УДК 81’33:811.161.1 DOI: 10.37482/2227-6564-V038

***МАЛИКОВА Алина Вячеславовна**, преподаватель кафедры романских языков и прикладной лингвистики Сибирского федерального университета (г. Красноярск)*[*]

ORCID: https://orcid.org/0000-0002-3438-1839

НЕВЕРБАЛЬНЫЕ МАРКЕРЫ ЭМОЦИЙ ДЛЯ СЕНТИМЕНТ-АНАЛИЗА РУССКОЯЗЫЧНЫХ ИНТЕРНЕТ-ТЕКСТОВ[1]

Статья посвящена описанию начальных этапов проекта по разработке классификатора интернет-текстов на русском языке по критерию эмоциональной тональности. Целью проекта является создание алгоритма сентимент-анализа, атрибутирующего тексты к одному из 8 классов эмоций по модели «Куб Лёвхейма». Необходимыми этапами проекта выступают тщательный отбор языкового материала для обучающей выборки, его независимая экспертная разметка, экспертный лингвистический анализ полученных данных для выделения маркеров эмоций, их валидация инструментами корпусной лингвистики и – при условии подтверждения значимости их показателей в корпусах эмоций – валидация в работе прототипа классификатора. Автор исследует возможность использования невербальных маркеров эмоций в качестве параметров классификации: в результате лингвистического анализа обнаруживаются два потенциальных параметра – фиксация лексем заглавными буквами и цифровой формат числительных. Двойная валидация выявленных маркеров позволяет определить, какой из данных маркеров вызывает положительную динамику точности классификации. Маркер графической передачи числительных приводит к увеличению общей точности работы алгоритма сентимент-анализа на 2 %, а также к приросту точности классификации для классов Интерес на 7 %, классов Удивление и Радость – на 3 %. Отмечается, что тип невербальных маркеров по своей эффективности для сентимент-анализа текстов незначительно отстает от лексико-семантических и пунктуационных вербальных маркеров и находится на одном уровне с синтаксическими вербальными маркерами. Результаты исследования указывают на необходимость рассмотрения данного типа маркеров наряду с вербальными маркерами эмоций и более подробного изучения конкретных маркеров для их использования в качестве параметров классификатора.

Ключевые слова: *русскоязычные интернет-тексты, классификатор текстов, машинное обучение, невербальные маркеры эмоций, сентимент-анализ.*

---

[*]*Адрес:* 660041, г. Красноярск, просп. Свободный, д. 82А; *e-mail:* malikovaav1304@gmail.

Для цитирования: Маликова А.В. Невербальные маркеры эмоций для сентимент-анализа русскоязычных интернет-текстов // Вестн. Сев. (Арктич.) федер. ун-та. Сер.: Гуманит. и соц. науки. 2020. № 4. С. 97–107. DOI: 10.37482/2227-6564-V038

[1]Работа выполнена при поддержке Российского фонда фундаментальных исследований (проект № 19-012-00205 «Разработка классификатора русскоязычных интернет-текстов по критерию их тональности на основе модели эмоций «Куб Лёвхейма»).

С развитием компьютерных технологий стало возможным приближенное к реальному интернет-общение, а созданные в результате такого общения тексты, хранящиеся в виде «больших данных», стали бесценным материалом многочисленных лингвистических исследований. Поскольку эмоции являются неотъемлемой частью речемыслительного процесса каждого человека, множество интернет-текстов сегодня эмоционально окрашены.

Интерес в научном сообществе к проблемам искусственного интеллекта и технологиям компьютерного моделирования способствовал появлению такого направления исследований, как распознавание эмоций в тексте и речи. Разработаны алгоритмы, способные соотносить фрагменты текста с той или иной характеристикой, в т. ч. с вербализованным в них эмоциональным состоянием говорящего.

Целью настоящего исследования стали выделение из текстовых данных маркеров эмоций и проверка валидности их использования в качестве параметров алгоритма классификации интернет-текстов по критерию их эмоциональной тональности.

На основе алгоритмов автоматической обработки текста созданы бинарные классификаторы текстов, определяющие позитивную или негативную тональность [1–4], тернарные классификаторы, дополнительно включающие опцию нейтральной тональности [5], а также незначительное количество многоклассовых классификаторов, относящих текст к конкретному классу эмоций [6, 7].

