УЧЕНЫЕ ЗАПИСКИ КАЗАНСКОГО УНИВЕРСИТЕТА.
СЕРИЯ ГУМАНИТАРНЫЕ НАУКИ

2021, Т. 163, кн. 1
С. 65–80

ISSN 2541-7738 (Print)
ISSN 2500-2171 (Online)

ОРИГИНАЛЬНАЯ СТАТЬЯ

УДК 81'27

doi: 10.26907/2541-7738.2021.1.65-80

СЕНТИМЕНТ-АНАЛИЗ ЧИТАТЕЛЬСКОГО КОММЕНТАРИЯ: АВТОМАТИЗИРОВАННАЯ VS РУЧНАЯ ОБРАБОТКА ТЕКСТА

Г.К. Гималетдинова, Э.Х. Довтаева

Казанский (Приволжский) федеральный университет, г. Казань, 420008, Россия

Аннотация

Статья посвящена изучению речевых и структурных особенностей читательского комментария как жанра интернет-коммуникации. Посредством автоматизированного анализа тональности текста (сентимент-анализа) с использованием прикладного программного интерфейса ParallelDots API была выявлена эмотивная составляющая англо- и русскоязычных читательских комментариев ($N = 3000$). Полученные результаты были верифицированы методом текстового лингвистического анализа ручным способом. В качестве экспертов выступили специалисты в области филологии английского и русского языков ($N = 6$), а также студенты-филологи, носители русского языка, владеющие английским языком как иностранным на уровне C1 ($N = 4$). Сопоставление автоматизированной и ручной обработки текста позволило выделить ряд факторов, снижающих уровень достоверности результатов автоматизированной обработки читательских комментариев. Трудности, препятствующие объективному определению тональности программой, были обнаружены в комментариях носителей как английского, так и русского языка, что указывает на схожесть структурного построения англоязычных и русскоязычных читательских комментариев на лексическом и синтаксическом уровнях. Сделан вывод о целесообразности синтезированного использования автоматизированного и ручного анализа текста с целью получения более детальных и объективных данных.

Ключевые слова: тональность текста, сентимент-анализ, интерактивная газетная статья, читательский комментарий, эмотивность

Введение

Интернет-пространство в эпоху глобализации существенным образом влияет на формы социальной адаптации личности к стремительно меняющимся условиям информационной реальности. Динамика общественных изменений, с одной стороны, и технологии – с другой, способствует более открытой и доступной коммуникации. Интернет-коммуникация по своей сути является глобальным полилогом, в котором каждый участник имеет широкие возможности для самовыражения, обсуждения актуальных событий местного и мирового масштаба. Развитие информационных технологий влечет за собой возникновение новых форм интернет-коммуникации, однако в большинстве случаев характерными ее особенностями остаются лаконичность и стандартизация, проявляющиеся параллельно с субъективностью и экспрессивностью (см. [1]).

В настоящем исследовании предметом специального рассмотрения является читательский комментарий (англ. *reader comment*) как жанр медийной коммуника-

ции, неразрывно связанный с понятием интерактивной газетной статьи (англ. *participatory news article*) [2, с. 144]. Важнейшей особенностью последней является наличие отправной точки – новостной статьи, поднимающей ту или иную общественно значимую проблему, тогда как комментарий читателя рассматривается как ответная реакция. Таким образом, диалогичность автора новостной статьи и интерактивность бесконечного множества читателей-комментаторов создает феномен интерактивной газетной статьи [2, с. 145]. Особенностью читательского комментария в структуре интерактивной газетной статьи является трансформация эмоции (психологическая категория) в эмотивность (языковая категория). В эмотиологии выражение эмоций предстает как «непосредственное речевое проявление, производимое при помощи специфических единиц-эмотивов, семантика которых индуцирует эмоциональное отношение» [3, с. 96].

Лингвистический анализ читательского комментария возможен при использовании специально разработанных инструментов (например, компьютерных программ) либо ручным способом. Определение эмоциональной составляющей затрудняется прежде всего сложностью «определения самого эмотивного пространства, количества и состава его измерений» [4, с. 511], а также обилием контекстно обусловленной оценочной лексики.

Компьютерная лингвистика применяет метод анализа тональности текста (англ. *sentiment analysis*), где тональность (сентимент) подразумевает выраженную в тексте эмоциональную оценку [5]. Тональность всего текста «определяется лексической тональностью составляющих его единиц и правилами их сочетания» [4, с. 511]. Основными преимуществами автоматизированного сентимент-анализа читательских комментариев считаются высокая скорость и отсутствие субъективности при измерении эмотивности.

Статья является частью общего исследования, посвященного изучению англоязычных и русскоязычных читательских комментариев в сопоставительном аспекте [2; 6–7]. В данной работе представлены результаты сопоставления текстов комментариев с применением метода сентимент-анализа. Цель настоящей работы – определить тональность англоязычных и русскоязычных читательских комментариев двумя способами: машинным (автоматизированным, с применением метода сентимент-анализа) и ручным (экспертным) – и выявить недостатки автоматизированного сентимент-анализа. Наиболее значимым достижением является систематизация причин, затрудняющих работу программы для корректного определения тональной оценки, на основе сопоставления результатов автоматизированного и ручного способов сентимент-анализа текстов англоязычных и русскоязычных читательских комментариев.

Обзор литературы по проблеме

В современных исследованиях, посвященных вопросам виртуального жанроведения (см. [8–12]), отмечается, что интернет-комментарий следует рассматривать как отдельный жанр интернет-коммуникации. Представляя собой метатексты – реакции на единый исходный текст, интернет-комментарии определяются как вторичные тексты, порождению которых способствуют восприятие и интерпретация информации пользователем [13–15]. Данный вид комментария является быстрой реакцией на происходящие события и отражает индивидуальное

субъективно-оценочное видение мира, выражаемое автором текста с помощью лингвистических средств [16]. Отсутствие визуального контакта с участниками виртуального полилога создает условия для свободного общения, зачастую вне социальных и культурных границ (см. [1; 17]). Интернет-комментарий характеризуется эмоциональной насыщенностью текста, а образ действительности, формирующийся у пользователя и выражающийся в комментарии, демонстрирует культурно и социально обусловленную эпистемическую модальность интернет-комментария, что способствует изучению лингвокультурной среды [18].

