ИНФОРМАЦИОННЫЕ СИСТЕМЫ И МЕДИАТЕХНОЛОГИИ

INFORMATION SYSTEMS AND MEDITECHNOLOGIES

УДК 81’33

А. А. Баркович, А. В. Антонов
Минский государственный лингвистический университет

МЕТОДОЛОГИЧЕСКИЙ ПОТЕНЦИАЛ СЕНТИМЕНТ-АНАЛИЗА: КОРПУСНЫЙ АСПЕКТ

Данная статья посвящена аспектам проведения процедуры сентимент-анализа на материале корпуса текстов. Сентимент-анализ текста традиционно ориентирован на оценку небольших речевых артефактов, типичным примером обрабатываемого текстового материала является пост блога или мессенджера. Вместе с тем постоянное развитие корпусного формата обусловливает актуальность реализации в референтном контексте всего спектра возможностей автоматизированной обработки речи, в том числе лингвопрагматической по сути практики сентимент-анализа. Данные возможности востребованы, однако их полноценная реализация предполагает предварительную апробацию существующих инструментов сентимент-анализа и их адаптацию для проведения исследований, в частности, в полностью совместимом с компьютерно-опосредованной коммуникацией корпусном аспекте – на качественно инновационном уровне. В процессе проведения соответствующего анализа был систематизирован и охарактеризован имеющийся методологический потенциал и предложены как пути тактической актуализации, так и стратегия совершенствования практики сентимент-анализа. Научная репрезентация релевантной проблематики поспособствует более полному раскрытию интердисциплинарного потенциала сентимент-анализа и позволит повысить результативность и эффективность его проведения.

Ключевые слова: сентимент-анализ, корпус текстов, тональность текста, тональный словарь, оценка, контекст.

Для цитирования: Баркович А. А. Методологический потенциал сентимент-анализа: корпусный аспект // Труды БГТУ. Сер. 4, Принт- и медиатехнологии. 2023. № 2 (273). С. 32–39. DOI: 10.52065/2520-6729-2023-273-2-5.

A. A. Barkovich, A. V. Antonov
Minsk State Linguistic University

METHODOLOGICAL POTENTIAL OF SENTIMENT ANALYSIS: THE CORPUS ASPECT

This article is devoted to the aspects of the sentiment analysis procedure on the material of a text corpus. Sentiment analysis of text is traditionally focused on the evaluation of small speech artifacts; a typical example of processed text material is a blog or a messenger post. At the same time, the constant development of corpus format considers the implementation of the full range of automated speech processing capabilities in the referential context, including the linguopragmatic in its essence practice of sentiment analysis. These possibilities are in demand, but their full-fledged realization presupposes a preliminary approbation of existing tools of sentiment analysis and their adaptation for conducting research, in particular, in a fully compatible with computer-mediated communication corpus aspect – at the qualitatively innovative level. In the course of the corresponding analysis, the available methodological potential was systematized and characterized, and both ways of tactical actualization and the strategy of improving the practice of sentiment analysis were proposed. The scientific representation of the relevant problems will contribute to the fuller disclosure of the interdisciplinary potential of sentiment analysis and will increase the effectiveness and efficiency of its implementation.

Keywords: sentiment analysis, text corpus, text tonality, tonal vocabulary, evaluation, context.

For citation: Barkovich A. A. Methodological potential of sentiment analysis: the corpus aspect. *Proceedings of BSTU, issue 4, Print- and Mediatechnologies*, 2023, no. 2 (273), pp. 32–39. DOI: 10.52065/2520-6729-2023-273-2-5 (In Russian).

Введение. Сентимент-анализ, или анализ тональности текста, – востребованный инструмент не только изучения отдельно взятого текста, но и база для научного осмысления важных социокультурных трендов, определения репутации брендов, выявления отзывов о продуктах и решения многих других задач [1, 2, 3]. В настоящее время – с ростом доступности информации и активным использованием социальных сетей – количество текстов, создаваемых и распространяемых каждый день, значительно возросло. С учетом лавинообразного расширения информационного континуума и постоянного совершенствования компьютерных технологий существующие в сфере обработки естественного языка задачи постоянно усложняются. В настоящее время динамично растущая востребованность сентимент-анализа во многом обусловлена экспансией сети интернет. При этом очевидно, что средства обработки языковых данных требуют компетентного сопровождения и постоянного совершенствования [4].

