2019. Том 11. Выпуск 3

УДК 811.111
doi 10.17072/2073-6681-2019-3-38-46

ЛЕКСИКО-ГРАММАТИЧЕСКИЕ МАРКЕРЫ ЭМОЦИЙ В КАЧЕСТВЕ ПАРАМЕТРОВ ДЛЯ СЕНТИМЕНТ-АНАЛИЗА РУССКОЯЗЫЧНЫХ ИНТЕРНЕТ-ТЕКСТОВ

Анастасия Владимировна Колмогорова
д. филол. н., профессор, зав. кафедрой романских языков и прикладной лингвистики
Сибирский федеральный университет
660041, Россия, г. Красноярск, Свободный просп., 79. nastiakol@mail.ru
SPIN-код: 4582-4134
ORCID: http://orcid.org/0000-0002-6425-2050
ResearcherID: D-9618-2017

Любовь Александровна Вдовина
магистрант I курса
Сибирский федеральный университет
660041, Россия, г. Красноярск, Свободный просп., 79. lu.vdovina@gmail.com
SPIN-код: 1651-8340
ORCID: http://orcid.org/0000-0001-8129-8408
ResearcherID: F-2690-2019

Статья поступила в редакцию 22.03.2019

Просьба ссылаться на эту статью в русскоязычных источниках следующим образом:
Колмогорова А. В., Вдовина Л. А. Лексико-грамматические маркеры эмоций как параметры для сентимент-анализа русскоязычных интернет-текстов // Вестник Пермского университета. Российская и зарубежная филология. 2019. Т. 11, вып. 3. С. 38–46. doi 10.17072/2073-6681-2019-3-38-46

Please cite this article in English as:
Kolmogorova A. V., Vdovina L. A. Leksiko-grammaticheskie markery emotsiy kak parametry dlya sentiment-analiza russkoyazychnykh internet-tekstov [Lexical and Grammatical Markers of Emotions as Parameters for Sentiment Analysis of Internet Texts in Russian]. *Vestnik Permskogo universiteta. Rossiyskaya i zarubezhnaya filologiya* [Perm University Herald. Russian and Foreign Philology], 2019, vol. 11, issue 3, pp. 38–46. doi 10.17072/2073-6681-2019-3-38-46 (In Russ.)

Рассматриваются промежуточные результаты создания автоматического классификатора русскоязычных интернет-текстов, распределяющего тексты на 8 классов в соответствии с 8 базовыми эмоциями, выделяемыми шведским биологом Гуго Левхеймом: «злость / гнев», «интерес / возбуждение», «удовольствие / радость», «брезгливость / отвращение», «удивление», «стыд / унижение», «страх / ужас», «страдание / тоска». Материалом для формирования обучающей выборки для классификатора послужили анонимные текстовые записи в жанре «интернет-откровения» пользователей в социальной сети «ВКонтакте». В основе работы классификатора лежит алгоритм машинного обучения с использованием метода опорных векторов. На вход классификатору подаются различные лингвистические параметры: например, частотность использования пунктуационных знаков «?», «!», «?!», «…», усилительных наречий, а также коллокации «когда люди говорят»; наличие в обрабатываемом тексте отрицательной частицы «не», конструкций «*такой* + прилагательное», «*так* + наречие», парцелляции, вопросительных слов, частицы «-то», лемм из лексико-семантических полей «смерть», «болезнь», «семья», «одиночество». На выходе получаем на основе учета статистической значимости «входящих» параметров текста его атрибуцию к одному из 8 эмоциональных классов текстов.

Результаты, рассматриваемые в публикации, заключаются в валидации дискриминантных черт текстов различных эмоциональных классов, выделенных исследовательской группой в предыдущих публикациях в качестве параметров для автоматической атрибуции текстов. Рассматривается степень

их влияния на точность работы классификатора. Достигнутая точность классификатора сравнивается с показателями фиктивного классификатора, осуществляющего атрибуцию случайным образом.

В заключение делаются выводы о наиболее эффективных для работы классификатора лингвистических параметрах, оценивается перспективность данного проекта с точки зрения практических задач, а также поднимается вопрос о продолжении исследования для увеличения точности атрибуции.

Ключевые слова: вербальные маркеры; машинное обучение; сентимент-анализ; эмоциональная тональность; ранжированный классификатор; классификация базовых эмоций; компьютерная лингвистика; социальные медиа.