В связи с растущей потребностью в определении конкретной эмоции в тексте Лабораторией прикладной лингвистики и когнитивных исследований Сибирского федерального университета был инициирован проект, направленный на создание классификатора, способного производить атрибуцию текстов по 8 классам эмоций. В основе подобной дистрибуции – модель «Куб Лёвхейма» [8, с. 342]. Она разработана шведским нейрофизиологом на основе психологической классификации С. Томкинса [9] и валидирована в экспериментах с животными и людьми. Классификация включает 8 классов эмоций: Интерес/Воодушевление, Удовольствие/Радость, Удивление, Грусть/Тоска, Гнев/Ярость, Страх/Ужас, Презрение/Отвращение, Стыд/Унижение.

Актуальность исследования связана с тем, что созданные к данному моменту многоклассовые классификаторы текстов по эмоциональной тональности, как правило, обрабатывают данные на английском языке – для русского языка подобная технология пока находится в стадии разработки: существуют работы по сентимент-анализу русскоязычных текстов, направленные на выявление в тексте ложной информации [10, 11] или психического состояния агрессии [12]. В рамках конкурса по анализу тональности в текстах на русском языке SentiRuEval также решаются задачи сентимент-анализа, в частности анализ тональности отзывов на товары и услуги [2]. На пути к созданию алгоритма классификации текстов по эмоциональной тональности стоит Лаборатория интернет-исследований НИУ ВШЭ, реализовавшая проект по разработке веб-словаря тональностей в 2015 году [13]. Краудсорсинговый ресурс продолжает деятельность по оценке тональности лексем в микротекстах от –2 (сильно отрицательная) до +2 (сильно положительная). Создаются в т. ч. базы данных вербальных и невербальных маркеров конкретных эмоций, например эмоции «страх» [14]. Однако попытка выделить маркеры и ранжировать тексты по 8 эмоциональным классам предпринимается впервые.

Для решения поставленной задачи создания классификатора была выбрана технология *машинного обучения с учителем*, главный принцип которой – «соединить» входы и выходы, т. е. найти закономерность между стимулом и реакцией, между текстом и «принудительно» присваиваемой ему эмоцией. При таком подходе наиболее важно собрать и разметить обучающую выборку, на основе которой затем строится статистический или вероятностный классификатор. Таким образом, основными задачами проекта стали отбор и аннотирование языкового материала, лингвистический анализ

полученных данных для выделения маркеров эмоций, конструирование параметров классификатора на базе маркеров, их валидация в работе прототипа классификатора, тестирование алгоритма на различных данных, а также оптимизация разработанного классификатора.

Материалом исследования послужили фрагменты текстов, отобранные методом сплошной выборки на страницах социальной сети «ВКонтакте» «Подслушано»[2] и «Карамель»[3]. Выбор источника данных обусловлен их жанром – редакторы пабликов позиционируют их как социальные проекты, целью которых является публикация текстов-откровений: «Все, что ты тут найдешь – это реальные женские секреты и откровения, в которых ты узнаешь себя» («Карамель»); «Мы – социальный развлекательный проект, в котором люди каждый день анонимно делятся своими секретами, откровениями и жизненными ситуациями перед огромной аудиторией» («Подслушано»). Как следствие, в таких текстах предполагается присутствие вербализованных эмоциональных переживаний пишущего. Использование сразу нескольких проектов обусловлено следующими причинами: 1) желанием включить тексты различной тематики – материалы «Подслушано» как описания необычных жизненных происшествий и экстремальных внутренних переживаний, а материалы страницы «Карамель» как квинтэссенции более повседневных, но не менее эмоциогенных ситуаций; 2) проверкой гипотезы о валидности эмотиконов в качестве предикторов эмоциональной тональности, присутствующих только во втором источнике; 3) стремлением проследить влияние гендерного фактора – второй паблик позиционирует себя как проект «исключительно для девочек». После первичного анализа языкового материала и обнаружения в нем большого количества эмотиконов исследовательской группой была поставлена гипотеза о наличии других невербальных маркеров эмоций и возможности их выявления с целью последующей их валидации для дальнейшего использования в качестве параметров для классификатора.

Как было отмечено выше, одним из этапов проекта стала тщательная подготовка обучающей выборки. Фрагменты из пабликов «Подслушано» и «Карамель» состоят в среднем из 2–5 предложений. Собранная коллекция данных насчитывает около 26 700 таких микротекстов, называемых *сверхфразовыми единствами* (СФЕ) (12 100 из одного источника и 14 600 – из другого). Данные были разделены на две равные части (по 13 350 СФЕ каждая), составившие обучающую и тестовую выборки. Часть обучающей выборки была авторазмечена редакторами проекта по тегам #Подслушано_успех, #Подслушано_счастье, #Подслушано_странное и #Подслушано_мистика, #Подслушано_одиночество, #Подслушано_БЕСИТ, #Подслушано_страшное, #Подслушано_фууу и #Подслушано_стыдно при публикации фрагментов. Указанные теги были отобраны 32 независимыми экспертами, которые получили на краудсорсинговой платформе «Яндекс.Толока» задание отобрать среди всех тегов, используемых на странице «Подслушано», наиболее подходящие каждому эмоциональному классу.