С 2000-х гг. оценка тональности текста оказалась в фокусе научных изысканий [5], но к настоящему времени в сентимент-анализе осталось множество нерешенных проблем. Одной из них является уже практически классическая для компьютерной лингвистики дилемма между статистическим и основанным на правилах моделированием обработки естественного языка. И, если в плане совершенствования статистической методики все более-менее понятно, – принципиальная схема соответствующей модели предельно проста – то основанное на правилах моделирование подразумевает определенную сложность и перманентно реализуемый методологический потенциал. Собственно, базирующееся на правилах моделирование речевой практики – тема неисчепаемая. В данном контексте сложности обусловлены ориентацией процедуры сентимент-анализа на оперирование тональными словарями, составленными по непрозрачным критериям, и изучение компактных текстов интернет-дискурса. Корпусный формат в данной связи позволяет провести уверенную верификацию сложившихся шаблонов и определить перспективную модель совершенствования существующей практики [6].

Основная часть. Итак, сентимент-анализ как инструментарий вполне уместно рассматривать в рамках корпусной методики и парадигмы компьютерной лингвистики. Данный подход позволяет идентифицировать методологический потенциал сентимент-анализа как объект изучения и его корпусный аспект как предмет рассмотрения. Для проведения исследования был задействован корпус *Instrument-independent text corpus "Avatar: The Way of Water" (movie reviews)* – в качестве источника языкового материала [7]. Корпус текстов, выбранный для анализа, представляет собой набор текстов-отзывов на кинофильм *Avatar: The Way of Water* на английском языке. В корпус методом сплошной выборки были собраны все относящиеся к данному кинофильму отзывы, оставленные посетителями на сайте *imdb.com* в течение марта 2023 г. Объем данного корпуса составил *130 239* токенов в 96 текстах – это вполне репрезентативный корпусный формат. Также использовались ресурсы языка *Python* на базе программы *VADER* [8]. *Цель* работы – выявление лингвистического методологического потенциала сентимент-анализа для распространения релевантной процедуры на корпусный формат языкового материала. *Объект* – сентимент-анализ. Профильная *методология* – методологическая парадигма компьютерной лингвистики при задействовании статистической, математической, корпусной методик и инструментария теоретической и прикладной лингвистики.

Методологическая специфика. В лингвистическом аспекте сентимент-анализ, или оценка тональности текста, – выявление эмоциональной составляющей речевой продукции [9, с. 117]. В функциональном аспекте сентимент-анализ представляет собой процедуру определения эмоциональной окраски текста: положительной, отрицательной или нейтральной. Подобная практика выявления определению «… отношения человека к определенному объекту или теме. Отношение может означать оценочное суждение – не только его положительную или отрицательную направленность, но и характер его эмоциональности: разочарование, радость, гнев, печаль, волнение и т. д.» [10, p. 201]. Например, сентимент-анализ актуален для выявления мнений о здравоохранении, политике, бренде, образовании, социальной сети, личности и многих других актуальных социокультурных институтах и феноменах.

На сегодняшний день реализуются 3 основные модели сентимент-анализа: основанная на правилах, статистическая и гибридная [9, с. 119]. Пока наиболее результативная статистическая модель практически безальтернативно подразумевает методику машинного обучения для определения тональности текста. При этом релевантны методы на основе наивного Байесовского классификатора и методы с использованием нейронных сетей. Но не менее востребовано базирующееся на правилах моделирование, фактически

основанное на практике задействования так называемых тональных словарей [11, 12]. При использовании основанной на правилах модели «… система анализа тональности ищет в рассматриваемом тексте слова, имеющие эмоционально-оценочный заряд, и, применяя заложенные в ней правила, учитывающие отрицание и слова-усилители, вычисляет тональность всего текста» [13, с. 1108]. Данная модель относительно несложна на практике, и ее очевидным достоинством является низкий порог ресурсозатратности – в отличие от статистической и содержащей статистическую составляющую гибридной моделей. Эта специфика обеспечивает ее методологическое доминирование, особенно в контексте лингвистической работы, которая далеко не всегда предполагает основательную компетенцию в компьютерно-информационных технологиях, в частности, необходимую для задействования методики машинного обучения.