Оценочная составляющая комментария определяется как основополагающий аспект при коммуникации, указывающий на ценность объекта с позиции комментирующего [19]. Отмечается, что оценка может быть выражена меньшими, чем слово, элементами [19; 20]. Эмоциональность и оценочность обусловлены актуализацией прагматической функции (функции воздействия), при этом доказано преобладание в тексте комментариев отрицательных оценочных номинаций [17; 21].

Анализ эмоциональности в текстах интернет-комментариев может осуществляться различными способами, в частности ручным (выявление эмоционально-оценочных единиц с последующим анализом их употребления в заданном отрезке коммуникации), а также автоматизированным (анализ тональности текста в результате его глубокого структурного разбора компьютерной программой). Оба способа традиционно относятся к методам контент-анализа, который получил название «анализ тональности текста» (сентимент-анализ) [22]. Автоматизированный поход базируется на комбинации методов, среди которых выделяются метод, основанный на использовании словарей; метод, основанный на применении правил; машинное обучение с учителем / без учителя; а также гибридный метод. Итогом автоматического анализа служат двухчастотные (позитив/негатив) или трехчастотные (позитив/негатив/нейтральность) классификаторы [22].

В ряде исследований, изучающих теоретические вопросы сентимент-анализа текстов различных областей в пределах сети Интернет, отмечается, что несоответствия между автоматическим и ручным анализом текста могут быть объяснены рядом факторов: наличием в текстах имплицитной оценки, эмотиконов, многозначностью лексических единиц и др. [4; 23–25]. Практическое обоснование влияния подобных факторов на тональность можно обнаружить в исследованиях, посвященных анализу текстов различной жанровой направленности [26; 27]. Несмотря на ряд преимуществ автоматической обработки текста, ученые отмечают возникающие в ходе исследований несоответствия между программной и ручной оценкой. Так, к примеру, в исследовании П.П. Зверевой десяти респондентам было предложено самостоятельно установить эмотивность текста. В результате тональность лишь двух фрагментов из десяти была оценена всеми участниками одинаково, что также совпало с оценкой, полученной автоматизированным способом. В остальных случаях толкование тональности текстов у респондентов разнилось, что указывает на неоднозначность восприятия текста не только программой, но и носителями языка (экспертами). В ходе анализа система обрабатывает отдельные слова и словосочетания, в то время как респондент рассматривает текст как целостное единство [26], кроме того, восприятие сообщения и поиск скрытого смысла иногда может вызывать трудности даже у человека [24].

Среди возможных причин, препятствующих высокому уровню достоверности результатов автоматизированного анализа, называют следующие: 1) наличие имплицитной оценки, полноценно воспринимающейся человеком, но представляющей трудности для системы; 2) использование иронии и сарказма, поскольку система оперирует лишь графемами и словоформами; 3) монотематичность системы (программирование на анализ лишь в определенной тематической сфере и, следовательно, учет лексики, свойственной только ее текстам), тональность может определяться предметной областью текста [25]; 4) дизамбигуация (решение многозначности), связанная с разной тональностью одного слова – одна и та же характеристика может иметь разную эмоциональную оценку для разных объектов [24]; 5) проблемы референции и кореференции, способствующие возникновению потенциальных неточностей в оценке; 6) синтаксические особенности текста [23].

Помимо этого, в ряду возможных причин отмечается разговорный стиль речи, присущий многим текстам интернет-пространства и содержащий сленг, специфичную пунктуацию и стилистику, опечатки и т. п., что вызывает трудности при работе автоматической системы анализа тональности. Способность к распознаванию эмотиконов также может увеличить точность определения тональности [24].

А.А. Юрганов утверждает, что употребление отрицания способно изменять тональность текста вплоть до противоположного результата, поскольку «значение тональности зависит от того, кто проводит анализ. К примеру, фраза «У KFC отлично идут дела» имеет положительный окрас для компании KFC и отрицательный для McDonald’s» [25, с. 40].

В настоящем исследовании автоматизированный анализ текста был применен непосредственно к англоязычным и русскоязычным читательским комментариям интерактивных газетных статей. Определение сентимент-составляющей читательского комментария посредством сопоставления автоматизированной и ручной обработки может способствовать не только выявлению культурной специфики представителей двух языковых сообществ, но и дальнейшему изучению автоматизированных способов анализа тональности теста.

Методы и материал исследования

В исследовании использован метод сентимент-анализа, позволяющий определить эмоциональную тональность читательского комментария на основании лексического анализа текстового материала. В качестве инструмента автоматизированной обработки текста применялся прикладной программный интерфейс ParallelDots API[1]. Одним из сервисов интерфейса является проведение сентимент-анализа текстов потребительских отзывов к блогам, статьям, форумам, услугам и продуктам на четырнадцати языках, включая английский и русский. Используемый метод классификации тональности – машинное обучение, модели обработки текстов на естественном языке (NLP) обучены на более чем миллиарде текстовых документов[2]. Тональность текста определяется при помощи алгоритмов

[1] URL: https://www.paralleldots.com/sentiment-analysis, свободный.
[2] URL: https://www.paralleldots.com/text-analysis-apis, свободный.

одной из разновидностей архитектуры рекуррентных нейронных сетей – долгой краткосрочной памяти (англ.: LSTM, Long short-term memory). Разработчики ParallelDots Sentiment Analysis API отмечают высокую точность результатов анализа и его устойчивость к семантически и структурно сложным предложениям, содержащим, например, двойное отрицание и нетипичный для языка порядок слов[3].

Результаты автоматизированной обработки англоязычных и русскоязычных читательских комментариев верифицировались ручным (экспертным) способом с применением метода текстового анализа. Исследование состояло из четырех этапов.

На первом этапе была произведена выборка электронных газетных публикаций, тематика которых, по нашему предположению, может вызвать значительный эмоциональный отклик у читателей – природные катаклизмы (ураган Мэтью, 2019 г.), авиакатастрофы (крушение Boeing 737, 2020 г.), техногенные катастрофы (взрыв в Бейруте, 2020 г.) и т. п. Использовались публикации в электронных газетных изданиях the Independent (I.), the Guardian (G.), Комсомольская правда (К.П.), РИА Новости (РИА). Затем была составлена картотека англоязычных и русскоязычных читательских комментариев к выбранным газетным статьям. Общее количество комментариев составило 3000, из них 1638 к изданиям на английском языке и 1362 к российским изданиям.