Для проверки корректности соотнесения тегов с эмоциональной тональностью текста был проведен дополнительный опрос: группе из 40 чел. были предложены в произвольном порядке случайно выбранные тексты из коллекции данных (три из каждого класса эмоций), которые они должны были отнести к одной из 8 эмоций по классификации Лёвхейма. Анализ результатов разметки текстов не показал значимых различий между разметкой независимых экспертов и авторазметкой редакторов. Единственным классом, в котором большинство экспертов выбрало отличную от авторазметки эмоцию, оказался класс Интерес (*табл. 1*). Подобный анализ позволил нам считать авторазметку корпуса

[2]Подслушано. URL: https://ideer.ru/ (дата обращения: 15.01.2020).

[3]Карамель. URL: https://vk.com/caramel6 (дата обращения: 15.01.2020).

Таблица 1

СОПОСТАВЛЕНИЕ АВТОМАТИЧЕСКОЙ РАЗМЕТКИ С РАЗМЕТКОЙ ЭКСПЕРТОВ В ТЕКСТАХ 8 ЭМОЦИОНАЛЬНЫХ КЛАССОВ

Разметка независимых экспертов, %	Авторазметка, %							
	Интерес	**Радость**	**Удивление**	**Тоска**	**Гнев**	**Страх**	**Отвращение**	**Стыд**
Интерес	5,56	6,06	6,67	0,00	8,57	0,00	2,86	3,03
Радость	19,44	**84,80**	0,00	8,33	0,00	0,00	0,00	6,06
Удивление	**36,11**	3,03	**36,67**	2,78	5,71	10,26	22,86	6,06
Тоска	11,11	0,00	20,00	**72,20**	5,71	30,77	0,00	21,21
Гнев	2,78	0,00	6,67	2,78	**57,10**	10,26	2,86	6,06
Страх	5,56	0,00	10,00	5,56	2,86	**33,30**	28,57	9,09
Отвращение	13,89	6,06	16,67	2,78	20,00	7,69	**40,00**	9,09
Стыд	5,56	0,00	3,34	5,56	0,00	7,69	2,86	**39,39**

«Подслушано» достаточно валидной. Вторая часть коллекции данных со страницы «Карамель» была разделена на эмоциональные классы в работе прототипа классификатора, предварительно обученного на размеченном корпусе «Подслушано».

Таким образом, обучающая выборка была отобрана и аннотирована с привлечением независимых экспертов – носителей русского языка. Этот метод – *метод экспертной оценки* – был выбран не случайно: использование эмоциональных реакций экспертов призвано увеличить валидность классификатора, максимально приблизить решения, принимаемые алгоритмом классификации, к решениям, принимаемым человеком.

Следующим этапом обработки материала стал так называемый метод *экспертного лингвистического анализа.* Он нацелен на то, чтобы учесть не только лексическую, но и синтаксическую, морфологическую, пунктуационную и другие особенности текстов при отборе вербальных и невербальных маркеров эмоций. Метод экспертного лингвистического анализа часто сопровождается методами статистического анализа или инструментами корпусной лингвистики. В данном исследовании, в частности, гипотезы, выдвинутые в ходе работы лингвиста с языковым материалом, были оценены благодаря применению корпусного менеджера Sketch Engine[4]. Sketch Engine – один из менеджеров четвертого поколения, особенность которых заключается в способности работать с большими объемами данных, что включает хранение в базе данных сервера коллекции корпусов до 1 млн слов и быстрый поиск, который обеспечивается предварительной индексацией языковых данных. Автоматическая разметка текстов, предоставляемая корпусным менеджером, оказывает значительную помощь при выявлении вербальных и невербальных маркеров эмоций.

Выявленные потенциальные маркеры эмоций были переконструированы в параметры в прикладной лингвистике – *features* (признаки) и использованы в работе прототипа классификатора для оценки их эффективности. Под *параметрами* в данном исследовании понимаются выделяемые в ходе лингвистического анализа вербальные и невербальные маркеры, выступающие признаками рассматриваемого объекта в алгоритме машинного обучения. Валидность таких маркеров отражалась в их влиянии на точность классификации, которая измерялась с применением f1-score – меры,

[4]Sketch Engine. URL: https://www.sketchengine.eu/ (дата обращения: 15.01.2020).