Введение

Развитие современных технологий в значительной степени повлияло на все научные дисциплины, в том числе и на языкознание. Сегодня особый интерес представляют исследования в области компьютерной лингвистики, поскольку они позволяют автоматически обрабатывать и анализировать большой объем языковых данных, в том числе веб-страницы, блоги и социальные медиа.

Публикация посвящена описанию вербальных маркеров, используемых в качестве параметров для автоматической атрибуции русскоязычных интернет-текстов к одному из 8 эмоциональных классов текстов. Работа выполняется в рамках сентимент-анализа. Подобные исследования предполагают выявление тональности текста при помощи методов NLP (обработки естественного языка), статистики, машинного обучения [Pang, Lee 2002; Pang, Lee 2008]. На сегодняшний день в исследовательской и технологической практике преобладают классификаторы, способные автоматически определять либо две тональности текста (позитивную и негативную), либо три, включая нейтральную (проекты см.: [Bollen, Mao, Zeng 2011; Chetviorkin, Loukachevitch 2013]). Атрибуция же текстов к разнообразным классам эмоций, в особенности на русскоязычном материале, пока представляет собой определенную лакуну.

В данной статье предметом обсуждения является статистическая релевантность ряда вербальных маркеров, предварительно полученных путем лингвистического анализа, в ходе их использования в качестве параметров для работы классификатора. Маркеры валидируются на основе выявления зависимостей между подаваемыми на вход модели машинного обучения по прецедентам параметрами и точностью атрибуции текстов в результате работы классификатора.

1. Классификация базовых эмоций

За основу классификации взята трехмерная модель базовых эмоций шведского биолога Г. Левхейма, визуализированная им в виде куба. Модель призвана описать корреляции между уровнем в крови субъекта эмоции специфических гормонов, выполняющих функции монаминных медиаторов, − допамина, норадреналина и серотонина − и эмоциональным состоянием, испытываемым субъектом. Основные эмоции упорядочены в ортогональной системе координат трех основных монаминных осей. Конец каждой из осей репрезентирует низкие и высокие уровни медиаторов, в то время как в каждом из 8 углов находится одна из базовых эмоций, выделяемых Г. Лёвхеймом вслед за С. Томкинском и обозначаемых двучленными терминами для дифференцирования слабой и сильной степеней интенсивности эмоции [Lövheim 2012].

Таким образом, в зависимости от уровней норадреналина, допамина и серотонина выделяются следующие базовые эмоции: «злость / гнев», «интерес / возбуждение», «удовольствие / радость», «брезгливость / отвращение», «удивление», «стыд / унижение», «страх / ужас», «страдание / тоска».

2. Представление данных и выбор алгоритма

Для создания компьютерного классификатора был выбран подход, основанный на машинном обучении с учителем (машинное обучение по прецедентам). При реализации такого подхода построение классификатора происходит на специально размеченном текстовом корпусе (обучающей выборке), в котором текстам приписаны метки, кодирующие важные признаки распознаваемых единиц / текстов. Обучение представляет собой по сути выявление общих закономерностей, присущих текстам, на основе данных обучающей выборки [Юсупова, Богданова, Бойко 2012].

В качестве используемого алгоритма классификации был выбран метод опорных векторов (SVM), поскольку это наиболее быстрый метод нахождения решающих функций [Wiebe, Riloff 2005; Witten, Frank 2005]. В данном методе используется разделяющая полоса максимальной ширины, благодаря чему в дальнейшем осуществляется более уверенная классификация. Программный код реализован на языке программирования Python, поскольку на данный момент он предоставляет наиболее широкий инструментарий для работы с естественным языком [Большакова 2017; VanderPlas 2017].

Обучающая выборка была сформирована нами из материала анонимных текстовых записей пользователей в паблике «Подслушано» социальной сети «ВКонтакте», повествующих об их личном опыте и эмоциональных переживаниях. Выборка состоит из 12123 постов, распределенных по классам следующим образом: «злость / гнев» – 1906, «интерес / возбуждение» – 2063, «удовольствие / радость» – 968, «брезгливость / отвращение» – 790, «удивление» – 2489, «стыд / унижение» – 902, «страх / ужас» – 2508, «страдание / тоска» – 497. Тексты частично были размечены по эмоциям экспертами-носителями русского языка на одной из краудсорсинговых платформ. Однако большая их часть была соотнесена с той или иной эмоцией благодаря соответствующим хештегам (например, #стыдно рассматривался как тематический маркер эмоции «стыд / унижение» и т. д.).

