ПРАКТИКИ СБОРА И АНАЛИЗА ФОРМАЛИЗОВАННЫХ ДАННЫХ

DOI: 10.19181/4m.2023.32.2.1
EDN: CRGFLH

СЕНТИМЕНТ-АНАЛИЗ КАК МЕТОД ИССЛЕДОВАНИЯ ИНФОРМАЦИОННОЙ ПОВЕСТКИ И ОБЩЕСТВЕННОГО МНЕНИЯ (НА ПРИМЕРЕ СМИ И СОЦИАЛЬНЫХ СЕТЕЙ КНР)

Анташева Мария Сергеевна
Национальный исследовательский университет «Высшая школа экономики», Москва, Россия
msantasheva@hse.ru
ORCID: 0000-0002-5255-8773

Лобанова Полина Александровна
Национальный исследовательский университет «Высшая школа экономики», Москва, Россия
plobanova@hse.ru
ORCID: 0000-0002-9878-9390

Исаева Юлия Камаловна
Национальный исследовательский университет «Высшая школа экономики», Москва, Россия
yisaeva@hse.ru
ORCID: 0000-0002-7974-8294

Сабидаева Елизавета Алексеевна
Национальный исследовательский университет «Высшая школа экономики», Москва, Россия
esabidaeva@hse.ru
ORCID: 0000-0001-9115-2285

Пиекалнитс Анна Сергеевна
Национальный исследовательский университет
«Высшая школа экономики», Москва, Россия
apiekalnits@hse.ru
ORCID: 0000-0003-0585-5350

Логинова Ирина Владимировна
Национальный исследовательский университет
«Высшая школа экономики», Москва, Россия
iloginova@hse.ru
ORCID: 0000-0002-3376-2728

Для цитирования: *Анташева М. С., Лобанова П. А., Исаева Ю. К., Сабидаева Е. А., Пиекалнитс А. С., Логинова И. В.* Сентимент-анализ как метод исследования информационной повестки и общественного мнения (на примере СМИ и социальных сетей КНР) // Социология: методология, методы, математическое моделирование (Социология:4М). 2023. № 57. С. 7–41. DOI: 10.19181/4m.2023.32.2.1. EDN: CRGFLH.

Информационная повестка, транслируемая китайскими медиаресурсами, является источником актуальных данных о мнении общества в отношении ключевых вопросов социального благосостояния. Вследствие технических особенностей организации китайских веб-сайтов и необходимости привлечения дополнительных ресурсов для автоматической обработки (парсинга) текстов на китайском языке, данная тематика не представлена достаточно широко в отечественных и зарубежных исследованиях. Целью настоящей работы является демонстрация методологии и результатов оценки общественного мнения на примере данных, собранных из китайских СМИ и социальных сетей, на основе обученной модели сентимент-анализа текстовых данных на китайском языке. При помощи ML-модели был проведен сравнительный анализ контента на китайском языке по проблематике развития городской инфраструктуры за период 2020–2022 гг. Результаты представлены в формате диаграмм распределения сентимента на основе данных СМИ и социальных сетей по месяцам за 2-летний период. Выявлено, что уровень сентимента

значительно отличается в зависимости от типа источника данных. Определено устойчивое преобладание позитивного сентимента в СМИ и негативного – в социальных сетях, что может объясняться различиями в составе авторов текстов, ограничениями, накладываемыми на публикуемый в источниках контент, а также разными целями использования ресурсов пользователями.

Ключевые слова: сентимент-анализ, эмоциональная окраска текстов, развитие городской инфраструктуры, общественное мнение, китайский язык, машинное обучение, интеллектуальный анализ данных, социальные сети

Благодарности: Статья подготовлена в рамках гранта, предоставленного Министерством науки и высшего образования Российской Федерации (№ соглашения о предоставлении гранта: 075-15-2022-325).

Введение

Постоянно увеличивающийся поток информации поспособствовал появлению и развитию прорывных технологий обработки данных [1]. За счет того, что значительные объемы новых данных появляются и потребляются ежедневно, обработать всю входящую информацию традиционными экспертными методами становится невозможно. Так, по данным международной консалтинговой компании McKinsey, 90% всех существующих в цифровом формате данных были созданы в течение последних двух лет, из них только 1% были проанализированы [2]. В то же время развитие методов текст-майнинга, семантического и сентимент-анализа текстов вышло на уровень, позволяющий оперативно и надежно применять данные методы для получения обоснованных выводов.

