УДК 81’33

А. А. Баркович, проф. д-р филол. наук
(БГТУ, г. Минск)

МОДЕЛИ АНАЛИЗА ТОНАЛЬНОСТИ ТЕКСТА

Аспекты научного осмысления речевой практики сегодня существенно зависят от эмпирической методологии нового поколения – компьютерно-информационной. Компьютерная локализация современной речевой практики и ее информационная актуализация, конечно же, обусловливает и приоритетность программной обеспеченности аналитической парадигмы. И в числе прочих достижений исследовательской мысли прикладного характера необходимо констатировать неослабевающий интерес сферы *IT* к такому продуктивному и перспективному инструментарию как анализ тональности текста. В лингвистическом контексте ***анализ тональности текста***, или сентимент-анализ, – выявление и обобщение лингвопрагматических особенностей речевой практики [1]. Данная методика применяется не только для рассмотрения явной субъективной специфики речевой продукции, – в значительной мере ее популярность обусловлена эффективностью задействования сентимент-инструментария для анализа и описания неочевидного потенциала текста.

Модели исследовательской практики формируются и культивируются на статистически значимой и структурно развитой эмпирической базе. Их наличие, как правило, свидетельствует о накоплении определенного опыта работы в той или иной сфере деятельности. И, несмотря на известную инерцию теоретического сопровождения сферы *IT*, соответствующие шаблоны уже достаточно уверенно можно идентифицировать и в практике сентимент-анализа. Выбор модели – важный процедурный аспект оценки тональности текста. При обобщении сложившихся стереотипов обработки текстов на предмет определения их тональности характерным образом реализуются три основных ***модели***:

– основанная на правилах;

– статистическая и

– гибридная.

Данный аспект исследовательской и «производственной» деятельности характеризуется неустоявшейся и фрагментарной рефлексией: несмотря на высокую степень регламентации компьютерно-опосредованной коммуникации наблюдается некоторый разнобой в подходах к классификациям моделей сентимент-анализа. Достаточно

популярен следующий подход: «Мы идентифицируем методы обнаружения настроений как принадлежащие к одной из трех категорий, каждая из которых имеет свои достоинства и недостатки: словарные методы; методы машинного обучения с учителем и методы машинного обучения без учителя» [2]. Есть немало сторонников еще одного подхода, базирующегося на трехэлементной схеме: согласно ему выделяют такие модели как машинное обучение, словарная и гибридная модели [3]. Видимым недостатком вышеупомянутых подходов является отсутствие в них места для фактически широко используемой «основанной на правилах модели». Вместе с тем, востребованность данной «классической» для связанной с компьютерными технологиями деятельности модели так или иначе подтверждается косвенно как отечественными, так и зарубежными исследователями [4]. Собственно, на практике «…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» [5, с. 1106]. *Правилом* анализа здесь является система индексов, в соответствии с которой может быть оценено каждое соотношение языковых единиц (лексем, словосочетаний, *n*-грамм). С учетом этого за лексемами закрепляются определенные цифровые значения, являющиеся дискретными репрезентациями языковых знаков в формализованной среде. Ориентация на уровень лексем здесь объяснима, с одной стороны, наличием обработанных массивов лексем в составе разнообразных словарей, а, с другой стороны, опорой на них в лингвистической традиции – слово традиционно находится в фокусе языкознания.

Основанная на правилах модель (англ. *rule-based model*) подразумевает использование в ней неких конвенциональных положений, или правил. Все это, конечно, применимо к такому конвенционально обусловленному объекту как язык и его реализация – речь. Таким образом, путем словарной агрегации правил речевого функционирования сентимент-анализ продолжает лингвистические традиции и оказывается совместимым со всем методологическим арсеналом языкознания. При этом, тем самым, он оказывается подвержен и системным недостаткам классической лингвистики, в числе которых и ориентация при анализе речевой практики на «словный», или лексический, формат. Конечно, в «словарях» отражаются человеческие знания в виде правил, однако для анализа живой речи подобных «костылей» очень часто оказывается недостаточно. Сентимент-анализ в определенной степени позволяет преодолеть условности словарной репрезентации языка посредством собственного статистического инстру-

ментария. Здесь тональность текста, или «документа» на компьютерном сленге, определяется по совокупности индексов оценочной лексики, представленной в специализированных словарях.

Словари с «оценочными» индексами имеют несколько параллельных номинаций, используемых в соответствии с терминологическими приоритетами исследователей: *семантические тезаурусы*, *тональные словари*, *лексиконами* – есть и другие вариант. Пожалуй, наиболее востребован термин *тональный словарь*, являющийся наиболее семантически «прозрачным» в данном контексте. Соответствующая лексикографическая работа выполнена для разных языков. По понятным причинам наиболее популярной языковой системой здесь оказалась английская: собственно, компьютерно-опосредованная коммуникация в весьма значительной мере англоязычна. Такими англоязычными источниками для сентимент-анализа в словарном формате являются *LIWC*, *Opinion Lexicon* и др. Впрочем, для русского языка в данной связи тоже сделано немало: вполне актуальны *РуСентиЛекс*, *Карта слов* и др. Тональные словари включают «датасеты», «списки слов», «списки слов только с положительными и отрицательными аннотациями», «списки слов со скалярными числовыми значениями» и т. д. [см. напр., 6, с. 348]. Подобные датасеты являются достаточно информативным подспорьем не только в плане системной репрезентации данных о задействованных в тексте языковых единиц, но также позволяют делать аргументированные выводы о полярной ориентации текста и степени выраженности в нем эмоциональной оценки.