Таким образом, на вход классификатора подаются текстовые данные, которые при помощи функций преобразуются в числовое представление и помогают классификатору провести атрибуцию текстов к одному из 8 классов эмоций. Для улучшения качества модели на вход классификатора подаются также определенные параметры. В качестве последних используются вербальные маркеры – лексические единицы, их сочетания – коллокации, синтаксические конструкции, пунктуационные знаки, предварительно оцененные в ходе экспертного лингвистического анализа (подробнее о методике см.: [Колмогорова, Калинин 2018]) как потенциально значимые для текстов определенного класса.

3. Базовые вербальные маркеры, используемые в качестве параметров

Поскольку в качестве критерия атрибуции служит преобладающая в тексте эмоция, в качестве параметров были выбраны языковые средства, предположительно репрезентирующие конкретные эмоциональные состояния [Болотнов 1981; Шаховский 2009].

Благодаря использованию методов контекстного, семантического, синтаксического анализа с привлечением инструментария корпусного менеджера Sketch Engine для создания функций классификатора были выбраны следующие вербальные маркеры, характерные для отдельных классов эмоций [Колмогорова, Калинин, Маликова 2018; Колмогорова 2018]:

1) использование сочетания слов «*так* / *такой* + прилагательное (полное либо краткое)»;

2) использование сочетания слов «*так* + наречие»;

3) частотность отрицательной частицы *не* или слова *нет*;

4-7) частотность пунктуационных знаков «?», «!», «?!», «...»;

8) наличие парцелляции;

9) наличие вопросительных слов *кто*, *что*, *почему*, *где*, *как*, *куда*, *откуда*, *когда*, *какой*, *чей*, *отчего*, *зачем*, *сколько*, *кого*;

10) наличие лексем из лексико-семантического поля «болезнь»: *врач*, *болезнь*, *боль*, *неизлечимый*, *неизлечимо*, *больница*, *лекарство*, *таблетка*;

11) наличие лексем из лексико-семантического поля «смерть»: *смерть*, *умирать*, *умереть*, *могила*, *похороны*, *кладбище*, *оплакивать*, *оплакать*, *скорбеть*, *хоронить*, *похоронить*, *скончаться*, *захоронить*, *погибнуть*, *погибать*, *кремировать*, *осиротеть*;

12) наличие лексем из лексико-семантического поля «семья»: *жена*, *муж*, *супруга*, *супруг*, *мама*, *мать*, *папа*, *отец*, *брат*, *сестра*, *дочь*, *сын*, *ребенок*, *бабушка*, *дедушка*, *тетя*, *дядя*, *семья*, *прабабушка*, *прадедушка*, *правнук*, *правнучка*, *внук*, *внучка*, *племянник*, *племянница*;

13) наличие лексем из лексико-семантического поля «одиночество»: «одиночество», «одинокий», «одиноко»;

14) частотность конструкции «*когда люди говорят*»;

15) частотность частицы «-то»;

16) частотность наречий меры степени: *очень*, *очень-очень*, *довольно-таки*, *достаточно*, *вполне*, *неслабо*, *настолько*, *сильно*, *невероятно*, *фантастически*, *удивительно*, *особенно*, *чертовски*, *столь*, *прямо-таки*, *необычайно*, *поистине*, *чрезвычайно*, *супер*, *исключительно*, *шибко*, *весьма*, *слишком*, *чересчур*, *чрезмерно*, *крайне*, *изрядно*;

17) наличие местоимений *сам*, *себя*;

18) частотность словоформ глагола *говорить*;

19) наличие слов, указывающих на «чужое слово», так называемых ксенопоказателей: *якобы*, *мол*, *дескать*.

Вышеперечисленные вербальные маркеры были проанализированы на предмет их частотности в текстах различных эмоциональных классов. По результатам тестирования были получены следующие результаты (табл. 1, где представлена доля текстов с ненулевым значением параметра в каждой категории эмоций).