Актуальность рассмотрения китайского сегмента интернета в настоящем исследовании обусловлена рядом факторов, в том числе связанных с динамикой развития использования сети вну-

три страны. Согласно 47-му «Статистическому отчету о развитии интернета в КНР», численность интернет-пользователей КНР на конец 2020 г. превысила 989 млн человек, а уровень покрытия сетью интернет составлял 70,4% [3], что создает прочную обширную базу для создания контента интернет-ресурсов различных типов. К сравнению, численность интернет-пользователей США за аналогичный период составила 299,8 млн человек [4]. За счет высокой степени интернет-покрытия и большого количества онлайн-пользователей информационное пространство Китая стало актуальным источником для исследования публичного мнения, в том числе с помощью метода сентимент-анализа.

Сентимент-анализ представляет собой тип обработки естественного языка, целью которого является анализ мнений, настроений, оценок, суждений и эмоций человека по отношению к некому объекту [5]. Метод подразумевает присвоение оценки эмоциональной окрашенности тексту (фрагменту текста) на основе используемой в нем лексики. Оценка может быть бинарной (позитивная или негативная, субъективная или объективная), также могут применяться шкалы, где степень позитивности/негативности утверждений определяется числовым значением в заданных пределах [6]. Подходы к сентимент-анализу с точки зрения техник присвоения оценки можно разделить на три группы: подходы на основе машинного обучения, на основе словарей [7], также выделяют ручное аннотирование [8]. Сентимент-анализ является универсальным методом с точки зрения его тематической применимости.

Процедуру анализа можно обобщить следующим образом. На начальных этапах исследователями определяются объекты, отношение к которым будет анализироваться, выбирается подход, осуществляется сбор и обработка текстов, определяется способ оценивания. Далее многое зависит от выбранного подхода – при машинном обучении, как правило, необходимо обучение или дообучение соответствующей модели, помимо исполнения технических нюансов требующее разметку обучающей, валидационной и трени-

ровочной выборок; при словарном подходе необходимо составление перечня слов с присвоенными им оценками; при ручной аннотации происходит разметка анализируемых единиц аннотаторами. Затем разработанные модели применяются непосредственно к подготовленному набору данных. На завершающем этапе полученные результаты обобщаются и интерпретируются исследователями.

Данный метод широко применяется в течение последних 10 лет как для двухчастной, так и для трехчастной классификации текстов [9] в различных сферах: медицина [10; 11], авиация [12], информационные технологии [13], ритейл [14] и т.д. Существует множество бенчмарков, позволяющих эффективно оценивать разрабатываемые модели для проведения автоматизированного сентимент-анализа [15].

Конструктная валидность применения сентимент-анализа на основе машинного обучения рядом авторов оценивается выше, чем более традиционные подходы, – использование словарей или ручное аннотирование текстов. Так, было выявлено, что по таким метрикам, как accuracy (доля всех правильных ответов), precision (точность) и recall (полнота), алгоритмы машинного обучения превосходят словарные подходы [16]. Рядом исследователей отмечается, что точность работы алгоритмов машинного обучения варьируется от 65,4 до 77,5%, в то время как словарных методов – от 50 до 60,4% [17]. Кроме того, подход на основе словарей требует учета контекста и различий в употреблении лексики в разных исследовательских областях [18]. Что касается ручного аннотирования, его качество может существенно зависеть от внутреннего состояния разметчиков (может сказываться усталость и др. факторы) [8]. Также при высокой гранулярности классов наблюдается тенденция к уменьшению согласованности между разными аннотаторами [17].

С точки зрения исследований эмоциональной окраски текстов, особый исследовательский интерес представляют общественные онлайн-площадки, публикующие контент типа «вопрос – ответ» или «текстовый пост – пользовательский комментарий». К таким

соцсетям, доступным для автоматического сбора данных, относится, например, Twitter. Результаты исследования контента Twitter на основе сентимент-анализа применялись для изучения демографических трансформаций [19], для анализа риторики политических дебатов [20], для определения превалирующего в обществе отношения к вакцинации против COVID-19 [21] и др.

Что касается сентимент-анализа публикаций китаеязычных социальных площадок, отдельно можно отметить работу с китайским микроблогом Sina Weibo (微博), применявшимся для изучения отношения жителей КНР к последствиям стихийных бедствий и их урегулированию [22], определения реакции людей на динамику цен на недвижимость [23], исследования мнения граждан КНР относительно ограничений, введенных в связи с пандемией COVID-19 [24], и др.

Показатели эмоциональной окраски текстов на китайском языке выражаются как в наличии слов-маркеров, указывающих на качественную характеристику оцениваемого объекта, так и в их совстречаемости. Так, слова-маркеры одной и той же эмоциональной окраски скорее всего будут часто встречаться в одних и тех же текстах, тогда как противоположные по сентименту слова в одних текстах скорее всего будут встречаться реже [25]. Также стоит отметить, что для китайского языка характерно и наличие инверторов тональности (Valence shifter indicator) – морфем или слов, изменяющих эмоциональную тональность предложения [25].

