Sentiment analysis in Twitter // In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). 2017. P. 502–518.
9. Pang B., Lee L., Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques // Proceedings of the ACL-02 conference on Empirical methods in natural language processing. V. 10. Association for Computational Linguistics, 2002. P. 79–86.
10. Turney P. D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews // Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002. P. 417–424.
11. Wilson T., Wiebe J., Hoffmann P. Recognizing contextual polarity in phrase-level sentiment analysis // Proceedings of the conference on human language technology and empirical methods in natural language processing. Association for Computational Linguistics, 2005. P. 347–354.
12. Jiang L. et al. Target-dependent twitter sentiment classification // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. V. 1. Association for Computational Linguistics, 2011. P. 151–160.
13. Лукашевич Н., Рубцова Ю. Объектно-ориентированный анализ твитов по тональности: результаты и проблемы // Труды Международной конференции DAMDID/RCDL-2015. Обнинск, 2015. С. 499–507.
14. Клековкина М. В., Котельников Е. В. Метод автоматической классификации текстов по тональности, основанный на словаре эмоциональной лексики // Труды конференции RCDL. 2012. С. 118–123.
15. Read J. Using emoticons to reduce dependency in machine learning techniques for sentiment classification // In Proceedings of ACL-05, 43rd Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2005.
16. Рубцова Ю. В. Метод построения и анализа корпуса коротких текстов для задачи классификации отзывов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды XV Всероссийской научной конференции RCDL'2013, Ярославль, Россия, 14–17 октября 2013 г. Ярославль: ЯрГУ, 2013. С. 269–275.
17. Рубцова Ю. В. Разработка и исследование предметно независимого классификатора текстов по тональности // Труды СПИИРАН. 2014. Т. 5. № 36. С. 59–77.
18. Рубцова Ю. Автоматическое построение и анализ корпуса коротких текстов (постов микроблогов) для задачи разработки и тренировки тонового классификатора // Инженерия знаний и технологии семантического веба. 2012. Т. 1. С. 109–116.
19. Rubtsova Y. Reducing the Degradation of Sentiment Analysis for Text Collections Spread over a Period of Time // International Conference on Knowledge Engineering and the Semantic Web. Springer, Cham, 2017. P. 3–13.
20. Рубцова Ю. В. Преодоление деградации результатов классификации текстов по тональности в коллекциях, разнесенных во времени // Системная информатика. 2016. С. 45–68.
21. Titov, I. Modeling Online Reviews with Multi-grain Topic Models // Proceedings of the 17th International Conference on World Wide Web (WWW'08). 2008. P. 111–120.
22. Levy, O. Improving Distributional Similarity with Lessons Learned from Word Em-beddings // Transactions of the Association for Computational Linguistics. 2015. P. 211–225.
23. Mikolov, T., Chen, K., Corrado, G., & Dean, J. Efficient estimation of word representations in vector space. arXiv preprint arXiv: 1301.3781. 2013.
24. Mikolov T., Sutskever I., Chen K., Corrado G., and Dean J. Distributed Representations of Words and Phrases and their Compositionality // In Proceedings of NIPS, 2013. P. 3111–3119.



**Юлия Рубцова.** E-mail: [yu.rubtsova@gmail.com](mailto:yu.rubtsova@gmail.com).

**Юлия Рубцова** закончила механико-математический факультет НГУ в 2007 году. С 2017 года работает младшим научным сотрудником в лаборатории искусственного интеллекта в Институте систем информатики имени А. П. Ершова. Темой классификации текстов занимается с 2012 года. В 2015–2016 годах Ю. Рубцова была соорганизатором международных соревнований по автоматической классификации текстов по тональности в рамках международной конференции „Диалог“. Ю. Рубцова состоит в программном комитете международной конференции „Knowledge engineering and semantic web“.

**Yuliya Rubtsova.** E-mail: [yu.rubtsova@gmail.com](mailto:yu.rubtsova@gmail.com).

**Yuliya Rubtsova** received her bachelor degree in Mathematics from the Mechanics and Mathematics Department of the Novosibirsk State University in 2007. Since 2017 she has been working as a junior researcher in the Laboratory of Artificial Intelligence in the Institute of system Informatics named after A. P. Ershov. The area of researchers interests is text analysis, she engaged this field since 2012. In 2015–2016 Yu. Rubtsova was a co-organizer of international competitions on the automatic sentiment classification the international conference „Dialogue“. Yu. Rubtsova is a program committee member of the international conference „Knowledge engineering and semantic web“.

*Дата поступления — 05.02.2018*