Таблица 1 / Table 1

Доля текстов (%), в которых присутствуют анализируемые вербальные маркеры, в корпусе текстов каждого из классов
The Percentage of Texts Containing Analyzed Verbal Markers among All Texts in Each of the Classes

Маркер / эмоция	Злость /гнев	Брезгливость / отвращение	Тоска / страдание	Удовольствие / радость	Интерес / возбуждение	Страх / ужас	Унижение / стыд	Удивление
1. Так / такой + прил.	5,299	4,9371	4,628	10,124	5,671	4,585	4,878	4,178
2. Так + наречие	5,089	3,165	5,634	5,372	4,653	5,502	14,856	5,022
3. Не, нет	76,337	69,114	72,032	68,492	73,534	81,140	74,723	74,166
4. ?	26,653	6,582	8,652	7,645	10,761	9,41	8,315	13,62
5. !	64,585	26,329	10,261	49,897	38,294	16,228	20,177	22,860
6. ?!	7,293	1,139	0,402	0,826	0,872	1,037	0,887	1,486
7. …	13,746	21,645	22,938	17,975	17,693	25,478	26,607	20,088
8. Парцел.	13,064	11,013	11,268	13,533	14,736	17,464	10,421	10,888
9. Вопрос. слова	93,022	86,709	88,33	85,640	87,203	91,108	90,687	88,871
10. ЛСП болезнь	4,407	3,291	3,420	6,302	5,041	20,734	6,430	6,187
11. ЛСП смерть	1,941	0,633	4,829	3,202	2,036	27,153	6,208	4,299
12. ЛСП семья	22,560	26,835	21,931	40,392	33,786	56,579	43,237	36,119
13. ЛСП одиноч.	0,787	0,126	50,1	2,273	0,921	1,196	0,111	0,643
14. Когда люди…	2,676	0	0,201	0,103	0,194	0	0,111	0,161
15. -то	22,350	28,354	20,724	17,562	25,4	28,35	26,94	25,512
16. Нареч. меры и степени	15,845	16,835	26,559	19,835	24,479	22,328	28,825	21,977
17. Сам, себя	25,498	20,126	28,773	27,996	26,854	24,442	25,277	21,776
18. Говорить	9,811	7,342	5,030	8,988	12,361	13,078	12,528	13,138
19. Ксено-показатели	4,407	2,405	1,006	1,756	4,459	3,628	3,991	3,937

Примечание. В каждой из строк темно-серым цветом выделена ячейка, соответствующая эмоции, для которой наиболее характерен анализируемый маркер, светло-серым – в наименьшей степени.

По результатам анализа можно отметить, что большинство маркеров наиболее ярко выражено в категориях «злость / гнев» и «страх / ужас», в то время как для категорий «брезгливость / отвращение», «удовольствие / радость», «интерес / возбуждение», «удивление» найдено всего по одному наиболее характерному маркеру.

Злость / гнев: наиболее частотное по сравнению с другими классами использование пунктуационных знаков «!», «?» и их комбинаций «?!», а также употребление вопросительных слов и конструкции «*когда люди говорят*». Наиболее редко встречаются многоточие, наречия меры и степени, а также лексемы из лексико-семантического поля (далее – ЛСП) «семья».

Брезгливость / отвращение: наиболее часто встречается частица «*-то*». Наиболее редко встречаются маркеры «*так* + наречие», «?», лексемы из ЛСП «болезнь», «смерть», лексемы *сам*, *себя*.

Тоска / страдание: по сравнению с другими классами здесь реже всего используются «!», «?», лексема *говорить* и слова-ксенопоказатели, в то же время наиболее часты лексемы *сам*, *себя*, а также маркеры из ЛСП «одиночество».

Удовольствие / радость: наиболее распространено сочетание «*так* / *такой* + прилагательное». По сравнению с другими классами здесь реже всего встречаются отрицательная частица *не* и слово *нет*, вопросительные слова, а также слова с частицей «*-то*».

Интерес / возбуждение: чаще, чем в других классах, здесь встречаются вербальные маркеры, выражающие недоверие к правдивости информации при передаче чужого слова: *якобы*, *мол*, *дескать*. Частотность остальных маркеров находится в среднем диапазоне.

Страх / ужас: в данном классе текстов наиболее высока частотность следующих маркеров: парцелляция, *не* или *нет*, а также лексем из ЛСП «болезнь», «смерть», «семья».

Унижение / стыд: здесь чаще, чем в других эмоциональных классах текстов, встречаются многоточие, наречия меры и степени, а также сочетание «*так* + наречие». Напротив, парцелляция и слова из ЛСП «одиночество» встречаются наиболее редко.

Удивление: для данного класса наиболее выраженным оказался маркер лексема *говорить*, а сочетание «*так* / *такой* + прилагательное» встречается реже всего.