Целью настоящего исследования является демонстрация методологии и результатов оценки общественного мнения на примере данных, собранных из китайских СМИ и социальных сетей, на основе специально обученной модели сентимент-анализа текстовых данных на китайском языке. Объектом исследования выступает общественное мнение жителей КНР в отношении развития городской инфраструктуры.

В качестве предмета исследования был выбран перечень крупных и доступных для автоматизированного анализа китайских

медиаресурсов: Xinhua (新华网), People.com (人民网), China Times (华夏时报网), Haiwai Net (海外网) и других, представляющих собой веб-сайты крупных новостных агентств, регулярно публикующих контент по широкому спектру тематик, в основном ориентированных на общество, бизнес, технологическое и инновационное развитие КНР и мира. Популярный китайский веб-сайт Zhihu (知乎)[1] был выбран в качестве второго типа источников исследования – социальных сетей. Zhihu – это социальная платформа, основанная в 2010 г. и публикующая контент формата «вопрос – ответ», где интересующие вопросы и ответы на них могут размещать все зарегистрированные пользователи. Всего, по данным за первый квартал 2022 г., число активных пользователей Zhihu в месяц составляет более 100 млн человек [26]. При этом 22% всех зарегистрированных в социальной сети пользователей составляют люди до 24 лет, 61% – от 25 до 35 лет, 14% – от 36 до 40 лет [26].

Для исследования были проанализированы данные за период с 01.01.2020 по 01.08.2022, в качестве метода расчета тональности было выбрано определение среднего сентимента за месяц. Были построены диаграммы распределения сентимента, отражающие тональность предложений по искомой тематике для двух типов источников.

Алгоритм исследования включает в себя обучение специальной модели сентимент-анализа текстовых данных на китайском языке, в том числе разметку данных, необходимых для обучения модели, составление ключевых слов для отбора релевантных теме исследования публикаций, аналитическую обработку полученных результатов. По итогам сравнительного анализа результатов по двум типам источников данных приводятся ключевые выводы, а также обозначаются перспективы дальнейших исследований тематики.

[1] 知乎 [Zhihu] [сайт]. URL: https://www.zhihu.com/explore (дата обращения: 30.09.2024).

1. Методология исследования

В рамках проведения настоящего исследования была обучена модель сентимент-анализа текстовых данных на китайском языке, делающая доступным автоматическое определение их тональности. На примерах размеченных данных модель, генерирующая паттерн поведения разметчика, научилась «предсказывать» семантическую оценку, представляющую собой число в диапазоне от -1 до 1 (включая его граничные значения), где значениям, близким к -1, соответствуют утверждения с негативной тональностью, а значениям, близким к 1, соответствуют утверждения с позитивной тональностью.

Первым этапом обучения модели стало формирование набора обучающих данных, состоящих из «тренировочной» и «валидационной» частей. «Тренировочная» часть включала в себя перечень размеченных данных формата «текст-оценка», а «валидационная» часть представляла собой подготовленную авторами базу текстовых данных на китайском языке формата «текст-оценка», служащую основой для финальной проверки корректности работы модели.

На этапе формирования «тренировочной» части данных для обучения модели авторами было размечено 3 тыс. коротких текстов на китайском языке. Для повышения точности экспертной оценки и в силу особенностей восприятия тональности человеком текстам присваивались три возможных значения сентимента: «-1» = «негативная оценка», «0» = «нейтральная оценка», «+1» = «позитивная оценка». Разметка «тренировочной» части данных осуществлялась несколькими из соавторов настоящего исследования (каждый из соавторов получил отдельный набор данных для разметки); вследствие однородности и содержательной однозначности текстовых данных и отсутствия в них такого формата преподнесения информации, как ирония или шутка, к оценке качества итоговой разметки внешние исследователи не привлекались. Авторами также не использовался метод автоматической разметки данных, результаты применения которого потенциально не могли бы повлиять на вид итоговой разметки по причине ее однозначности.

Тексты, размечаемые для формирования «тренировочной» выборки, выгружались из базы источников системы iFORA по заданным авторами ключевым словам. Корректность работы обученной модели-регрессора была проверена на «валидационной» выборке, представляющей собой 100 собранных вручную из китаеязычных источников разных типов абзацев, также размеченных на предмет значений сентимента. Затем экспертные оценки сравнивались со значениями сентимента, присвоенными предложениям из «валидационной» выборки моделью, из сопоставления полученных оценок был сделан вывод об эффективности работы модели.