Конечно, основанная на правилах модель не учитывает контекст напрямую, но обладает существенным лингвистически обусловленным потенциалом развития. Определенный потенциал имеется и для опосредованного учета контекста (см., в том числе, ниже). Реализация такого потенциала, несомненно, будет востребована в технологическом совершенствовании сентимент-анализа в целом.

Базовый для данного исследования инструментарий *VADER* – типичная программа анализа тональности текста, работающая по модели, основанной на правилах [8]. Эта программа изначально была разработана для текстов на английском языке. С точки зрения компьютерной лингвистики аналитический английский язык – удобная языковая среда с минимумом, по сравнению с синтетическими языками, словоизменительных правил. Поскольку программа *VADER* написана на языке программирования *Python*, это позволяет адаптировать и совершенствовать ее возможности широкому кругу пользователей. Например, путем несложных манипуляций можно добавлять новые слова и выражения в базовый тональный словарь, подстраивать процедуру анализа под специфические особенности анализируемого текста: «Наш подход, названный *VADER*, представляет собой лексикон и инструмент анализа тональности на основе правил, который специально настроен на анализ тональности в социальных сетях» [14, p. 216]. Однако соответствующая «настройка», а конкретно «нормализация» (см. ниже), достаточно своеобразна и для анализа больших массивов данных потребует верификации и адаптации.

Каждой языковой единице в тональном словаре – базе данных программы *VADER* – присвоено значение тональности по шкале от «–4» до «+4». При этом «–4» – самый негативный показатель, а «+4» – самый позитивный. В дополнение к тональному рейтингу обычных в речи лексем и эмотиконов данный инструмент также учитывает другие маркеры, которые могут влиять на тональность текста, например слова-усилители, слова-отрицания, знаки препинания, капитализацию и др. В частности, ***слово-усилитель*** – это языковая единица, усиливающая эмоциональность связанной с ним лексемы (*very*, *extremely* и т. д.). ***Слово-отрицание*** – это языковая единица, которая частично или полностью меняет вектор тональности включающей ее синтаксической конструкции (в частности, *no* или *never*). Эффект данных слов-модификаторов зависит от их «расстояния» до слова, которое они дополняют. Более «далекие» модификаторы оказывают относительно меньшее воздействие на оригинальное слово. В используемой в нашем исследовании программе один модификатор рядом с базовым словом добавляет или вычитает «0,293» балла настроения предложения, в зависимости от того, является ли модифицируемое слово положительным или отрицательным. Второй модификатор рядом со связанным словом добавляет / вычитает 95% от «0,293», а третий – 90%.

При этом, хотя отдельные языковые единицы оцениваются в диапазоне значений от «–4» до «4», общая тональность предложений и более крупных речевых фрагментов традиционно измеряется в интервале от «–1» до «1». Для совместимости этих систем координат общая сумма оценок языковых единиц «нормализуется», чтобы адаптировать интервал значений «–4» – «4» к интервалу значений «–1» – «1». Это предусматривает использование так называемой ***функции нормализации***, которая наделяет бóльшим «весом» слова с высокими значениями тональности. Математически функция нормализации выглядит следующим образом: $x/\sqrt{(x^2 + \alpha)}$. Здесь x – сумма значений тональности слов, а α – параметр нормализации. Характерно, что с увеличением значения x оценка все больше приближается к «–1» или «1». Аналогично, если в анализируемом тексте много слов, то мы получаем оценку, близкую к абсолютным значениям «–1» или «1». Использование данной функции при сентимент-анализе вполне оправдано для обработки коротких текстов, таких как твиты или отзывы. Однако, как показало проведенное исследование, этот подход в корпусном масштабе накапливает погрешности частных оценок и существенно искажает оценку тональности для корпуса текстов.