На втором этапе исследования отобранные комментарии читателей подверглись автоматизированной обработке с помощью программного интерфейса ParallelDots API (метод сентимент-анализа). В результате программной обработки текстового материала эмотивная составляющая комментариев была подразделена на три вида тональности – позитивную, нейтральную и негативную.

На третьем этапе результаты автоматизированной обработки читательских комментариев были верифицированы с применением ручного текстового лингвистического анализа. В качестве экспертов были привлечены специалисты в области филологии английского и русского языков из числа профессорско-преподавательского состава Казанского федерального университета ($N = 6$), а также студенты-филологи, носители русского языка, владеющие английским языком как иностранным на уровне C1 ($N = 4$). Экспертам было предложено ознакомиться с текстом англоязычных и русскоязычных комментариев и определить их тональность как позитивную, нейтральную или негативную. При расхождении в определении тональности одного и того же комментария отдельными экспертами окончательное решение принималось исходя из большинства (60% и более) полученных оценок. Необходимость проведения ручного анализа была связана с обнаружением неточностей программных оценочных результатов вследствие особенностей лексического, грамматического и графического оформления текста. На данном этапе внимание уделялось также эмпирическому подходу, который не является компонентом автоматизированного анализа.

На четвертом этапе исследования было проведено сопоставление результатов, полученных в ходе второго и третьего этапов, с выделением так называемых «сентимент-пар». Установление процентного расхождения между автоматизирован-

[3] URL: https://blog.paralleldots.com/product/contextual-sentiment-analysis-applications/, свободный.

ным и неавтоматизированным анализом тональности читательских комментариев имело своей целью, во-первых, обнаружение причин ошибочной программной оценки, во-вторых, выявление культурологических особенностей англоязычных и русскоязычных комментариев.

Результаты исследования

Проведенный сентимент-анализ читательских комментариев с применением ParallelDots API выявил процентное соотношение трех видов тональности: позитивной, негативной и нейтральной – со значительным преобладанием негативной тональности в обоих языках. Сопоставление результатов программной обработки с данными ручного анализа позволило установить значительное преобладание расхождений в определении всех трех видов тональности, в особенности негативной. Статистические данные по обнаруженным совпадениям/расхождениям в определении трех видов тональности автоматизированным и ручным способами в обоих языках представлены в табл. 1.

Табл. 1

Данные по обнаруженным совпадениям/расхождениям в определении тональности автоматизированным и ручным способами

	Англоязычные комментарии		Русскоязычные комментарии	
	Совпадения	Расхождения	Совпадения	Расхождения
Позитивная тональность	27	106	32	99
Негативная тональность	709	490	615	252
Нейтральная тональность	171	135	104	260
Общее кол-во совпадений/ расхождений	907	731	751	611
Общее кол-во комментариев	1638		1362	

Данные в табл. 1 показывают, что количество расхождений по результатам сентимент-анализа, проведенного автоматизированным и ручным способами, значительно преобладает над количеством совпадений, и это характерно для обоих языков, но в особенности английского.

Подробнее результаты программной и ручной оценки читательских комментариев представлены в процентах на рис. 1.

В англоязычном тексте в результате автоматизированной обработки выявлено меньше комментариев позитивной тональности по сравнению с ручной, при этом разница составила 1%. В русскоязычном тексте больше комментариев позитивной тональности обнаружено в результате ручной обработки, разница составила 2.6% (рис. 1).

Существенные расхождения выявлены в отношении англоязычных читательских комментариев негативной и нейтральной тональности. Негативно окрашенные комментарии составили 73.2% по итогам программной обработки и 54.2%

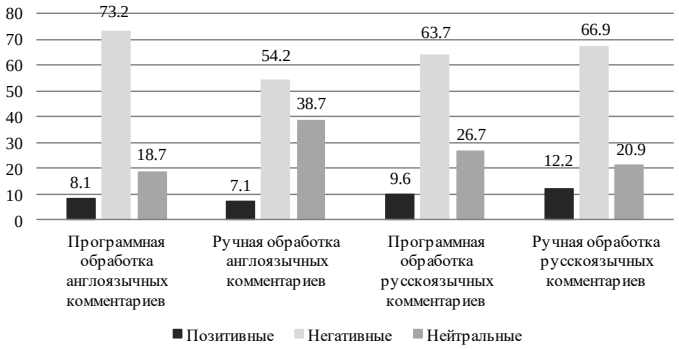

Рис. 1. Общее процентное соотношение автоматизированного и ручного сентимент-анализа англоязычных и русскоязычных читательских комментариев

по итогам ручной (расхождение в 19%). Более того, 18.7% комментариев программа определила как нейтральные, тогда как последующий ручной анализ показал наличие нейтральной тональности в 38.7% комментариев (расхождение в 20%) (см. рис. 1). Полученные данные свидетельствуют о том, что программная обработка недостаточно корректно выявляет негативную и нейтральную тональность в англоязычном тексте.

В отношении русскоязычных читательских комментариев негативной и нейтральной тональности подобных несоответствий результатов автоматизированной и ручной обработки текста не выявлено. Подробнее результаты анализа представлены на рис 1.

Таким образом, верификация результатов автоматизированной обработки англоязычных читательских комментариев ручным способом показала, что программа завысила количество негативно окрашенных комментариев и занизила число нейтральных. В отношении русскоязычного текста программа выявила меньший процент негативных, но больший процент нейтральных комментариев.

По причине значительных расхождений между результатами программной и ручной обработки текста авторы приняли решение определить процентное количество совпадений и несовпадений между двумя способами сентимент-анализа, для чего были составлены так называемые тональные пары (сентимент-пары) по принципу: «системно-определенная эмотивность + экспертно-определенная эмотивность», где под первой понимается программная обработка, а под второй – ручная (экспертная). Анализ частотности тональных пар был применен в ходе сопоставления англоязычных и русскоязычных читательских комментариев.