Данный вывод подкрепляется показателями «средняя абсолютная ошибка» (англ. Mean Absolute Error, MAE) и «средняя квадратичная ошибка» (англ. Mean Squared Error, MSE), представляющими собой ключевые метрики эффективности работы модели-регрессора. MAE для обученной модели составляет 0,335, MSE составляет 0,212. Оба значения близки к 0, поэтому можно сделать вывод о том, что модель дает сравнительно небольшое количество ошибок прогноза.

Фрагмент «валидационной» части данных (включающей экспертную оценку тональности предложений (колонка «Оценка авторов») и значения сентимента, присвоенные предложениям моделью) представлен в табл. 1.

Для автоматического выявления релевантных предложений были заданы ключевые слова по тематике развития городской среды в КНР. Перечень исследуемых тематических областей, а также подобранные к ним ключевые слова (с переводом на русский язык) представлены в табл. 2. В качестве тематических областей для исследования были выбраны явления, критически важные для устойчивого развития города или региона и отражающие уровень их экономического развития, являющиеся фундаментальными элементами формирования благополучной для жизни среды: «Здравоохранение», «Транспорт» и «Инфраструктура». Исследование подобных тематических областей помогает получить представление

Таблица 1

«ВАЛИДАЦИОННАЯ» ЧАСТЬ ДАННЫХ ДЛЯ ОБУЧЕНИЯ МОДЕЛИ СЕНТИМЕНТ-АНАЛИЗА (ФРАГМЕНТ)

№	Оригинальный текст	Перевод (справочно)	Оценка авторов	Оценка модели
1	即使像上海这种大城市，在市区中心道路，也是非常的混乱。我开车的时候有一个感觉：咱们中国的交通标志和各种设施，基本上都是不开车的人在办公室里面瞎想出来的。	Даже в таких крупных городах, как Шанхай, дороги в центре города очень запутанные. Когда я за рулем, у меня складывается ощущение, что дорожные знаки и инфраструктура в Китае как будто придуманы людьми, которые сами не водят машину, а только сидят в офисах	-1	-0,978818
2	乡镇污水处理正在发展，因为获利很难，需要数量规模上去才有收益，所以政策有导向，但要发展到村是没有必要的，只需铺设污废水管路，通向集中式处理设施就可以了，应该有部分工程内容还是农民自己集资。	Система очистки сточных вод в населенных пунктах развивается, но так как получение прибыли затруднено, локально получить выгоду возможно только в случае увеличения масштаба очистных работ. Конечно, правительство проводит соответствующие меры, однако вовсе не обязательно развивать систему очистки сточных вод в мелких населенных пунктах, достаточно проложить от них сточные трубопроводы к центральным очистным сооружениям. Таким образом, достаточно решать проблему централизованно, а какие-то мелкие технические проблемы могут решаться жителями мелких населенных пунктов уже самостоятельно	0	-0,036462

Окончание табл. 1

№	Оригинальный текст	Перевод (справочно)	Оценка авторов	Оценка модели
3	在经济发达的省份及直辖市管辖范围之内的基础医疗设施还是不错的，不见得比国外差，为什么看病难，可以学习港澳台嘛，就能解决大部分问题了，其中香港做的最好，用最低的投入获得最大的收益，香港人平均寿命好像是全球第一，医疗投入也不多，民众抱怨也不多，就是做的了公立医院，私立医院加私人诊所的多种医疗组合，做到了有效分流，现在去港澳台便宜方便去看看就知道了，不用在网上瞎抱怨，旅游时适当体验一下当地民众生活，可以开阔眼界，毕竟读万卷书不如行万里路！	В экономически развитых провинциях и в городах центрального подчинения непосредственно под юрисдикцией основных медицинских учреждений медицинское обслуживание обладает высоким качеством. Оно отнюдь не хуже, чем за границей. Вы зря говорите, что сейчас трудно попасть к хорошему врачу: для медицинской консультации вы можете рассмотреть Гонконг, Макао и Тайвань – местные врачи точно смогут решить большинство ваших проблем. Из них самое хорошее медицинское обслуживание – вГонконге. Поход к врачу в Гонконге не обойдется вам слишком дорого, зато польза, которую вы получите от медицинской консультации, будет максимальна: средняя продолжительность жизни гонконгцев, кажется, самая высокая в мире, на медицинское обслуживание люди не жалуются – ни в государственных поликлиниках, ни в частных. Различные частные медицинские организации очень эффективно распределяют работу между собой. Сейчас поехать в Гонконг, Макао и Тайвань за медицинской консультацией действительно недорого и очень полезно. Не нужно жаловаться в интернете! Нужно выбраться и посмотреть на все своими глазами!	1	0,704115

о сильных и слабых сторонах местного управления, ключевых проблемах в регионах, вызывающих наибольший общественный резонанс, и о возможностях для роста уровня жизни.