Программная специфика. Рассмотрим вышеописанную специфику на примере [7]. Анализ тестового текста исследуемого корпуса с использованием программы *VADER* показал следующие результаты: *{'neg': 0.184, 'neu': 0.702, 'pos': 0.114, 'compound': -0.8342}* (рис. 1).

```
from nltk.sentiment.vader import SentimentIntensityAnalyzer

analyzer = SentimentIntensityAnalyzer()

text = "Actually cheap visual effects. Unexplainable average rating actually bec
ause of bad visual effects.Cgi " \
       "effects are just not convincing. The whole movie looks like 2007 compute
r game, in fact 2007 Crysis games " \
       "graphics were better. Secondary elements look cheap, especially floor\gr
ound it's just stretched textures " \
       "with no ambient occlusion, no normal or displacement map, no particles e
ffects what so ever. Movement and " \
       "dynamics design is also bad and dull. Everything is bright and colorful
though if you pause at any moment and " \
       "start deconstruction you can see how hopeless situation is. Most people
can't see visual effects problem " \
       "because of epileptic colorful blinking nature of this movie, just like a
6 month old baby laughing at shaking " \
       "keys. "
result = analyzer.polarity_scores(text)
print(result)

#Результат:
{'neg': 0.184, 'neu': 0.702, 'pos': 0.114, 'compound': -0.8342}
```

Рис. 1. Представление сентимент-анализа текста в программе *VADER*

Здесь можно отметить следующие особенности тональности текста:

– отрицательная оценка (‘*neg*’) со значением «0,184» свидетельствует о наличии отрицательных слов, но они не доминируют;

– нейтральная оценка (‘*neu*’) со значением «0,702» свидетельствует о доминировании нейтральных по своей сущности языковых единиц;

– положительная оценка (‘*pos*’) со значением «0,114» свидетельствует о наличии положительной оценки, но она слабо выражена;

– комплексный показатель (‘*compound*’) со значением «–0,8342» указывает на то, что текст содержит ярко выраженную отрицательную тональность.

Таким образом, учитывая присутствие в тексте эмоционально окрашенных лексических единиц, наличие усилителей, отсутствие слов-отрицаний и эмоционально окрашенных идиом, *VADER* сформировал оценку тональности для данного текста с комплексным показателем «–0,8342». Однако стоит отметить, что данный показатель – *compound* (см. рис. 1) – не является предельно отрицательным, что технически объяснимо ввиду не только присутствия большой группы нейтральных слов, о чем говорит соответствующее значение «0,702», но и наличия положительных слов, таких как *colorful* и *bright*, которые в определенной степени уравновешивают общую отрицательную тональность текста.

Характерно, что опора только на тональный словарь в данной ситуации не позволяет получить объективную оценку тональности текста в целом. Если ориентироваться на тональный словарь, то показатели положительно- («0,114»), нейтрально- («0,702») и отрицательно- («0,184») ориентированной лексики свидетельствуют о якобы нейтральной тональности текста. Разница между положительной и отрицательной оценкой составила лишь «0,07» в пользу отрицательного значения. Виртуально значение «–0,07» могло бы свидетельствовать о практически нейтральной оценке тональности текста, что никак не отражает его очевидно негативную тональность. Однако усложнение процедуры позволило программе сформировать ярко выраженное отрицательное и более близкое к действительному значение тональности текста «–0,8342». Этому способствовал учет дополнительных «металексических» параметров языковых единиц [15]. В частности, была задействована функция нормализации и учтены маркеры, влияющие на тональность текста, например:

– *эмоционально окрашенные лексические единицы:* в тексте присутствуют такие отрицательно окрашенные слова, как *cheap*, *bad*, *dull*, *hopeless*, *problem* и *epileptic*; несомненно, они внесли свой вклад в отрицательную тональность всего текста;

– *усилители:* в тексте есть усилители, такие как *just*, например *just not convincing*; *actually*, например, *actually cheap*; *unexplainable*, например, *unexplainable average rating*, которые усилили отрицательную тональность семантически зависимых от них языковых единиц.

Вместе с тем интуитивно-понятийная оценка данного текста не полностью соответствует данным компьютерной процедуры (рис. 2).