4. Результаты работы классификатора

В рамках данного исследования для оценки качества работы классификатора использована мера, комбинирующая точность и полноту (f1-score) классификации для каждого из классов в отдельности, а также значения macro avg (среднее арифметическое всех значений независимо от класса), micro avg (среднее арифметическое, учитывающее количество анализируемых фрагментов для каждого из классов) и weighted avg (взвешенное среднее арифметическое). В табл. 2–3 представлены значения точности работы классификатора при последовательном добавлении на вход параметров, соответствующих вышеуказанным функциям. В каждом последующем столбце на вход подаются все предыдущие параметры, а также один новый.

Для определения эффективности полученные результаты сравниваются с показателями точности фиктивного классификатора, выполняющего атрибуцию в случайном порядке (dummy classifier – DC).

Таблица 2 / Table 2

Оценка точности работы классификатора при добавлении новых параметров (маркеры 1–9)
Estimation of the Classifier's Accuracy When Adding New Parameters (Markers 1–9)

f1-score	**Список параметров, подаваемых на вход**									**DC**
	так + прил.	**так + нареч.**	**не, нет**	**«?»**	**«!»**	**«?!»**	**«...»**	**парцел.**	**вопрос. слова**	
Гнев			0.24	0.29	0.44	0.44	0.44	0.44	0.47	0.13
Отвращение										0.06
Тоска			0.03	0.12	0.14	0.14	0.15	0.13	0.13	0.07
Радость	0.11	0.10	0.07	0.08	0.05	0.05	0.04	0.03	0.06	0.12
Интерес			0.29	0.28	0.15	0.15	0.21	0.20	0.29	0.15
Страх	0.34	0.34	0.21	0.29	0.38	0.38	0.38	0.38	0.35	0.17
Стыд		0.15	0.11	0.12	0.09	0.08	0.10	0.10	0.08	0.09
Удивление	0.01	0.01	0.13	0.01	0.05	0.05	0.05	0.06	0.14	0.16
micro avg	0.20	0.20	0.20	0.21	0.27	0.27	0.27	0.26	0.28	0.13
macro avg	0.06	0.07	0.13	0.15	0.16	0.16	0.17	0.17	0.19	0.12
weighted avg	0.08	0.09	0.17	0.17	0.20	0.20	0.21	0.21	0.24	0.14

Таблица 3 / Table 3

Оценка точности работы классификатора при добавлении новых параметров (маркеры 10–19)
Estimation of the Classifier's Accuracy When Adding New Parameters (Markers 10–19)

f1-score	**Список параметров, подаваемых на вход**										**DC**
	ЛСП болезнь	**ЛСП смерть**	**ЛСП семья**	**ЛСП один.**	**когда люди…**	**-то**	**нареч. меры и степени**	**сам, себя**	**говор.**	**ксено-показатели**	
Гнев	0.47	0.48	0.48	0.48	0.48	0.48	0.47	0.48	0.48	0.48	0.13
Отвращ.						0.04	0.05	0.05	0.06	0.06	0.06
Тоска	0.13	0.12	0.13	0.45	0.45	0.46	0.44	0.41	0.40	0.40	0.07
Радость	0.06	0.06	0.10	0.10	0.10	0.09	0.09	0.07	0.07	0.07	0.12
Интерес	0.31	0.33	0.35	0.34	0.34	0.33	0.33	0.32	0.32	0.32	0.15
Страх	0.37	0.44	0.48	0.48	0.48	0.48	0.48	0.47	0.48	0.48	0.17
Стыд	0.08	0.10	0.06	0.10	0.10	0.11	0.14	0.14	0.14	0.14	0.09
Удивление	0.16	0.19	0.10	0.23	0.23	0.22	0.21	0.19	0.20	0.20	0.16
micro avg	0.28	0.31	0.31	0.35	0.35	0.35	0.35	0.34	0.34	0.34	0.13
macro avg	0.19	0.21	0.21	0.27	0.27	0.28	0.28	0.27	0.27	0.27	0.12
weighted avg	0.24	0.28	0.27	0.31	0.31	0.31	0.31	0.30	0.31	0.31	0.14

По результатам работы классификатора со всеми заявленными функциями видим достаточно высокий уровень точности для эмоций «злость / гнев» – 0.48, «страх / ужас» – 0.48 и «тоска» – 0.40 и низкую точность работы для эмоций «брезгливость / отвращение» – 0.06, и «радость / удовольствие» – 0.07.