Таблица 2

КЛЮЧЕВЫЕ СЛОВА ДЛЯ ПОИСКА РЕЛЕВАНТНЫХ ТЕКСТОВЫХ ДОКУМЕНТОВ НА КИТАЙСКОМ ЯЗЫКЕ ПО ТЕМАТИКАМ ИССЛЕДОВАНИЯ

Тематическая область	Ключевые слова для запроса	Перевод ключевых слов для запроса на русский язык (справочно)
Здравоохранение	«医院» OR ((«公立» OR «儿童» OR «市立» OR «综合» OR «地段» OR «多科性») AND «医院») OR «医疗» OR «医疗保险» OR «门诊部» OR ((«儿科» OR «综合性» OR «市立» OR «分科») AND «门诊部») OR «分科诊疗所» OR «诊疗所» OR «医疗改革» OR «医疗质量» OR «医疗效果» OR «医疗服务» OR «医疗技术»	«Больница» OR ((«Государственная» OR «Детская» OR «Муниципальная» OR «Универсальная» OR «Местная» OR «Многопрофильная») AND «Больница») OR «Медицина» OR «Медицинская страховка» OR «Амбулаторное отделение» OR ((«Педиатрическое» OR «Общее» OR «Муниципальное» OR «Специализированное») AND «Отделение») OR «Педиатрия» OR «Клиника» OR «Реформирование здравоохранения» OR «Качество медицинской помощи» OR «Результаты медицинской помощи» OR «Медицинские услуги» OR «Медицинские технологии»

Окончание табл. 2

Тематическая область	Ключевые слова для запроса	Перевод ключевых слов для запроса на русский язык (справочно)
Транспорт	(«交通» OR ((«城市» OR «公共») AND «交通») OR («交通» AND («投资» OR «改革» OR «运输» OR «规划»)) OR «高速公路» OR «公路» OR «公共汽车» OR «公交车» OR «地铁»)	(«Транспорт» OR ((«Городской» OR «Общественный») AND «Транспорт») OR («Транспорт» AND («Инвестиции» OR «Реформа» OR «Перевозка» OR «Планирование»)) OR «Автомагистраль» OR «Дорога» OR «Общественный транспорт» OR «Автобус» OR «Метро»)
Инфраструктура	«基础设施» OR («基础设施» AND («投资» OR «建设»)) OR»城市规划» OR «基础建设» OR «基础结构» OR «公共工程» OR «公共建设» OR «设施»	«Инфраструктура» OR («Инфраструктура» AND («Инвестиции» OR «Строительство»)) OR «Городское планирование» OR «Базовая инфраструктура» OR «Общественная инфраструктура» OR «Социальные инфраструктурные проекты» OR «Строительство общественных объектов» OR «Благоустройство»

Тематическая область «Здравоохранение» включает в себя ряд факторов, связанных с предоставлением медицинских услуг и общим состоянием здоровья городского населения: доступность медицинских учреждений, таких как больницы или медицинские центры, а также качество обслуживания, предоставляемого этими учреждениями. Тематическая область «Транспорт» представлена такими аспектами, как доступность общественного транспорта и его экологичность, качество автодорог, управление дорожным движением и его безопасность. Тематическая область «Инфраструктура»

включает в себя широкий спектр компонентов, таких как транспортные системы, системы связи (телекоммуникации, интернет и т.д.), энергетические системы, системы водоснабжения и другие.

2. Результаты

Результаты исследования представлены ниже в виде диаграмм с распределением показателей сентимента по месяцам по трем исследуемым тематическим направлениям. Для проведения анализа по каждому из тематических направлений были выгружены утверждения, впоследствии оцениваемые моделью на предмет эмоциональной окраски. Так, по корпусу социальных сетей по направлению «Здравоохранение» было выгружено 11 635 утверждений, по «Транспорту» – 14 927 утверждений, по «Инфраструктуре» – 6665 утверждений. В то же время по корпусу СМИ по тематическому направлению «Здравоохранение» было выгружено 25 963 утверждения, по направлению «Транспорт» – 8094 утверждения, по «Инфраструктуре» – 18 196 утверждений. Итого суммарно по всем корпусам и по всем тематическим разрезам анализ проводился на основе 85 480 утверждений, собранных в китайских источниках. Значения выше 0 имеет позитивный сентимент, ниже 0 – негативный сентимент, также на графиках показана линия тренда.