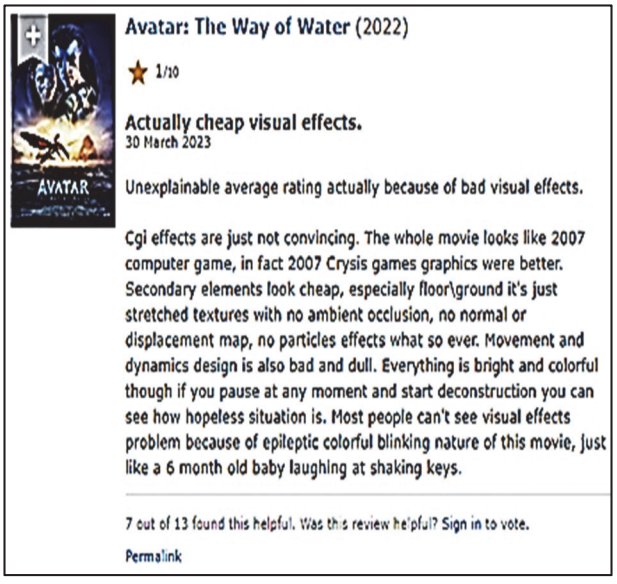

Рис. 2. Отзыв пользователя на сайте *imdb.com*

Здесь можно видеть, что данный текст, по мнению его автора, должен был отразить максимально критическое и негативное мнение об анализируемом фильме. Автор текста оценил свои впечатление значением «1» из «10» возможных баллов.

И подобная оценка, на самом деле, приблизительно должна была бы соответствовать значению «–1», а не «–0,8342»: в результате искажение результата составило более 16%. Эта разница вполне может быть отнесена на счет лингвопрагматических особенностей идиостиля автора текста. Например, вполне резонно предположить, что автор по каким-то причинам не смог или не захотел отразить в тексте отзыва всю глубину своего разочарования качеством анализируемого медийного продукта.

Корпусная специфика. Анализ корпуса по-новому репрезентирует отмеченные на уровне его тестового фрагмента неточности сентимент-анализа. Так, при анализе объектного корпуса текстов с помощью базового программного кода *VADER* было получено абсолютное значение «1». Это можно объяснить тем, что при увеличении количества анализируемых данных сумма оценок тональности этих слов также увеличивается, усиливая заложенное в функции нормализации «стремление» к целым значениям. Однако логичное с точки зрения эстетики «совершенство» функции в данной ситуации приводит, как оказывается (см. ниже), к весьма значительному искажению фактических данных.

Для того чтобы верифицировать данный результат, можно проанализировать каждый текст по отдельности, затем сложить все полученные комплексные значения и разделить их на количество проанализированных текстов. По результатам выполненной таким образом «ручной» верификации общая тональность корпуса была оценена значением «0,4518», что разительно отличается от данных совокупной оценки со значением «1». Выявленная дифракция, искажение суммарного показателя оценок, превысила 50% – «1» вместо «0,4518».

Аналогичный результат может быть получен и «автоматически», путем совершенствования программного обеспечения (рис. 3).

```
import nltk
from nltk.sentiment.vader import SentimentIntensityAnalyzer

# Загрузка корпуса текстов
with open('avatar corpus.txt', 'r', encoding='utf-8') as f:
    corpus = f.read().split('#\n')

# Инициализация анализатора тональности VADER
analyzer = SentimentIntensityAnalyzer()

# Анализ тональности всего корпуса
compound_scores = []
for text in corpus:
    score = analyzer.polarity_scores(text)
    compound_scores.append(score['compound'])

# Расчет общей тональности корпуса
average_compound_score = sum(compound_scores) / len(compound_scores)

print(f"Общая тональность корпуса: {average_compound_score}")

#Результат:
Общая тональность корпуса: 0.45180736842105274
```

Рис. 3. Фрагмент программного кода для выявления оценки тональности корпуса текстов

Полученное значение на самом деле также является положительным, но более близким к нейтральному.

Еще одним этапом верификации может служить средневзвешенное представление субъективной оценки тональности текстов самими авторами (рис. 4).