являющейся средним гармоническим значением между точностью (precision) и полнотой охвата (recall).

Выбор исследователями Лаборатории описанной методологии обусловлен стремлением получить наиболее объективные результаты, а также нацелен на двойную валидацию маркеров – с применением инструментария корпусной лингвистики и непосредственно в машинном обучении.

В результате было обнаружено, что параметром для классификатора текстов по эмоциональной тональности могут быть как вербальные, так и невербальные маркеры. *Невербальным маркером* эмоции выступает воспринимаемая визуально, но не являющаяся обязательной частью кода данного естественного языка семиотическая единица или характеристика, самостоятельно и/или в совокупности с другими единицами с определенной частотностью присутствующая в тексте и указывающая на наличие в нем некоторой эмоции.

При анализе вербальной составляющей текстов [15] было обнаружено, что авторы значительного числа текстов, артикулируя свои эмоции, использовали наряду с вербальными невербальные маркеры, в частности эмоджи (подробнее об их валидации см. [16]), или ненормативное графическое представление слова, например фиксировали лексемы заглавными (прописными) буквами:

(1) *Гуляла с другом, и как-то разговор зашел о Луне, звездах. Этот кадр на полном серьезе доказывал мне, что ЛУНА – это просто СОЛНЦЕ, которое ночью не так ярко светит! Как же я тогда знатно офигела! Один аргумент был круче другого: во-первых, Луна – это спутник, но она же не светится, значит, видеть мы ее НЕ МОЖЕМ; во-вторых, Солнце же такое яркое, оно же не может ночью просто взять и скрыться. До сих пор не понимаю, как он защитился с красным дипломом и что у него вместо мозга напихано было* (класс Удивление).

Довольно часто подобный маркер встречался в гневных (2–3) и, напротив, радостных текстах (4–5):

(2) *Бесит, что ВСЕГДА, когда громко слушаю музыку в наушниках, кажется, что меня кто-то зовет! Даже если я дома один* (класс Гнев);

(3) *У меня очень плохое зрение. Когда же я научусь СНАЧАЛА класть очки в поле видимости, а уже ПОТОМ снимать линзы!* (класс Гнев).

В данных СФЕ верхним регистром выделены наречия времени, которые находятся в рематической части предложения-высказывания (2) или в позиции контрастной темы (3) и которые при их устной артикуляции подверглись бы акцентному выделению. В текстах класса Радость такие невербальные графические средства выражения экспрессивности использованы как на отдельной лексеме, так и у целых словосочетаний и законченных фраз:

(4) *В детстве мне несколько раз снился сон, будто я встаю с кровати и хожу по облакам в нашей квартире. В этом сне я была очень счастлива. Прошло 15 лет, этим летом поехала с парнем отдыхать. Поднялись в горы на высоту 2300 м, я вышла из кабинки фуникулера и увидела прямо напротив себя ОБЛАКА. На уровне моего роста. Я в слезах от восторга рассказала ему тот сон. На что он отвел меня дальше от смотровой и... Я ПОТРОГАЛА ОБЛАКО! Сказка, ставшая реальностью. Теперь я самый счастливый человек в мире :)* (класс Радость);

(5) *К своим 27 годам встретила идеального мужчину. Из-за него развелась с мужем спустя месяц знакомства. Уже больше года живу с ним и наслаждаюсь его идеальностью, не знаю, как передать словами, я СЧАСТЛИВА ЧЕРТ ВОЗЬМИ!!! Хочется орать об этом каждый день с балкона нашего 15-го этажа* (класс Радость).

Поиск лексем, зафиксированных заглавными буквами, по Sketch Engine продемонстрировал значительную погрешность в автоматической разметке: менеджер аннотирует

таким образом как строчную фиксацию слов, так и знаки препинания. По этой причине для получения статистических данных о частотности использования ненормативного графического представления слова был использован обычный скрипт на языке программирования Python. Проведенный благодаря созданной функции анализ продемонстрировал отличный от других показатель по указанному маркеру только в классе Гнев – 0,61 % от всех единиц класса при средней доле единиц, фиксированных верхним регистром; среди прочих классов – от 0,1 до 0,3 % (*табл. 2*).

(6) *В детстве на чердаке я нашел какую-то высокую большую банку. Спросил у родителей про нее и не получив внятного ответа, забрал ее себе. Хранил я в этой банке мелочь на протяжении (!) 20 лет! И вот она наполнилась до самого верху, и я решил посчитать, сколько же там. Высыпал все деньги, а там... 68 000 р... Я в шоке сижу и думаю, что мне делать с этими деньгами. На карту не положишь (все ближайшие банкоматы принимают только купюры), а все банки слишком далеко. Пойду в магазине разменяю, что ли...* (класс Интерес).