Комментарии позитивной тональности. Среди англоязычных читательских комментариев совпадение результатов системно-определенной позитивной тональности с аналогичной экспертной оценкой составило 1.7%, тогда как в случае с русскоязычными комментариями это число достигло 2.4% (рис. 2). Количество случаев, когда система оценила комментарии как позитивно окрашенные, а ручная обработка выявила нейтральную тональность, в англоязычных комментариях составило 2.9%, в русскоязычных – 1.6% (см. рис. 2). В отношении комментариев, определенных системой как позитивные, тогда как экспертная

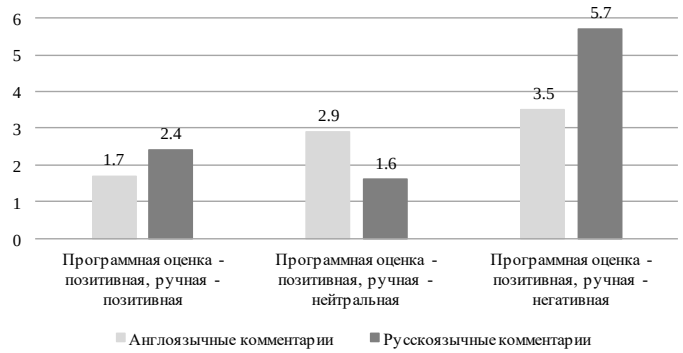

Рис. 2. Процентное расхождение позитивных (согласно программной обработке) комментариев с их экспертной оценкой

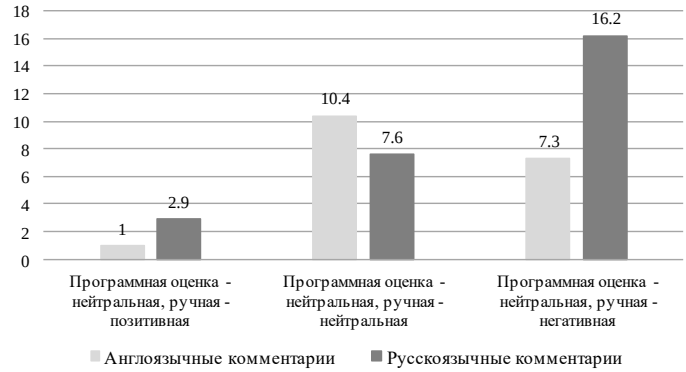

Рис. 3. Процентное расхождение нейтральных (согласно программной обработке) комментариев с их экспертной оценкой

оценка показала наличие негативной тональности, было выявлено 3.5% англоязычных сентимент-пар и 5.7% русскоязычных. Результаты анализа сентимент-пар представлены на рис. 2.

Комментарии нейтральной тональности. Сопоставительный анализ выявил 1% англоязычных и 2.9% русскоязычных комментариев, в которых программа выявила нейтральную тональность, а экспертный анализ показал позитивную. Совпадение результатов в отношении нейтральной тональности наблюдается в 10.4% англоязычных комментариев и 7.6% русскоязычных. В 7.3% англоязычных комментариев и 16.2% русскоязычных программная обработка выявила нейтральную тональность, а ручная – негативную. Данные анализа нейтральной тональности обобщены на рис. 3.

Комментарии негативной тональности. Количество комментариев, тональность которых вследствие программной обработки была определена как негативная, однако экспертная оценка выявила позитивную эмотивность, составило 4.6% для англоязычного текста и 6.9% для русскоязычного. 25.3% англоязычных комментариев и 11.6% русскоязычных, определенных программой как негативно

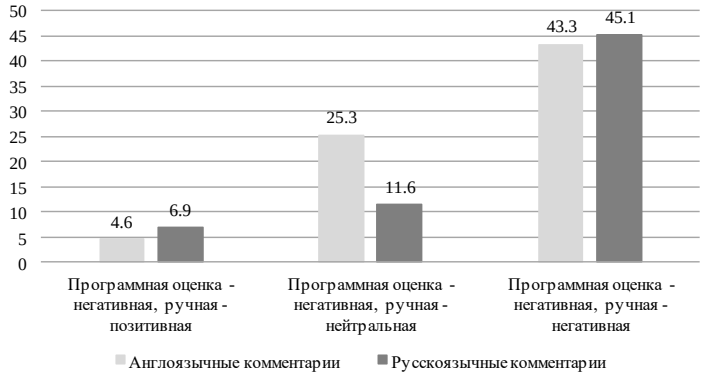

Рис. 4. Процентное расхождение негативных (согласно программной обработке) комментариев с их экспертной оценкой

окрашенные, в результате экспертной оценки отнесены к комментариям нейтральной тональности. Совпадение результатов автоматизированного и экспертного анализа комментариев негативной тональности выявлено в 43.3% англоязычных и 45.1% русскоязычных комментариев. Данные анализа негативной тональности читательских комментариев представлены на рис. 4.

Обсуждение полученных результатов

Авторы проанализировали корректность работы программы, определившей комментарии как позитивные, негативные или нейтральные, сопоставив данные с результатами ручного экспертного анализа.

В результате проведенного сопоставительного анализа были выделены следующие закономерности, осложняющие корректное определение тональности читательского комментария автоматизированным методом.

1. Использование читателями лексических единиц, значение которых может полностью менять интерпретацию текста, несмотря на наличие в них ярко выраженной эмотивности (как позитивной, так и негативной). Наглядным примером являются случаи поддержки пострадавшей стороны (позитивная эмотивность) с использованием негативно-оценочной лексики (англ.: *dreadful (ужасный), cruel blow (жестокий удар),* рус.: *ужас* и др.), как, например, в комментариях (1) и (2), отмеченных программой как негативные, а экспертами – как позитивные:

(1) *Haiti has been in crisis for a very long time. This is just another cruel blow. They have not yet "recovered" from Katrina! If ever there was a hellhole created by imperialism, Haiti is the poster child. Dreadful. Just dreadful*[4] (G.). (Гаити уже давно переживает кризис. Это просто еще один жестокий удар. Они еще не «оправились» от Катрины! Если когда-то и была адская бездна, созданная империализмом, то Гаити является живым примером. Ужасно. Просто ужасно[5].)