Рис. 4. Статистические данные относительно распределения отзывов разной тональности в корпусе (по оценкам пользователей)

Отнесенные на основании указаний авторов к 10 разным рейтинговым категориям оценки свидетельствуют об очень близкой к нейтральной тональности текстов всего корпуса: около половины отзывов положительны или частично положительны, остальные полностью или частично отрицательны. При обобщении оценок тональности анализируемых текстов, сделанных самими пользователями, хорошо заметно, что референтные данные далеки от симметрии, хотя и распределены достаточно равномерно между крайне отрицательной оценкой «1» и крайне положительной «10» (рис. 4). При этом отмечается большее количество текстов с полярно высокими оценками («7» и «9» баллов) и высокий рейтинг оценок «1», «3» и «5». Общий балл пользовательских оценок оказался равен «5,5», что, в общем, подтвердило умеренно-позитивную тональность всего корпуса.

Вместе с тем оба полученных значения – и «0,4518», и «1» существенно превысили субъективные оценки авторов – предсказуемым было бы значение, приблизительно соответствующее «0,1» (это приблизительный эквивалент значения «5,5»).

Необходимо отметить, что не менее очевидны погрешности при совокупной оценке текстов, содержащих разнонаправленную (сложную) тональность. Компьютерный сентимент-анализ не всегда корректно определяет тональность текста, особенно если в нем присутствуют комбинации разнонаправленных эмоций (позитивных, нейтральных и негативных), присущих элементам исследуемого материала.

Таким образом, рассмотрение специфики проведения сентимент-анализа для больших массивов текста показало, что процедура использования существующих программных средств в данном аспекте нуждается в существенной адаптации и совершенствовании.

Очевидно, что если для небольшого текста получаемые данные вполне могут быть откорректированы в процессе постобработки «вручную», то на больших массивах текста исправлять погрешности – уже на порядок большие – непродуктивно и нецелесообразно. Оптимальным решением тут будет предметная адаптация и совершенствование программного обеспечения.

Полученные и представленные результаты исследования перспективны и предполагают их дальнейшую интерпретацию.

Заключение. Проведенный в корпусном ключе анализ тональности текста показал, что сформированные стереотипы практики сентимент-анализа методологически обладают существенным, но недостаточно научно осмысленным и апробированным потенциалом развития. Особенно выразительно соответствующая специфика проявляется на высокорепрезентативном материале и материале, содержащем сложную разнонаправленную семантику тональности (эмоциональности). В целом выявленный потенциал совершенствования процедуры сентимент-анализа может быть реализован в четырех аспектах: трех тактических и одном стратегическом. На уровне тактических мероприятий проблематика совершенствования процедуры лежит, во-первых, в плоскости качественной настройки языкового инструментария посредством коррекции параметров слов-маркеров. Во-вторых, целесообразна модификация самого программного «кода» для повышения совместимости механизмов разносистемного шкалирования. В-третьих, традиционно актуально максимальное количественное расширение самого тонального словаря: это в любом случае обеспечит улучшение репрезентативности материала. Проведенное исследование подтвердило данную актуальность: далеко не все языковые единицы проанализированных текстов были автоматически оценены по причине их отсутствия в тональном словаре. В-четвертых, стратегически актуальным является учет контекста для полноценной интерпретации результатов. Этот недостаток присущ всем программам автоматической обработки текстов на естественном языке, и изжить его при современном уровне развития технологий можно только путем подключения статистических ресурсов – частности, посредством методики машинного обучения. В том числе задействование методики машинного обучения является эффективным при использовании программ, работающих по основанной на правилах модели сентимент-анализа. Методологически рациональным в данном аспекте представляется совершенствование процедуры сентимент-анализа посредством задействования гибридной модели. Гибридная модель, собственно, и подразумевает сочетание возможностей тональных словарей с технологиями машинного обучения. Опыт проведенного исследования полностью подтверждает данный тезис в контексте обработки больших объемов языковых данных – корпусов.