Результаты корпусного анализа, однако, показали, что в классах Интерес, Радость, Тоска, Страх и Гнев только около половины числительных представлены в цифровой форме. Еще меньше этот показатель в классе Удивление, значительно меньше – в классе Отвращение (35,87 %), а в классе Стыд он, напротив, выше среднего – 67,66 % (*табл. 2*). Это позволило предположить, что цифровой, а не буквенный код для передачи числительных можно использовать как потенциальный параметр при атрибуции текстов к последним двум классам.

Таблица 2

СООТНОШЕНИЕ НЕВЕРБАЛЬНЫХ МАРКЕРОВ В ТЕКСТАХ 8 ЭМОЦИОНАЛЬНЫХ КЛАССОВ

Невербальный маркер	Класс							
	Интерес	Радость	Удивление	Тоска	Гнев	Страх	Отвращение	Стыд
Заглавные буквы,%	0,26	0,25	0,20	0,15	**0,61**	0,17	0,19	0,15
Цифровой формат, % от числ.	48,14	48,49	45,00	48,00	53,67	47,64	**35,87**	**67,66**

С другой стороны, при работе с показателями по морфологическим маркерам эмоциональных классов в корпусном менеджере внимание исследователей привлек следующий невербальный маркер: цифровой формат числительных. В ряде текстов авторы фиксировали числительные в буквенной форме, в то время как в других текстах количество передавалось цифрами.

Изначально возникло предположение, что для удобства восприятия интернет-текста авторы в целом будут стремиться к компактному изложению своих мыслей, избегая использования буквенной формы сложных и составных числительных, а если лексемы, передающие значение большого количества, с большей вероятностью встречаются в текстах, описывающих необычный опыт, удивительные происшествия и т. п., то, следовательно, наибольшее количество числительных в цифровом формате ожидаемо было увидеть в классах Удивление или Интерес, как, например, в следующем СФЕ:

Полученные потенциальные маркеры были в дальнейшем валидированы в работе прототипа классификатора, а именно основанного на модели bag-of-words Наивного байесовского классификатора с использованием TF-IDF взвешивания. В формировании параметров алгоритма участвовали как вербальные, так и невербальные маркеры – параметры подавались на вход в порядке их предполагаемой эффективности: лексико-семантические, синтаксические,

Таблица 3

РЕЗУЛЬТАТЫ ВАЛИДАЦИИ ВЕРБАЛЬНЫХ И НЕВЕРБАЛЬНЫХ МАРКЕРОВ В ТЕКСТАХ 8 ЭМОЦИОНАЛЬНЫХ КЛАССОВ

f1-score	**Параметры, подаваемые на вход**														
	ЛСП Семья	**ЛСП Болезнь**	**ЛСП Смерть**	**ЛСП Одиночество**	**Сам, себя**	**ADV+ADJ/ ADV**	**Парцелляция**	**!**	**?**	**?!**	**Многоточие**	**Заглавные буквы**	**Цифровой формат**	**Сравнительная степень**	**Превосходная степень**
Гнев	0,29	0,30	0,30	0,31	0,30	0,31	0,30	0,41	0,44	0,44	0,44	0,45	0,45	0,46	0,46
Отвращение	–	–	–	–	–	–	–	–	0,17	0,17	0,17	0,17	0,18	0,18	0,18
Тоска	–	–	–	0,54	0,51	0,46	0,47	0,44	0,44	0,44	0,48	0,48	0,49	0,47	0,48
Радость	–	–	–	–	–	–	–	0,01	0,02	0,02	0,04	0,03	**0,06**	0,07	0,07
Интерес	–	–	–	–	–	–	–	–	0,07	0,09	0,13	0,13	**0,20**	0,21	0,20
Страх	0,37	0,40	0,44	0,45	0,45	0,45	0,45	0,48	0,49	0,49	0,49	0,49	0,49	0,49	0,49
Стыд				–	–	0,10	0,11	0,09	0,10	0,10	0,10	0,10	0,07	0,08	0,08
Удивление	0,02	0,05	0,08	0,09	0,08	0,12	0,13	0,33	0,19	0,19	0,17	0,17	**0,20**	0,19	0,19
Micro avg	0,22	0,23	0,24	0,26	0,26	0,26	0,26	0,33	0,31	0,31	0,31	0,31	0,32	0,32	0,32
Macro avg	0,09	0,09	0,10	0,17	0,17	0,18	0,18	0,22	0,24	0,24	0,25	0,25	0,27	0,27	0,27
Weighted avg	0,13	0,14	0,16	0,18	0,18	0,19	0,20	0,26	0,26	0,26	0,27	0,27	**0,29**	0,29	0,29

пунктуационные, невербальные, морфологические (*табл. 3*).