Для 6 эмоциональных классов текстов из 8 представленных разработанный классификатор оказался значительно эффективнее фиктивного; для текстов, вербализующих эмоцию «отвращение», точность работы обоих классификаторов одинакова, а для текстов, выражающих эмоцию «радость», фиктивный классификатор произвел более точную атрибуцию.

Влияние параметров, подаваемых классификатору, не всегда является положительным; так, при добавлении параметра частотности «?» эффективность определения эмоции «удивление» упала с 0.13 до 0.01, эмоции «интерес» – с 0.29 до 0.28, т. е. на основании данного параметра классификатор построил некую ложную модель, которая в действительности не свидетельствует о принадлежности текста к данному эмоциональному классу. При этом эффективность для классов «гнев», «тоска», «радость», «страх» и «стыд» в целом возросла на 0.24 пункта. Таким образом, сумма значений точности в общей сложности увеличилась на 0.12 пунктов.

В свою очередь, параметр частотности конструкции «*когда люди говорят*» сам по себе не привел ни к каким изменениям, т. е. все значения эффективности классификации остались на прежнем уровне. Вероятнее всего, данный эффект является результатом низкой встречаемости самой конструкции и данный параметр может оказаться эффективным применительно к корпусам большего объема.

Параметр частотности «?!», напротив, при его добавлении снижает точность для класса «стыд», не влияя при этом на другие классы. Казалось бы, из этого можно сделать вывод, что данный параметр оказывает на классификатор отрицательный эффект, и убрать его, однако при тестировании классификатора со всеми остальными параметрами, за исключением данного, точность для класса «удивление» падает на 0.01, а для класса «стыд» не увеличивается. В данном случае мы видим эффект построения классификатором моделей на основе всех имеющихся параметров и их взаимосвязей. Таким образом, несмотря на кажущуюся «бесполезность» отдельных параметров, в случае их совместного применения получается положительный эффект.

Можно заметить также, что при совместном взаимодействии добавление параметра, характерного для одного из классов и не характерного для другого, не всегда оказывает положительное

или отрицательное воздействие на прирост точности для этих классов соответственно.

Заключение

Валидация параметров на основе ряда вербальных маркеров показала, что частотные вербальные маркеры являются более эффективными и показательными для улучшения точности работы классификатора, чем редко встречающиеся.

Практическая значимость полученных результатов по анализу вербальных маркеров и их использованию для повышения точности работы классификатора заключается в том, что данный набор маркеров может использоваться как основа для дальнейшей разработки эффективной компьютерной программы, определяющей преобладающую в тексте эмоцию. Такая разработка может быть полезной для создания текстов, репрезентирующих то или иное эмоциональное состояние, и улучшения качества машинного перевода путем выбора наиболее подходящего лексического эквивалента с учетом тональности текста.

Классификатор, учитывающий частотность встречаемости выявленных нами вербальных маркеров, работает лучше, чем алгоритм фиктивного классификатора, однако данный набор маркеров не является достаточным для полного решения задачи сентимент-анализа текстов, поскольку общая взвешенная эффективность работы составляет всего 31 %, т. е. в правильную категорию входит лишь каждый третий текст. Тем не менее существенное повышение эффективности работы классификатора представляется возможным за счет выявления новых дискриминантных черт, характерных для каждого из эмоциональных классов текстов, в особенности для классов, точность определения которых является минимальной, поэтому работа над проектом будет продолжаться.

Примечание

[1] Исследование выполнено при поддержке гранта РФФИ (проект «Разработка классификатора русскоязычных интернет-текстов по критерию их тональности на основе модели эмоций “Куб Левхейма”» № 19-012-00205).

Список литературы

Болотнов В. И. Эмоциональность текста в аспектах языковой и неязыковой вариативности: основы эмотивной стилистики текста. Ташкент: Фан, 1981. 116 с.

Большакова Е. И. и др. Автоматическая обработка текстов на естественном языке и анализ данных / Е. И. Большакова, К. В. Воронцов, Н. Э. Ефремова, Э. С. Клышинский, Н. В. Лукашевич, А. С. Сапин. М.: Изд-во НИУ ВШЭ, 2017. 269 с.

Колмогорова А. В. Вербальные маркеры эмоций в контексте решения задач сентимент-анализа // Вопросы когнитивной лингвистики. 2018. № 1. С. 83–93.

Колмогорова А. В., Калинин А. А. Частотность и сочетаемость соматизмов в текстах различной эмоциональной тональности // Компьютерные и интеллектуальные технологии. 2018. Вып. 17. С. 317–330.