(2) *Нет слов, чтобы выразить все свое сострадание японцам! Ужасная трагедия. Всем сердцем с вами, Люди! Дай Бог вам сил, подняться и их этой Беды!* (К.П.)

2. Наличие иронии и сарказма, которые не могут быть распознаны при автоматизированном сентимент-анализе. Данное явление особенно характерно для

[4] Во всех примерах сохранена орфография и пунктуация оригинала.
[5] Здесь и далее перевод наш. – *Г.К., Э.Д.*

комментариев позитивной или нейтральной тональности (согласно автоматизированной обработке). В примере (3) используется прием иронии (*johnny boy – малыш Джонни*), примеры (4) и (5) содержат положительно-оценочную лексику (*умник, молодцы*), которая используется в данном контексте саркастически, поэтому текстовый анализ экспертами выявил негативную тональность, тогда как программа указала на позитивную:

(3) *no Plain English Award for you johnny boy* (I.). (нет тебе награды в номинации ‘Простой английский’, малыш Джонни.)

(4) *Sure. Now Google “Hillary Clinton Pizzagate” and count the hits. Genius…* (G.) (Конечно. Теперь погуглите “Хиллари Клинтон Пиццагейт” и посчитайте попадания. Умник…)

(5) *260 тонн воды? это один бассейн. Молодцы* (РИА).

3. Графическое и пунктуационное оформление текста комментария и невозможность однозначной интерпретации графических символов при автоматизированном анализе текста. В ходе исследования было обнаружено, что многократное использование восклицательного и вопросительного знаков и многоточия чаще всего демонстрирует отрицательную тональность читательского комментария. Наличие графических эмотиконов (в большей мере свойственных русскоязычным читательским комментариям) способствует кардинальному изменению эмотивности текста, зачастую являясь интенсификатором иронии. Примеры (6) и (7) иллюстрируют негативную тональность, созданную графическими средствами и пунктуацией, усиливающими отрицательные эмоции:

(6) *У природы нет, плохой погоды!)* (РИА);

(7) *STOP AGGRESSION, azerbaijan!!! Nagorno Karabakh Republic has never been and will never be a part of azerbaijan, REMEMBER!!!!!!!* (I.) (ОСТАНОВИ АГРЕССИЮ, Азербайджан!!! Нагорно-Карабахская Республика никогда не была и не будет в составе Азербайджана, ЗАПОМНИ!!!!!!!)

4. Прослеживается тенденция различной интерпретации читательского комментария в зависимости от того, является ли текст первичным комментарием или представляет собой обращение к другому участнику обсуждения. Отмечается, что в ответных комментариях вероятность негативной тональности больше, чем позитивной или нейтральной. Тональность комментариев (8) и (9) определена системой как нейтральная, тогда как эксперты указали на негативную:

(8) *А теперь представь что твои, например, дети или родители там жили. Красиво?* (РИА)

(9) *Do you want them to rename it Hurricane Jamal instead then?* (G.) (Ты хочешь, чтобы вместо этого они переименовали его в ураган Джамал?)

5. Обнаруженные случаи несовпадения автоматизированной и экспертной оценки комментариев, содержащих междометия, позволяют предположить, что они также могут затруднять верное определение тональности комментария. Например, комментарий (10) отмечен программой как негативный, а экспертами как позитивный, тогда как пример (11) в результате атоматизированной обработки определен как нейтральный, а экспертами выявлен негативный сентимент:

(10) *Эх, Граждане! Вы тут начинаете путать одно с другим. Моя семья все эти дни сочуствует Японцам и желает им быстрого восстановления* <…> (К.П.);

(11) *Ooh look a Fox commenter.* (I.) (Ооо, посмотрите-ка, что пишет пользователь Fox.)

6. Автоматизированному анализу свойственно сканирование текста посредством выделения ранее запрограммированных и, следовательно, «знакомых»

лексических единиц и групп, в то время как экспертная обработка подразумевает рассмотрение текста комментария как единого целого. Поэтому степень объективности выше в результате экспертного анализа. В примерах (12) и (13), согласно ручной обработке, положительное отношение комментаторов демонстрируется посредством лексики (*I feel sorry – мне жаль; жаль искренне* и др.), использованной в текстах для выражения сочувствия по отношению к пострадавшей стороне. Полагаем, что программа обработала текст комментариев (12) и (13) только на предмет лексических единиц, определив тональность как негативную, в то время как экспертный анализ был применен к тексту как к единому целому, и в результате была выявлена позитивная тональность:

(12) *It is cñear that the situation in Lebanpn cannot ven start to improve, as long as the situation in Israel is unresolved. As there are no signs of that happening, becuase Israel simply does not know ehat is wants, I feel sorry for Lebanon* (G.). (Понятно, что ситуация в Ливане не может начать улучшаться, пока не урегулирована ситуация в Израиле. Нет никаких признаков того, что это произойдет, так как Израиль просто не знает, чего он хочет, мне жаль Ливан.)

(13) «*Будем давить…» Погибших, конечно жаль, искренне. Но и лицо свое терять нельзя. Иран, на моей памяти – единственная страна, безоговорочно признавшая свою вину за сбитый пассажирский самолет. Остальные так или иначе вечно увиливали от ответственности. Даже если и отпираться было глупо. Примеров – масса… Самые яркие – США, сбившие иранский лайнер, да и сама Украина, сбившая российский борт* (РИА).

Таким образом, в ходе исследования было выявлено в некоторых случаях несоответствие результатов автоматизированного и неавтоматизированного анализа тональности читательских комментариев. Как в англоязычных, так и в русскоязычных комментариях при обоих подходах к обработке наблюдалось доминирование негативной тональности. Небольшое процентное расхождение в определении позитивной тональности свидетельствует о том, что оба способа текстового анализа в целом опираются на наличие в тексте эмоционально маркированных лексических единиц.

Русскоязычные комментарии, оцененные экспертным способом, чаще содержат эмоционально окрашенную лексику, что приводит к преобладанию позитивной или негативной тональности по сравнению с нейтральной. Примечательно, что процент англоязычных читательских комментариев превысил процент русскоязычных именно в данной категории, а по всем остальным параметрам процент русскоязычных комментариев превалировал.