Альтернативным основанной на правилах модели инструментарием для оценки тональности текста является «статистическая модель». ***Статистическая модель*** (англ. *statistical model*) подразумевает учет для презентации объекта лишь его формальных показателей – без учета конвенционального «смыслового» содержания объекта, в данном контексте речи. Задействование для сентимент-анализа – безотносительно характерного при этом камуфляжа под некие абстракции типа «обучения» – сегодня актуализируется на фоне все упрощающейся доступности «больших данных» (англ. *big data*). практически безальтернативно базируется на методике *машинного обучения*. ***Машинное обучение*** – методика задействования для компьютерных вычислений больших массивов обработанных предварительно однотипных данных для решения алгоритмических задач. Смысл «обучения» здесь – в агрегации статистически значимой совокупности уже состоявшихся однотипных решений, на основе чего формируется алгоритм решения подобных задач в дальнейшем. Такое «обучение» –

не более, чем метафора: компьютер – машина. Тем не менее, в машинном обучении позиционируются как методически самостоятельные два основных его типа: «без учителя» (англ. *unsupervised learning*) или «с учителем» (англ. *supervised learning*). По уже сложившемуся стереотипу эти шаблоны дифференцируются как существенно отличающиеся, однако, с методической точки зрения, это, конечно, варианты все той же статистической модели. Характерно, что исследователи, отмечают необходимость дальнейшей дифференциации форматов машинного обучения: пакетное и динамическое обучение, обучение на основе образцов или на основе моделей, обучение с подкреплением и др. [7]. Это все возможно, но пока не подкреплено широкомасштабным задействованием. Так или иначе, статистический подход используется в прикладной лингвистике не только для оценки тональности текста: соответствующая группа методов – с развитием информационно-компьютерных технологий – приобретает особую научную актуальность [8]. Принципиально, задействование статистической модели для сентимент-анализа ничем не отличается от форматов использования этой модели при решении самых разнообразных задач *NLP*, или автоматизированной обработки естественного языка.

Гибридная модель (англ. *hybrid model*) подразумевает синтетическую общность уже сформированных и апробированных моделей. В сентимент-анализе в качестве гибридного шаблона используется та или иная пропорция двух вышеупомянутых моделей – основанной на правилах и статистической. Гибридный подход оказывается все более популярным в последнее время. При этом зачастую подобное усложнение базовых компонентов предлагается рассматривать как нечто принципиально новое. О принципиальной новизне в данной связи, видимо, можно будет говорить при достижении некоего симбиотического слияния базовых моделей, что пока не просматривается. Вместе с тем, доминирование в синтетическом формате той или иной модели не означает ее исключительности и может сочетаться с другими подходами к анализу тональности текста.

В сентимент-анализе приоритет использования какой-либо модели зависит от целой совокупности факторов: характера данных, возможностей их обработки, наличия программных ресурсов, компетенции исследователя и ряда других. Теоретически есть все основания предполагать, что – при должной подготовке инструментария – основанная на правилах модель обладает большим потенциалом, чем статистическая модель, однако практика не так категорична. Усредненные показатели эффективности данных моделей вполне сопоставимы.

При этом на практике результативность статистической модели сентимент-анализа зачастую оказывается более высокой.

Таким образом, современная практика автоматизированной обработки естественного языка является сложной и многоаспектной. Несмотря на динамичное развитие сферы информационных технологий в прикладном русле, принципиальные основы такой деятельности попрежнему тесно коррелируют с лингвистической системой знаний: в методологическом контексте оценка тональности текста основывается на лингвистически обусловленной парадигматике моделей.

ЛИТЕРАТУРА

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

2. Reagan A. J. et al. Sentiment analysis methods for understanding large-scale texts: a case for using continuum-scored words and word shift graphs // EPJ Data Science. 2017, vol. 6 (28). pp. 1–21.

3. Poria S., Hazarika D., Majumder N., Mihalcea R. Beneath the tip of the Iceberg: Current challenges and new directions in sentiment analysis research // IEEE Transactions on Affective Computing. 2020, vol. 14. pp. 108–132.

4. Barkovich A. Informational Linguistics: The New Communicational Reality. Newcastle upon Tyne: Cambridge Scholars Publishing, 2020. 271 p.

5. Кулагин, Д. И. Открытый тональный словарь русского языка КартаСловСент / Д. И. Кулагин // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог». – Москва: Изд-во РГГУ. –2021, вып. 20. – С. 1106–1119.

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

7. Géron A. Hands-On Machine Learning with Scikit-Learn and TensorFlow. Boston: O'Reilly Media, Inc. 2017. 574 p.

8. Баркович, А. А. Интернет-дискурс: метаязыковые модели практики / А. А. Баркович // Вестник Волгоградского государственного университета. – Серия 2, Языкознание. 2015, № 5 (29). – С. 171–183.