Колмогорова А. В., Калинин А. А., Маликова А. В. Лингвистические принципы и методы компьютерной лингвистики для решения задач сентимент-анализа русскоязычных текстов // Актуальные проблемы филологии и педагогический лингвистики. 2018. № 1(29). С. 139–148.

Шаховский В. И. Эмоции как объект исследования в лингвистике // Вопросы психолингвистики. 2009. № 9. С. 29–42.

Юсупова Н. И., Богданова Д. Р., Бойко М. В. Алгоритмическое и программное обеспечение для анализа тональности текстовых сообщений с использованием машинного обучения // Вестник Уфимского государственного авиационного технического университета. 2018. № 16 (6(51)). С. 91–99.

Bollen J., Mao H., Zeng X. Twitter mood predicts the stock market // Journal of Computational Science. 2011. № 1(2). P. 1–8.

Chetviorkin I. I., Loukachevitch N. V. Sentiment analysis track at romip-2012 // Компьютерная лингвистика и интеллектуальные технологии, по материалам конференции «Диалог-2013». 2013. Т. 2. С. 40–50.

Lövheim H. A New Three-dimensional Model for Emotions and Monoamine Neurotransmitters // Medical hypotheses. 2011. № 78. P. 341–348.

Pang B., Lee L. Opinion Mining and Sentiment Analysis // Foundations and Trends in Information Retrieval. 2008. Vol. 2, № 1–2. P. 1–135.

Pang B., Lee L., Vaithyanathan Sh. Thumbs up? Sentiment classification using machine learning techniques // Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). 2002. P. 79–86.

VanderPlas J. Python Data Science Handbook: Essential Tools for Working with Data. Sebastopol: O’Reilly Media, 2017. 548 p.

Wiebe J., Riloff E. Creating subjective and objective sentence classifiers from unannotated texts // Computational Linguistics and Intelligent Text Processing. Berlin: Springer, 2005. 486 p.

Witten I. H., Frank E. Data Mining: Practical Machine Learning Tools and Techniques (Second Edition) // Burlington: Morgan Kaufmann, 2005. P. 56–63.

References

Bolotnov V. I. *Emotsional'nost' teksta v aspektakh yazykovoy i neyazykovoy variativnosti: osnovy emotivnoy stilistiki teksta* [Emotionality of text in the aspects of linguistic and non-linguistic variability: basics of text emotivity]. Tashkent, Fan Publ., 1981. 116 p. (In Russ.)

Bol'shakova E. I., Vorontsov K. V., Efremova N. E., Klyshinskiy E. S., Lukashevich N. V., Sapin A. S. *Avtomaticheskaya obrabotka tekstov na estestvennom yazyke i analiz dannykh* [Automatic natural language text processing and data analysis]. Moscow, HSE Publishing House, 2017. 269 p. (In Russ.)

Kolmogorova A. V. Verbal'nye markery emotsiy v kontekste resheniya zadach sentiment-analiza [Verbal markers of emotions in sentiment analysis researches]. *Voprosy kognitivnoy lingvistiki* [Issues of Cognitive Linguistics], 2018, issue 1, pp. 83–93. (In Russ.)

Kolmogorova A. V., Kalinin A. A. Chastotnost' i sochetaemost' somatizmov v tekstakh razlichnoy emotsional'noy tonal'nosti [Frequency and compatibility of somatisms in texts of different emotional tonality]. *Komp'yuternye i intellektual'nye tekhnologii* [Computer and Intellectual Technologies], 2018, issue 17, pp. 317–330. (In Russ.)

Kolmogorova A. V, Kalinin A. A., Malikova A. V. Lingvisticheskie printsipy i metody komp'yuternoy lingvistiki dlya resheniya zadach sentiment-analiza russkoyazychnykh tekstov [Linguistic principles and computational linguistics methods for the purposes of sentiment analysis of Russian texts]. *Aktual'nye problemy filologii i pedagogicheskiy lingvistiki* [Current Issues in Philology and Pedagogical Linguistics], 2018, issue 1(29), pp. 139–148. (In Russ.)

Shahovskiy V. I. Emotsii kak ob"ekt issledovaniya v lingvistike [Human emotions as an object of the study in linguistics]. *Voprosy psikholingvistiki* [Journal of Psycholinguistics], 2009, issue 9, pp. 29–42. (In Russ.)