Достаточно неожиданно в результате валидации была выявлена положительная динамика общей точности классификации с включением маркера «цифровой формат числительных»: рост средневзвешенного f1-score составил 2 %. Данный маркер оказал значительное влияние на классы Интерес (+7 %), Удивление и Радость (+3 %) в отличие от маркера «заглавные буквы», который не только никак не повлиял на средневзвешенный показатель точности по всем классам эмоций вместе, но и не повлиял положительно ни на один из классов в отдельности: изменение точности на 1 % (положительно в случае с классом Гнев и негативно – в классе Радость) не свидетельствует о валидности маркера.

Что касается эффективности обоих невербальных маркеров, отмечена одна точка резкого роста f1-score (+7 % к классу Интерес), прирост среднего взвешенного на 2 % и сумма приростов по классам – на 0,12, что совпадает с данными об эффективности синтаксических маркеров и позволяет нам включить невербальные маркеры в типологию эффективных маркеров эмоций.

Таким образом, в ходе реализации стадии экспертного лингвистического анализа проекта по разработке многоклассового классификатора русскоязычных интернет-текстов по эмоциональной тональности были выявлены не только потенциальные вербальные маркеры, но и невербальные, а именно фиксация лексем заглавными буквами и фиксация числительных цифрами. Валидация данных маркеров в работе прототипа классификатора показала эффективность маркера, связанного с форматом числительных, однако опровергла значимость заглавных букв в тексте для передачи эмоции.

В контексте типологизации вербальных и невербальных маркеров по их эффективности для определения эмоциональной тональности текста невербальные маркеры наряду с синтаксическими составляют группу среднеэффективных, отставая по своим показателям от лексико-семантических и пунктуационных маркеров. Этот вывод позволяет нам рассматривать вербальные и невербальные маркеры как равноправных кандидатов в параметры классификатора русскоязычных интернет-текстов по эмоциональной тональности.

Список литературы

1. *Hogenboom A., Frasincar F., de Jong F., Kaymak U.* Polarity Classification Using Structure-Based Vector Representations of Text // Decis. Support Sys. 2015. № 74. P. 46–56.

2. *Loukachevitch N.V., Blinov P.D., Kotelnikov E.V., Rubtsova Y.V., Ivanov V.V., Tutubalina E.* SentiRuEval: Testing Object-Oriented Sentiment Analysis Systems in Russian // Computational Linguistics and Intellectual Technologies: Proceedings of the Annual International Conference “Dialogue”, Moscow, 27–30 May 2015. Moscow: RSUH, 2015. Iss. 14 / ed. by V.P. Selegey. P. 3–15.

3. *Vasilyev V.G., Denisenko A.A., Soloviev D.A.* Aspect Extraction and Twitter Sentiment Classification by Fragment Rules // Computational Linguistics and Intellectual Technologies: Proceedings of the Annual International Conference “Dialogue”, Moscow, 27–30 May 2015. Moscow: RSUH, 2015. Iss. 14 / ed. by V.P. Selegey. P. 76–88.

4. *Karpov I.A., Kozhevnikov M.V., Kazorin V.I., Nemov N.R.* Entity Based Sentiment Analysis Using Syntax Patterns and Convolutional Neural Network // Computational Linguistics and Intellectual Technologies: Proceedings of the Annual International Conference “Dialogue”, Moscow, 1–4 July 2016. Moscow: RSUH, 2016. Iss. 15 / ed. by V.P. Selegey. P. 225–236.

5. *Lucas G.M., Gratch J., Malandrakis N., Szablowski E., Fessler E., Nichols J.* GOAALLL!: Using Sentiment in the World Cup to Explore Theories of Emotion // Image Vis. Comput. 2017. Vol. 65. P. 58–65.

6. *Staiano J., Guerini M.* DepecheMood: A Lexicon for Emotion Analysis from Crowd-Annotated News // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL), Baltimore, USA, 22–27 June 2014 / ed. by K. Toutanova, H. Wu. N. Y.: Association for Computational Linguistics, 2014. P. 427–433.

7. *Alm C.O., Roth D., Sproat R.* Emotions from Text: Machine Learning for Text-Based Emotion Prediction // Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Vancouver, Canada, 6–8 October 2005 / ed. by R.J. Mooney. Stroudsburg: Association for Computational Linguistics, 2005. P. 579–586.