Заключение

Автоматический анализ тональности текста читательских комментариев продемонстрировал достаточно объективные результаты, особенно в отношении позитивной тональности. Основное преимущество автоматической обработки заключается в технической возможности анализа большего количества материала в значительно меньшие сроки по сравнению с ручным способом. Используемая в настоящем исследовании система ParallelDots API также предоставляет возможность получить информацию о тональности комментария, содержащую процентное количество присутствующей позитивной, нейтральной и негативной эмотивности одновременно, что не представляется возможным при ручной (экспертной) оценке.

Трудности, препятствующие более совершенному определению тональности программой, были обнаружены в случаях с комментариями представителей обеих анализируемых языковых групп, что указывает на схожесть структурного построения англоязычных и русскоязычных читательских комментариев как на лексическом, так и на синтаксическом уровне. Больший объем (как в содержательном, так и в количественном плане) комментариев англоязычных читателей свидетельствует об активной вовлеченности англоговорящих пользователей в интернет-дискурс и об их стремлении к выражению личного мнения в полной и содержательной форме. Русскоязычные пользователи, напротив, склонны к написанию кратких и лаконичных комментариев. Авторы предполагают, что большой объем англоязычного текстового материала в некоторых случаях затруднял работу системы ParallelDots API.

Показательным является доминирование негативной тональности читательских комментариев не только в обоих языковых группах, но и при программной и экспертной оценке. При экспертной оценке выявлено превалирование приемов иронии и сарказма в русскоязычных комментариях, что, вследствие невозможности системы проанализировать имплицитность текста, было оценено ею как положительная тональность.

Проведенное исследование позволило определить некоторые сложности, препятствующие максимально корректной работе автоматизированного сентимент-анализа читательского комментария. Мы пришли к выводу об эффективности синтезированного использования программного (автоматизированного) и ручного (экспертного) анализа текста с целью получения более детальных и объективных данных.

Результаты могут быть полезными при проведении сопоставительных исследований двух и более языков методом сентимент-анализа, при сравнении лексической, грамматической, культурологической составляющих языков, а также для исследований в области когнитивной лингвистики.

Источники

I. – News. The Independent. Today’s Headlines and Latest Breaking News. – URL: https://independent.co.uk, свободный.

G. – News, Sport and Opinion from the Guardian’s Global Edition. The Guardian. – URL: https://theguardian.com/international, свободный.

К.П. – Комсомольская правда в Москве. – URL: https://msk.kp.ru/, свободный.

РИА – РИА Новости – события в Москве, России и мире сегодня: темы дня, фото, видео, инфографика, радио. – URL: https://ria.ru, свободный.

Литература

1. *Топчий И.В.* Креативное комментирование в социальных медиа: обзор исследований // Изв. Урал. фед. ун-та. Сер. 1. Проблемы образования, науки и культуры. – 2020. – Т. 26, № 2. – С. 22–28.
2. *Гималетдинова Г.К.* Лингвистические основы интерактивной газетной статьи: к постановке вопроса // Полит. лингвистика. – 2012. – № 3. – С. 143–148.

3. *Нашхоева М.Р.* Лингвистическая концепция эмоций и эмотивности текста // Вестн. ЮУрГУ – 2011. – № 1. – С. 95–98.
4. *Пазельская А.Г., Соловьев А.Н.* Метод определения эмоций в текстах на русском языке // Компьютерная лингвистика и интеллектуальные технологии: Материалы ежегодной междунар. конф. «Диалог». – М.: Изд-во РГГУ, 2011. – Вып. 10. – С. 510–522.
5. *Pang B., Lee L.* Opinion mining and sentiment analysis // Foundations and Trends in Information Retrieval. – 2008. – V. 2, No 1–2. – P. 1–135. – doi: 10.1561/1500000011.
6. *Гималетдинова Г.К.* О монологичности и диалогичности англоязычного читательского комментария // Филология и культура. Philology and Culture. – 2012. – № 3. – С. 30–34.
7. *Гималетдинова Г.К., Довтаева Э.Х.* Сентимент-анализ читательского интернет-комментария к политическому тексту // Полит. лингвистика. – 2020. – № 1. – С. 42–51.
8. *Горошко Е.И., Полякова Т.Л.* К построению типологии жанров социальных медий // Жанры речи. – 2015. – № 2. – С. 119–127.
9. *Горошко Е.И., Землякова Е.А.* Виртуальное жанроведение: становление теоретической парадигмы // Уч. зап. Таврич. нац. ун-та им. В.И. Вернадского. Сер. «Филология. Социальные коммуникации». – 2011. – Т. 24, № 1–1. – С. 225–237.
10. *Дахалаева Е.Ч.* Интернет-комментарий и интернет-отзыв: параметры жанрового разграничения // Современные проблемы науки и образования. – 2014. – № 6. – URL: http://www.science-education.ru/ru/article/view?id=16222, свободный.
11. *Митягина В.А.* Интернет-комментарий как коммуникативное воздействие // Жанры и типы текста в научном и медийном дискурсе: Межвуз. сб. науч. тр. – Орел: ФГБОУ ВПО «ОГИИК», ООО «Горизонт», 2012. – Вып. 10. – С. 188–197.
12. *Щипицина Л.Ю.* Классификация жанров компьютерно-опосредованной коммуникации по их функции // Изв. РГПУ им. А. И. Герцена. – 2009. – № 114. – С. 171–178.
13. *Мельник Н.В.* Лингвоперсонология политического интернет-комментария // Полит. лингвистика. – 2017. – № 5. – С. 47–51.
14. *Савельева И.В.* Лингвоперсонологический потенциал интернет-комментария // Сиб. филол. журн. – 2017. – № 4. – С. 192–201.
15. *Степанова Л.Н.* Комментарий в современном информационно-коммуникативном пространстве: перспективы лингвистического исследования // Современная филология: Материалы II Междунар. науч. конф. – Уфа: Лето, 2013. – С. 94–97.
16. *Кураш В.В.* Стилистический потенциал интернет-комментария // Вестн. МГУП им. Ивана Фёдорова. – 2015. – № 2. – С. 99–103.
17. *Шхумишхова А.Р., Калашаова А.А.* Лингвостилистические особенности интернет-комментария в СМИ // Cross Cultural Studies: Education and Science. – 2018. – № 3. – С. 387–392.
18. *Карпоян С.М.* Эпистемическая модальность в интернет-комментарии: Автореф. дис. … канд. филол. наук. – Ростов н/Д, 2014. – 26 с.
19. *Пасечная Л.А., Щербина В.Е.* Способы выражения оценки в интернет-комментариях // Изв. ВГПУ. – 2020. – № 9. – С. 87–92.
20. *Вольф Е.М.* Функциональная семантика оценки. – М.: URSS, 2020. – 278 с.
21. *Ляпун С.В., Соколова Г.В.* Вербализация эмоциональной напряженности в интернет-комментариях читателей «Новой газеты» // Вестн. Адыг. гос. ун-та. Сер. 2: Филология и искусствоведение. – 2016. – № 4. – С. 231–235.
22. *Baccianella S., Esuli A., Sebastiani F.* Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining // Proc. 7th Int. Conf. on Language Resources and Evaluation (LREC'10). – Valletta: Eur. Lang. Resour. Assoc., 2010. – V. 10. – P. 2200–2204.