Список литературы

1. Майорова Е. В. О сентимент-анализе и перспективах его применения // Социальные и гуманитарные науки. Отечественная и зарубежная литература. Сер. 6: Языкознание. 2020. № 4. С. 78–87.

2. Семина Т. А. Анализ тональности текста: современные подходы и существующие проблемы // Социальные и гуманитарные науки. Отечественная и зарубежная литература. Сер. 6: Языкознание. 2020. № 4. С. 47–64.

3. Beigi G., Hu X., Maciejewski R., Liu H. An overview of sentiment analysis in social media and its applications in disaster relief // Sentiment analysis and ontology engineering: an environment of computational intelligence / eds. W. Pedrycz, S.-M. Chen. Cham: Springer, 2016. P. 313–340.

4. Araque O. Zhu G., Iglesias C. A. A semantic similarity-based perspective of affect lexicons for sentiment analysis // Knowledge-Based Systems. 2019. No. 165. P. 346–359.

5. Liu B. Sentiment analysis and opinion mining // Synthesis lectures on human language technologies. 2012. Vol. 5, no. 1. P. 1–16.

6. Баркович А. А., Ван Ц. Лингвистические корпусы китайского языка: функциональный аспект // Вестник МГЛУ. Сер. 1. Филология. 2015, № 5 (78). С. 105–113.

7. Antonov A. V., Barkovich A. A. Instrument-independent text corpus “Avatar: The Way of Water” (movie reviews). URL: https://drive.google.com/file/d/1Y7V15sEmH0NI6rAFSbyIjAXd0wYY5h/view (accessed: 20.05.2023).

8. VADER Sentiment Analysis: A Complete Guide, Algo Trading and More. URL: https://blog.quantinsti.com/vader-sentiment (accessed: 20.05.2023).

9. Баркович А. А. Сентимент-анализ: лингвистический потенциал регламентации предобработки // Виртуальная коммуникация и социальные сети. 2023. Т. 2, № 3. С. 116–123.

10. Mohammad S. M. Sentiment analysis: Detecting valence, emotions, and other affectual states from text // Emotion measurement. Elsevier. 2016. P. 201–237.

11. Пазельская А. Г., Соловьев А. Н. Метод определения эмоций в текстах на русском языке // Компьютерная лингвистика и интеллектуальные технологии: ежегодная Междунар. конф. «Диалог», Бекасово, 25–29 мая 2011 г. М., 2011. Вып. 10. С. 510–522.

12. Taboada M., Brooke J., Tofiloski M., Voll K., Stede M. Lexicon-based methods for sentiment analysis // Computational linguistics. 2011. No. 37(2). P. 267–307.

13. Кулагин Д. И. Открытый тональный словарь русского языка КартаСловСент // Компьютерная лингвистика и интеллектуальные технологии: ежегодная Междунар. конф. «Диалог», Москва, 16–19 июня 2021 г., М., 2021. Вып. 20. С. 1106–1119.

14. Hutto C. J., Gilbert E. VADER: A parsimonious rule-based model for sentiment analysis of social media text // Proceedings of the 8th international conference on weblogs and social media (ICWSM), 2014, May. Ann Arbor, Michigan USA: PKP Publishing Services Network. 2014. P. 216–225.

15. Баркович А. А. Компьютерно-опосредованная коммуникация: потенциал металексической значимости // Ученые записки Петрозаводского государственного университета. Общественные и гуманитарные науки. 2015, № 7 (152). С. 38–43.

References

1. Mayorova E. V. On sentiment analysis and prospects for its application. *Sotsial'nyye i gumanitarnyye nauki. Otechestvennaya i zarubezhnaya literatura* [Social and human sciences. Domestic and foreign literature], series 6, Linguistics, 2020, no. 4, pp. 78–87 (In Russian).

2. Semina T. A. Sentiment analysis: modern approaches and existing problems. *Sotsial'nyye i gumanitarnyye nauki. Otechestvennaya i zarubezhnaya literatura* [Social and human sciences. Domestic and foreign literature], series 6, Linguistics, 2020, no. 4, pp. 47–64 (In Russian).