В тематической области «Здравоохранение» наблюдается значительное расхождение в показателях сентимента по СМИ и социальным сетям: так, для СМИ характерно практически абсолютное преобладание позитивного сентимента за исследуемый период, где пиковое положительное значение сентимент принимает в августе 2021 г. (что может быть связано с успешным завершением масштабной пилотной программы внедрения инновационных medtech-решений в деятельность некоторых медицинских учреждений провинции Гуандун и района Большого залива [27]) (рис. 1).

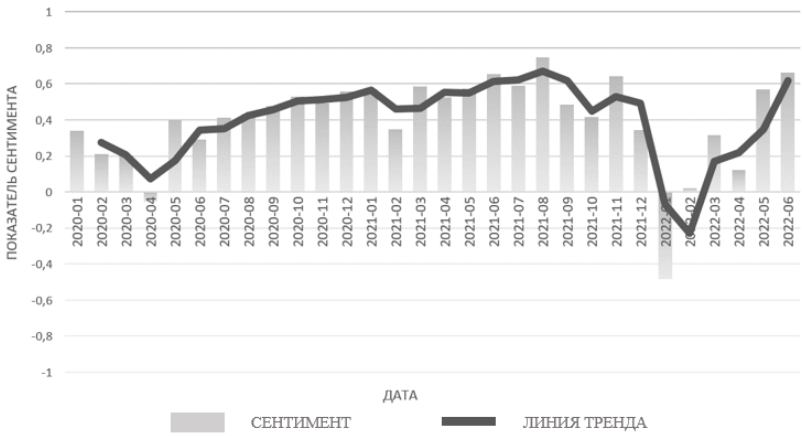

Рис. 1. Распределение сентимента по тематическому направлению «Здравоохранение» по типу источника «СМИ»

По результатам анализа социальных сетей (рис. 2) можно наблюдать противоположную тенденцию: волнообразные колебания сентимента с явным преобладанием негатива, колеблющегося в диапазоне от -0,01 до -0,7, где пиковое значение негатива приходится на октябрь 2021 г. (время обсуждения пользователями социальных сетей резонансного кейса, когда в одном из городов провинции Цзилинь напротив больницы произошло серьезное ДТП с участием нескольких автотранспортных средств, но сотрудники больницы отказали в помощи пострадавшим [28]).

Данные по тематическому направлению «Транспорт» продемонстрировали, что показатели сентимент-анализа по двум типам источников также существенно различаются. Распределение сентимента в СМИ по тематике держится на уровне средней положительной отметки в 0,7 (рис. 3) с некоторыми колебаниями: пиковым положительным значением в марте 2021 г. (что может быть связано с существенным увеличением инвестиций в транспортную инфраструктуру провинции Хэйлунцзян, понесшей особый урон на фоне пандемии коронавируса [29]) и снижением

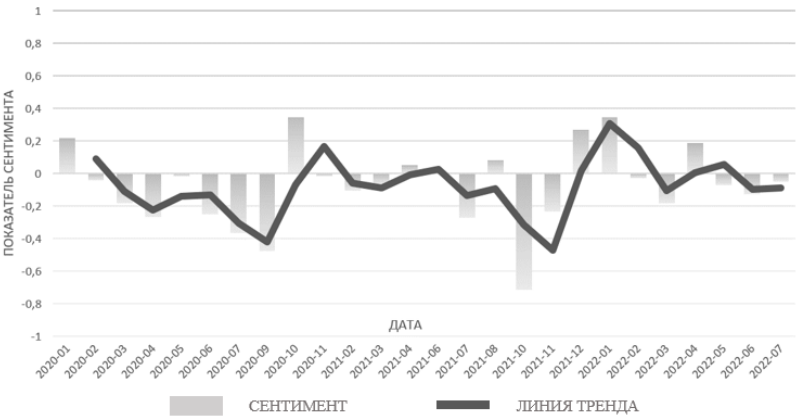

Рис. 2. Распределение сентимента по тематическому направлению «Здравоохранение» по типу источника «Социальные сети»

пика в сентябре 2021 г. (что может быть связано с увольнением с государственных должностей и исключением из КПК представителей высшего менеджмента Транспортной инвестиционной компании Внутренней Монголии [30]).

Рис. 3. Распределение сентимента по направлению «Транспорт» по типу источника «СМИ»

Для социальных сетей показатели сентимента за анализируемый период только в течение нескольких месяцев превышают отметку «0», средний уровень сентимента остается на уровне -0,2, а пиковое значение негативного сентимента, как и в случае со СМИ, приходится на сентябрь 2021 г. (рис. 4).