8. *Lövheim H.* A New Three-Dimensional Model for Emotions and Monoamine Neurotransmitters // Medical Hypotheses. 2012. Vol. 78, № 2. P. 341–348.

9. *Tomkins S.S.* Affect Theory // Emotion in the Human Face / ed. by P. Ekman. Cambridge: Cambridge University Press, 1982. P. 353–395.

10. *Potapova R., Lykova O.* Verbal Representation of Lies in Russian and Anglo-American Cultures // Procedia – Soc. Behav. Sci. 2016. Vol. 236. P. 114–118.

11. *Pisarevskaya D.* Rhetorical Structure Theory as a Feature for Deception Detection in News Reports in the Russian Language // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue", Moscow, 31 May – 3 June 2017. Moscow: RSUH, 2017. Vol. 1, iss. 16 / ed. by V.P. Selegey. P. 191–200.

12. *Potapova R., Komalova L.* Multimodal Perception of Aggressive Behavior // Lecture Notes in Computer Science. Cham: Springer, 2016. Vol. 9811. P. 499–506.

13. Koltsova O.Y., Alexeeva S.V., Kolcov S.N. An Opinion Word Lexicon and a Training Dataset for Russian Sentiment Analysis of Social Media // Computational Linguistics and Intellectual Technologies: Proceedings of the Annual International Conference "Dialogue", Moscow, 1–4 July 2016. Moscow: RSUH, 2016. Iss. 15 / ed. by V.P. Selegey. P. 259–268.

14. *Колосов Я.В.* Лингвистические корреляты эмоционального состояния «страх» в русской и английской речи: формирование базы данных: дис. … канд. филол. наук. М., 2004. 214 с.

15. *Колмогорова А.В.* Вербальные маркеры эмоций в контексте решения задач сентимент-анализа // Вопр. когнит. лингвистики. 2018. № 1(54). С. 83–93. DOI: 10.20916/1812-3228-2018-1-83-93

16. *Kolmogorova A., Kalinin A., Malikova A.* Emojis as Predictors in Lövheim Cube Backed Multi-Class Sentiment Analysis: Can We Really Trust Them? // 6th SWS International Scientific Conference on Arts and Humanities ISCAH 2019. Sofia, 2019. Vol. 6. P. 645–652.

References

1. Hogenboom A., Frasincar F., de Jong F., Kaymak U. Polarity Classification Using Structure-Based Vector Representations of Text. *Decis. Support Syst.*, 2015, no. 74, pp. 46–56.

2. Loukachevitch N.V., Blinov P.D., Kotelnikov E.V., Rubtsova Y.V., Ivanov V.V., Tutubalina E. SentiRuEval: Testing Object-Oriented Sentiment Analysis Systems in Russian. Selegey V.P. (ed.). *Computational Linguistics and Intellectual Technologies*. Moscow, 2015. Iss. 14, pp. 3–15.

3. Vasilyev V.G., Denisenko A.A., Soloviev D.A. Aspect Extraction and Twitter Sentiment Classification by Fragment Rules. Selegey V.P. (ed.). *Computational Linguistics and Intellectual Technologies*. Moscow, 2015. Iss. 14, pp. 76–88.

4. Karpov I.A., Kozhevnikov M.V., Kazorin V.I., Nemov N.R. Entity Based Sentiment Analysis Using Syntax Patterns and Convolutional Neural Network. Selegey V.P. (ed.). *Computational Linguistics and Intellectual Technologies*. Moscow, 2016. Iss. 15, pp. 225–236.

5. Lucas G.M., Gratch J., Malandrakis N., Szablowski E., Fessler E., Nichols J. GOAALLL!: Using Sentiment in the World Cup to Explore Theories of Emotion. *Image Vis. Comput.*, 2017, no. 65, pp. 58–65.

6. Staiano J., Guerini M. DepecheMood: A Lexicon for Emotion Analysis from Crowd-Annotated News. Toutanova K., Wu H. (eds.). *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*. New York, 2014, pp. 427–433.

7. Alm C.O., Roth D., Sproat R. Emotions from Text: Machine Learning for Text-Based Emotion Prediction. Mooney R.J. (ed.). *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, 2005, pp. 579–586.

8. Lövheim H. A New Three-Dimensional Model for Emotions and Monoamine Neurotransmitters. *Med. Hypotheses*, 2012, vol. 78, no. 2, pp. 341–348.