Yusupova N. I., Bogdanova D. R., Boyko M. V. Algoritmicheskoe i programmnoe obespechenie dlya analiza tonal'nosti tekstovykh soobshcheniy s ispol'zovaniem mashinnogo obucheniya [Algorithms and software for sentiment analysis of text messages using machine learning]. *Vestnik Ufimskogo gosudarstvennogo aviatsionnogo tekhnicheskogo universiteta* [Herald of Ufa State Aviation Technical University], 2018, issue 16 (6(51)), pp. 91–99. (In Russ.)

Bollen J., Mao H., Zeng X. Twitter mood predicts the stock market. *Journal of Computational Science*, 2011, pt. 1(2), pp. 1–8. (In Eng.)

Chetviorkin I. I., Loukachevitch N. V. Sentiment analysis track at romip-2012. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii, po materialam konferentsii «Dialog-2013»* [Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference 'Dialogue' (2013)], 2013, vol. 2, pp. 40–50. (In Eng.)

Lövheim H. A New Three-dimensional Model for Emotions and Monoamine Neurotransmitters. *Medical Hypotheses,* 2011, pt. 78, pp. 341–348. (In Eng.)

Pang B., Lee L. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2008, vol. 2, issues 1–2, pp. 1–135. (In Eng.)

Pang B., Lee L., Vaithyanathan Sh. Thumbs up? Sentiment classification using machine learning techniques. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2002, pp. 79–86. (In Eng.)

VanderPlas J. *Python data science handbook: Essential tools for working with data*. Sebastopol, O'Reilly Media, 2017. 548 p. (In Eng.)

Wiebe J., Riloff E. *Creating subjective and objective sentence classifiers from unannotated texts. Computational Linguistics and Intelligent Text Processing.* Berlin, Springer, 2005. 486 p. (In Eng.)

Witten I. H., Frank E. *Data mining: Practical machine learning tools and techniques* (Second Edition). Burlington, Morgan Kaufmann, 2005, pp. 56–63. (In Eng.)

LEXICAL AND GRAMMATICAL MARKERS OF EMOTIONS AS PARAMETERS FOR SENTIMENT ANALYSIS OF INTERNET TEXTS IN RUSSIAN

Anastasia V. Kolmogorova
Professor, Head of the Department of Romance Languages and Applied Linguistics
Siberian Federal University
79, Svobodnyy prospekt, Krasnoyarsk, 660041, Russian Federation. nastiakol@mail.ru
SPIN-code: 4582-4134
ORCID: http://orcid.org/0000-0002-6425-2050
ResearcherID: D-9618-2017

Lyubov A. Vdovina
Master's Student
Siberian Federal University
79, Svobodnyy prospekt, Krasnoyarsk, 660041, Russian Federation. lu.vdovina@gmail.com
SPIN-code: 1651-8340
ORCID: http://orcid.org/0000-0001-8129-8408
ResearcherID: F-2690-2019

Submitted 22.03.2019

The article covers intermediate results of the creation of an automatic classifier for Russian-language Internet texts, which distributes those into 8 classes, in accordance with 8 basic emotions proposed by the Swedish biologist Hugo Levheim: 'anger / rage', 'interest / excitement', 'enjoyment / joy', 'contempt / disgust', 'surprise', 'shame / humiliation', 'fear / terror', 'distress / anguish'. The material of the training sample are anonymous texts in the genre of 'Internet revelations' posted by users of the social network VKontakte. The operation of the classifier is based on the machine learning algorithm using the support vector machine method. The input parameters are the frequency of the punctuation marks '?', '!', '?!', '...' used, the presence of the negative particle '*ne'* <*not*> , the use of constructions '*takoi* <*such*> + adjective', '*tak* <*so*> + adverb', the collocation '*kogda lyudi govoryat'* <*when people say*>, the presence of parceling, question words, particle '*-to'*, lexemes from lexical fields 'death', 'disease', 'family', 'loneliness', as well as measure and degree adverbs.

The results considered in the paper consist in the validation of the most characteristic verbal markers of specific emotions as parameters that determine the accuracy of the classifier. We conclude that there is a dependence between the efficiency of parameters and the frequency of correlating verbal markers occurrence within emotional text corpora. The achieved accuracy of the classifier is compared with the results of a dummy classifier that performs attribution randomly.

In conclusion, the paper highlights the most useful verbal markers, assesses the prospects of this project in terms of practical problems, and raises the question of continuing the study to increase the accuracy of attribution.

Key words: verbal markers; machine learning; sentiment analysis; ranked classifier; classification of basic emotions; computational linguistics; social media.