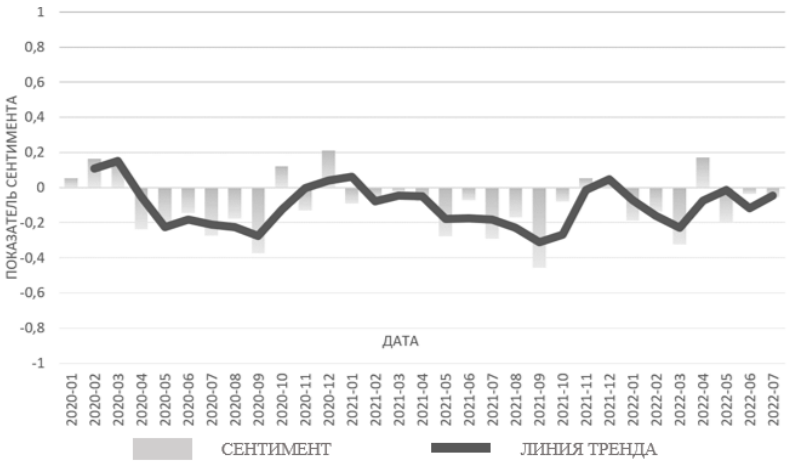

Рис. 4. Распределение сентимента по направлению «Транспорт» по типу источника «Социальные сети»

Для тематической области «Инфраструктура» закономерность распределения сентимента по источникам данных аналогична рассмотренным выше тематикам: данные по СМИ показывают высокий уровень позитивного сентимента, в среднем превышающего отметку в 0,7 (пиковое значение позитивного сентимента – сентябрь 2021 г., когда Госсовет КНР утвердил план развития инновационной инфраструктуры нового типа в рамках программы 14-го пятилетнего плана [31]) (рис. 5).

В распределении данных по социальным сетям наблюдаются волнообразные колебания в диапазоне от -0,4 до 0,3, пик негативного сентимента приходится на март 2022 г., что может быть

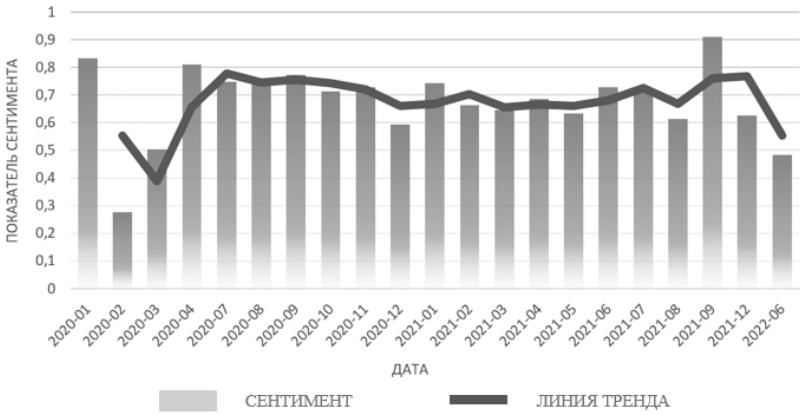

Рис. 5. Распределение сентимента по направлению «Инфраструктура» по типу источника «СМИ»

связано с всплеском недовольства в отношении инфраструктуры провинции Шаньдун (в данный период особенно ярко обсуждается в социальных сетях неэффективность инфраструктурного планирования в Цзинане и нерезультативность стратегии предотвращения наводнений в провинции [32]) (рис. 6).

3. Дискуссия и выводы

В рамках проведенного исследования была обучена модель сентимент-анализа текстов на китайском языке, с помощью которой были проанализированы тексты китайских СМИ и социальных сетей. На примере анализа таких социально значимых тематических областей, как «Здравоохранение», «Транспорт» и «Инфраструктура», было установлено, что тональность текста существенно зависит от типа источников данных. Результаты ряда других исследований демонстрируют похожие выводы: так, сентимент статей, посвященных COVID-19, из официальных новостных источников Китая преимущественно положительный, тогда как сентимент публикаций