23. *Семина Т.А.* Анализ тональности текста: современные подходы и существующие проблемы // Социальные и гуманитарные науки. Отечественная и зарубежная литература. Сер. 6, Языкознание: Реферативный журн. – 2020. – № 4. – С. 47–63.

24. *Сметанина Н.И.* Основные задачи анализа тональности текстов в социальных сетях // Скиф. – 2017. – № 15. – С. 96–99.

25. *Юрганов А.А.* Сентимент-анализ как инструмент исследования текстов // Проблемы современной науки и образования. – 2017. – № 29. – С. 39–41.

26. *Зверева П.П.* Сентимент-анализ текста (на материале печатных текстов газеты «The New York Times» о России и россиянах) // Вестн. МГОУ. Сер. Лингвистика. – 2014. – № 5. – С. 32–37.

27. *Колмогорова А.В.* Использование текстов жанра «Интернет-откровение» в контексте решения задач сентимент-анализа // Вестн. НГУ. Сер. Лингвистика и межкультурная коммуникация. – 2019. – № 3. – С. 71–82.

Поступила в редакцию
28.12.2020

Гималетдинова Гульнара Камилевна, кандидат филологических наук, доцент кафедры романо-германской филологии

Казанский (Приволжский) федеральный университет
ул. Кремлёвская, д. 18, г. Казань, 420008, Россия
E-mail: *gim-nar@yandex.ru*

Довтаева Эмили Хамзатовна, студент Института филологии и межкультурной коммуникации

Казанский (Приволжский) федеральный университет
ул. Кремлёвская, д. 18, г. Казань, 420008, Россия
E-mail: *emily_dovtaeva@mail.ru*

ISSN 2541-7738 (Print)
ISSN 2500-2171 (Online)

UCHENYE ZAPISKI KAZANSKOGO UNIVERSITETA. SERIYA GUMANITARNYE NAUKI
(Proceedings of Kazan University. Humanities Series)

2021, vol. 163, no. 1, pp. 65-80

ORIGINAL ARTICLE

doi: 10.26907/2541-7738.2021.1.65-80

Sentiment Analysis of Reader Comments: Automated vs Manual Text Processing

G.K. Gimaletdinova[*], *E.Kh. Dovtaeva*[**]

Kazan Federal University, Kazan, 420008 Russia
E-mail: [*]*gim-nar@yandex.ru*, [**]*emily_dovtaeva@mail.ru*

Received December 28, 2020

Abstract

The verbal and structural features of the reader comment, a genre of Internet communication, were studied. The method of sentiment analysis (ParallelDots API) was used to reveal and measure the emotive component of the reader comments ($N = 3000$) in the English and Russian languages. The results obtained were verified by the manual linguistic text analysis. The experts were specialists in the field of philology of the English and Russian languages ($N = 6$), students of philology, as well as native speakers of the Russian language for whom English is a foreign language, i.e., their level of proficiency is C1 ($N = 4$). As a result of the comparison of the data collected through the automated and manual text pro-

cessing, a number of factors that reduce the reliability of the results of automated sentiment analysis of the reader comments were singled out. Difficulties hindering the objective determination of the sentiment by the program were found in the reader comments in both analyzed languages. This is indicative of the structural similarities between the English and Russian reader comments at the lexical and syntactic levels. The feasibility of the mixed automated and manual text processing in order to obtain more detailed and objective data was demonstrated. The results of this work can be used for comparative studies of two or more languages performed by the method of sentiment analysis, as well as for drawing parallels between the lexical, grammatical, and cultural components of languages.

Keywords: text sentiment, sentiment analysis, participatory news article, reader comment, emotiveness

Figure Captions

Fig. 1. Total percentage obtained as a result of the automated and manual sentiment analysis of the reader comments in the English and Russian languages.

Fig. 2. Percentage discrepancy between the positive (according to the software processing) comments and their expert assessment.

Fig. 3. Percentage discrepancy between the neutral (according to the software processing) comments and their expert assessment.

Fig. 4. Percentage discrepancy between the negative (according to the software processing) comments and their expert assessment.