3. Beigi G., Hu X., Maciejewski R., Liu H. An overview of sentiment analysis in social media and its applications in disaster relief. *Sentiment analysis and ontology engineering: an environment of computational intelligence* / eds. W. Pedrycz, S.-M. Chen, Cham: Springer, 2016, pp. 313–340.

4. Araque O. Zhu G., Iglesias C. A. A semantic similarity-based perspective of affect lexicons for sentiment analysis. *Knowledge-Based Systems*, 2019, no. 165, pp. 346–359.

5. Liu B. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 2012, vol. 5, no.1, pp. 1–16.

6. Barkovich A. A., Wang, Q. Linguistic corpora of the Chinese language: a functional aspect. *Vestnik MGLU* [Bulletin of the MSLU], series 1, Philology, 2015, no. 5 (78), pp. 105–113 (In Russian).

7. Antonov A. V., Barkovich A. A. Instrument-independent text corpus "Avatar: The Way of Water" (movie reviews). Available at: https://drive.google.com/file/d/1Y7V15sEmH0NI6rAFSbyIjAXd0wYY5h/view (accessed 20.05.2023).

8. VADER Sentiment Analysis: A Complete Guide, Algo Trading and More. Available at: http://www.multitran.ru (accessed 20.05.2023).

9. Barkovich A. A. Sentiment Analysis: Linguistic Potential of Preprocessing Regimentation. *Virtual'naya kommunikatsiya i sotsial'nyye seti* [Virtual Communication and Social Networks], 2023, no. 2(3), pp. 116–123 (In Russian).

10. Mohammad S. M. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. *Emotion measurement*, Elsevier, 2016, pp. 201–237.

11. Pazel'skaya A. G., Solov'ev A. N. A method for determining emotions in texts in Russian. *Komp'yuternaya lingvistika i intellektual'nyye tekhnologii: ezhegodnaya Mezhdunarodnaya konferentsiya "Dialog"* [Computational Linguistics and Intelligent Technologies: materials of the annual International Conference "Dialogue"], 2011, issue 10, pp. 510–522 (In Russian).

12. Taboada M., Brooke J., Tofiloski M., Voll K., Stede M. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 2011, no. 37(2), pp. 267–307.

13. Kulagin D. I. Open Tonal Dictionary of the Russian Language KartaSlovSent. *Komp'yuternaya lingvistika i intellektual'nyye tekhnologii: materialy ezhegodnoy Mezhdunarodnoy konferentsii "Dialog"* [Computational Linguistics and Intelligent Technologies: materials of the annual International Conference "Dialogue"], 2021, issue 20, pp. 1106–1119 (In Russian).

14. Hutto C. J., Gilbert E. VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the 8th international conference on weblogs and social media (ICWSM)*, 2014, May, Ann Arbor, Michigan USA: PKP Publishing Services Network, 2014, pp. 216–225.

15. Barkovich A. A. Computer-mediated communication: the potential of metalexical significance. *Uchenyye zapiski Petrozavodskogo gosudarstvennogo universiteta. Obshchestvennyye i gumanitarnyye nauki* [Scientific notes of Petrozavodsk State University. Social and human sciences], 2015, no. 7 (152), pp. 38–43 (In Russian).

Информация об авторах

Баркович Александр Аркадьевич – доктор филологических наук, доцент, заведующий кафедрой информатики и прикладной лингвистики. Минский государственный лингвистический университет (220034, г. Минск, ул. Захарова, 21, Республика Беларусь). E-mail: barkovichaa@gmail.com

Антонов Андрей Владимирович – студент. Минский государственный лингвистический университет (220034, г. Минск, ул. Захарова, 21, Республика Беларусь). E-mail: andrey56735472@gmail.com

Information about the authors

Barkovich Aliaksandr Arkad'yevich – DSc (Philology), Associate Professor, Head of the Department of Informatics and Applied Linguistics. Minsk State Linguistic University (21, Zakharova str., 220034, Minsk, Republic of Belarus). E-mail: barkovichaa@gmail.com

Antonov Andrey Vladimirovich – student. Minsk State Linguistic University (21, Zakharova str., 220034, Minsk, Republic of Belarus). E-mail: andrey56735472@gmail.com

Поступила 26.06.2023