9. Tomkins S.S. Affect Theory. Ekman P. (ed.). *Emotion in the Human Face*. Cambridge, 1982, pp. 353–395.

10. Potapova R., Lykova O. Verbal Representation of Lies in Russian and Anglo-American Cultures. *Procedia – Soc. Behav. Sci.*, 2016, vol. 236, pp. 114–118.

11. Pisarevskaya D. Rhetorical Structure Theory as a Feature for Deception Detection in News Reports in the Russian Language. Selegey V.P. (ed.). *Computational Linguistics and Intellectual Technologies*. Moscow, 2017. Iss. 16. Vol. 1, pp. 191–200.

12. Potapova R., Komalova L. Multimodal Perception of Aggressive Behavior. Ronzhin A., Potapova R., Németh G. (eds.). *Speech and Computer. SPECOM 2016*. Cham, 2016. Vol. 9811, pp. 499–506.

13. Koltsova O.Y., Alexeeva S.V., Kolcov S.N. An Opinion Word Lexicon and a Training Dataset for Russian Sentiment Analysis of Social Media. Selegey V.P. (ed.). *Computational Linguistics and Intellectual Technologies*. Moscow, 2016. Iss. 15, pp. 259–268.

14. Kolosov Ya.V. *Lingvisticheskie korrelyaty emotsional'nogo sostoyaniya "strakh" v russkoy i angliyskoy rechi: formirovanie bazy dannykh* [Linguistic Correlates of the Emotional State of Fear in Russian and English Speech: Database Formation: Diss.]. Moscow, 2004. 214 p.

15. Kolmogorova A.V. Verbal'nye markery emotsiy v kontekste resheniya zadach sentiment-analiza [Verbal Markers of Emotions in Sentiment Analysis Researches]. *Voprosy kognitivnoy lingvistiki*, 2018, no. 1, pp. 83–93. DOI: 10.20916/1812-3228-2018-1-83-93

16. Kolmogorova A., Kalinin A., Malikova A. Emojis as Predictors in Lövheim Cube Backed Multi-Class Sentiment Analysis: Can We Really Trust Them? *6th SWS International Scientific Conference on Arts and Humanities ISCAH 2019*. Sofia, 2019. Vol. 6, pp. 645–652.

DOI : 10.37482/2227-6564-V038

Alina V. Malikova
Siberian Federal University;
prosp. Svobodnyy 82A, Krasnoyarsk, 660041, Russian Federation;
ORCID: https://orcid.org/0000-0002-3438-1839 *e-mail:* malikovaav1304@gmail.com

NON-VERBAL EMOTION MARKERS IN THE SENTIMENT ANALYSIS OF RUSSIAN-LANGUAGE INTERNET TEXTS

This article describes the initial stages of the project aiming to design a classifier of Internet texts in Russian by emotional tonality. To create a sentiment analysis algorithm that attributes texts to one of the 8 basic emotions according to Lövheim's cube model, it is necessary to do the following: carefully select the language material for the training sample; label its tonality with the assistance of an independent expert; carry out an expert linguistic analysis of the data in order to determine the emotion markers; validate the markers using corpus analysis tools; and, subject to their quantitative significance in the emotion corpora, validate them in the work of the prototype classifier. The author examined the possibility of using non-verbal emotion markers as classification parameters. The linguistic analysis revealed two potential parameters: lexemes written in capital letters and numbers written in figures. Double validation of the markers identified allows us to determine which of them

For citation: Malikova A.V. Non-Verbal Emotion Markers in the Sentiment Analysis of Russian-Language Internet Texts. *Vestnik Severnogo (Arkticheskogo) federal'nogo universiteta. Ser.: Gumanitarnye i sotsial'nye nauki*, 2020, no. 4, pp. 97–107. DOI: 10.37482/2227-6564-V038

improves the accuracy of classification. The marker of writing numbers in figures leads to a 2 % increase in the overall accuracy of the sentiment analysis algorithm, as well as to a 7 % increase in the classification accuracy for the basic emotion of interest/excitement, and a 3 % increase for the basic emotions of surprise/startle and enjoyment/joy. It is noted that non-verbal markers are slightly less effective for the sentiment analysis of texts than lexical, semantic or punctuation markers, but are as much effective as syntactic markers. The results indicate the need to consider this type of markers along with verbal markers of emotions and explore in more detail concrete non-verbal markers as potential classifier parameters.

Keywords: *Russian-language Internet texts, text classifier, machine learning, non-verbal emotion markers, sentiment analysis.*

Поступила: 23.01.2020
Принята: 20.04.2020

Received: 23 January 2020
Accepted: 20 April 2020