References

1. Topchii I.V. Creative commenting in social media: A review of studies. *Izvestiya Ural'skogo Federal'nogo Universiteta. Seriya 1. Problemy Obrazovaniya, Nauki i Kul'tury*, 2020, vol. 26, no. 2, pp. 22–28. (In Russian)
2. Gimaletdinova G.K. Linguistic background of participatory news article: On statement of the problem. *Politicheskaya Lingvistika*, 2012, no 3, pp. 143–148. (In Russian)
3. Nashkhoeva M.R. Linguistic concept of emotions and text emotiveness. *Vestnik YuUrGU*, 2011, no. 1, pp. 95–98. (In Russian)
4. Pazel'skaya A.G., Solov'ev A.N. The method for identifying emotions in Russian texts. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Materialy ezhegodnoi mezhdunarodnoi konferentsii "Dialog"* [Computational Linguistics and Intelligent Technologies: Proc. Annu. Int. Conf. "Dialogue"], 2011, no. 10, pp. 510–522. (In Russian)
5. Pang B., Lee L. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2008, vol. 2, nos. 1–2, pp. 1–135. doi: 10.1561/1500000011.
6. Gimaletdinova G.K. Monologue and dialogue of the reader comments in the English language. *Filologiya i kul'tura. Philology and Culture*, 2012, no. 3, pp. 30–34. (In Russian)
7. Gimaletdinova G.K., Dovtaeva E.Kh. Sentiment analysis of the reader Internet comment on the political text. *Politicheskaya Lingvistika*, 2020, no. 1, pp. 42–51. (In Russian)
8. Goroshko E.I., Polyakova T.L. On the development of the genre typology for social media. *Zhanry Rechi*, 2015, no. 2, pp. 119–127. (In Russian)
9. Goroshko E.I., Zemlyakova E.A. Virtual genre studies: Establishment of a theoretical paradigm. *Uchenye Zapiski Tavricheskogo Natsional'nogo Universiteta imeni V.I. Vernadskogo. Seriya "Filologiya. Sotsial'nye Kommunikatsii"*, 2011, vol. 24, no. 1-1, pp. 225–237. (In Russian)
10. Dakhalaeva E.Ch. Internet comment and Internet review: Parameters of genre distinction. *Sovremennye Problemy Nauki i Obrazovaniya*, 2014, no. 6. Available at: http://www.science-education.ru/ru/article/view?id=16222. (In Russian)
11. Mityagina V.A. Internet commenting as a communicative influencing. In: *Zhanry i tipy teksta v nauchnom i mediinom diskurse* [Genres and Types of Texts in Scientific and Media Discourse]. Orel, FGBOU VPO "OGIIK", OOO "Gorizont", 2012, no. 10, pp. 188–197. (In Russian)
12. Shchipitsina L.Yu. Classification of computer-mediated communication genres based on their functions. *Izvestiya RGPU imeni A.I. Gertsena*, 2009, no. 114, pp. 171–178. (In Russian)

13. Mel'nik N.V. Linguopersonology of the political Internet comment. *Politicheskaya Lingvistika*, 2017, no. 5, pp. 47–51. (In Russian)
14. Savel'eva I.V. Linguopersonological potential of the Internet comment. *Sibirskii Filologicheskii Zhurnal*, 2017, no. 4, pp. 192–201. (In Russian)
15. Stepanova L.N. Commenting in the modern information and communication space: Prospects of linguistic research. *Sovremennaya filologiya: Materialy II mezhdunar. nauch. konf.* [Modern Philology: Proc. II Int. Sci. Conf.]. Ufa, Leto, 2013, pp. 94–97. (In Russian)
16. Kurash V.V. Stylistic potential of Internet comments. *Vestnik MGUP imeni Ivana Fedorova*, 2015, no. 2, pp. 99–103. (In Russian)
17. Shkhumishkhova A.R., Kalashaova A.A. Linguistic features of an Internet comment in mass media. *Cross Cultural Studies: Education and Science*, 2018, no. 3, pp. 387–392. (In Russian)
18. Karpoyan S.M. Epistemic modality and Internet comments. *Extended Abstract of Cand. Philol. Diss.* Rostov-on-Don, 2014. 26 p. (In Russian)
19. Pasechnaya L.A., Shcherbina V.E. The ways for expressing evaluative attitudes in Internet comments. *Izvestiya VGPU*, 2020, no. 9, pp. 87–92. (In Russian)
20. Volf E.M. *Funktsional'naya semantika otsenki* [Functional Semantics of Evaluation]. Moscow, URSS, 2020. 278 p. (In Russian)
21. Lyapun S.V., Sokolova G.V. Verbalization of emotional tension in the Internet comments by readers of the Novaya Gazeta. *Vestnik Adygeiskogo Gosudarstvennogo Universiteta. Seriya 2: Filologiya i Iskusstvovedenie*, 2016, no. 4, pp. 231–235. (In Russian)
22. Baccianella S., Esuli A., Sebastiani F. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. *Proc. 7th Int. Conf. on Language Resources and Evaluation (LREC'10)*. Valletta, Eur. Lang. Resour. Assoc., 2010, vol. 10, pp. 2200–2204.
23. Semina T.A. Sentiment analysis: Modern approaches and existing problems. *Sotsial'nye i Gumanitarnye Nauki. Otechestvennaya i Zarubezhnaya Literatura. Seriya 6. Yazykoznanie*, 2020, no. 4, pp. 47–63. (In Russian)
24. Smetanina N.I. Basic tasks of sentiment analysis in social networks. *Skif*, 2017, no. 15, pp. 96–99. (In Russian)
25. Yurganov A.A. Sentiment analysis as a tool for studying texts. *Problemy Sovremennoi Nauki i Obrazovaniya*, 2017, no. 29, pp. 39–41. (In Russian)
26. Zvereva P.P. Sentiment analysis of text (based on texts about Russia and the Russians from the New York Times). *Vestnik MGOU. Seriya: Lingvistika*, 2014, no. 5, pp. 32–37. (In Russian)
27. Kolmogorova A.V. Texts of "Internet confession" with regard to solving the tasks of sentiment analysis. *Vestnik NGU. Seriya: Lingvistika i Mezhkul'turnaya Kommunikatsiya*, 2019, no. 3, pp. 71–82. (In Russian)

***Для цитирования*:** *Гималетдинова Г.К., Довтаева Э.Х.* Сентимент-анализ читательского комментария: автоматизированная vs ручная обработка текста // Учен. зап. Казан. ун-та. Сер. Гуманит. науки. – 2021. – Т. 163, кн. 1. – С. 65–80. – doi: 10.26907/2541-7738.2021.1.65-80.

***For citation*:** Gimaletdinova G.K., Dovtaeva E.Kh. Sentiment analysis of reader comments: Automated vs manual text processing. *Uchenye Zapiski Kazanskogo Universiteta. Seriya Gumanitarnye Nauki*, 2021, vol. 163, no. 1, pp. 65–80. doi: 10.26907/2541-7738.2021.1.65-80. (In Russian)