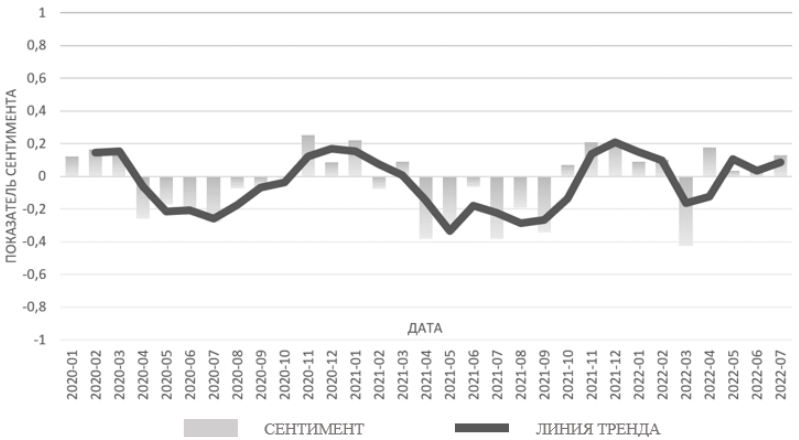

Рис. 6. Распределение сентимента по направлению «Инфраструктура» по типу источника «Социальные сети»

пользователей в Weibo – отрицательный [33]. Отмечается, что в китайских СМИ вопросы здравоохранения в основном освещаются не в негативном свете [34].

Результаты исследования показали, что в среднем в социальных сетях по всем тематикам негативный сентимент преобладает над позитивным, обратный процесс наблюдается в СМИ. В первую очередь было проведено тестирование данных на нормальность с помощью теста Шапиро. В качестве нулевой гипотезы выдвигается утверждение, что данные распределены нормально, как альтернативная гипотеза – данные не распределены нормально. Результатом вычисления стало *p*-значение, равное 0,09132879227399826, что позволяет нам подтвердить нулевую гипотезу. Распределение средних значений в разрезе СМИ и социальных сетей представлено на рис. 7.

В условиях нормального распределения для выявления значимости различия уровня сентимента между двумя типами источников данных был проведен *t*-тест. В качестве нулевой гипотезы выступило суждение о том, что представленные результаты не явля-

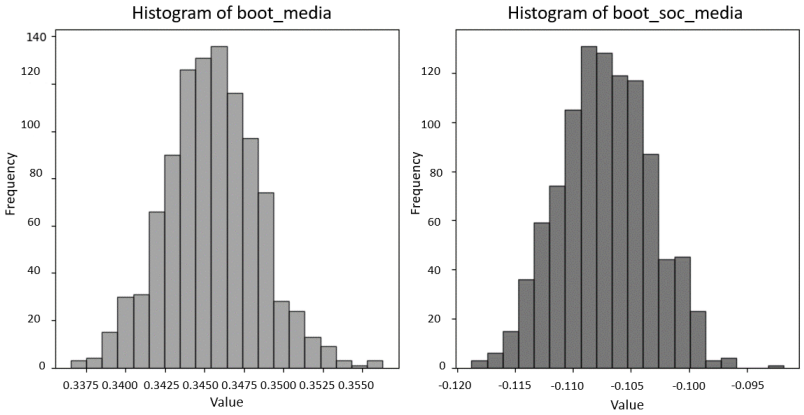

Рис. 7. Распределение средних значений в разрезе СМИ и социальных сетей

ются статистически значимыми (отсутствует существенное различие в значении сентимента для двух типов источников данных). По результатам исследования было получено низкое *p*-значение (49494109979×10 - 1800), что свидетельствует о том, что есть достаточно оснований для того, чтобы отвергнуть нулевую гипотезу. Таким образом, можно установить, что между двумя источниками данных существует значимое различие в уровне сентимента.

Подобные различия между показателями сентимента в СМИ и социальных сетях могут быть вызваны целым рядом факторов. Во-первых, в создании публикуемого в социальных сетях контента, в отличие от СМИ, принимает участие большее число авторов – таким образом, социальные сети как тип источника данных отличаются многообразием представленных точек зрения. Во-вторых, вследствие мобильности в пользовании социальные сети более гибко реагируют на актуальные события.

Существенные различия в уровне сентимента в зависимости от источника данных могут быть вызваны и тем, что социальные сети принимают на себя роль «предохранительных клапанов»

(safety valves) – механизмов снятия общей социальной напряженности, позволяющих обществу время от времени «выпускать пар» и таким образом препятствующих активизации нежелательных для государства внутренних точек напряженности [35]. Данная метафора нередко используется в научной литературе для объяснения механизма действия государственных институтов, в том числе и китайских. Так, в работе Д. Чэнь анализируются особенности теории «предохранительных клапанов», которые, как считается, используют центральные государственные образования КНР [36]. Учитывая важность открытых интернет-сообществ для снятия социальной напряженности, сообщения, связанные с актуальными для значительной части общества проблемами, позволяют пользователям соцсетей в Китае выражать свое недовольство в отношении волнующих их вопросов постепенно и на регулярной основе, без накопления негативных эмоций.

При проведении исследования авторы столкнулись с рядом ограничений, связанных со сложностью получения некоторых потенциально ценных данных из-за несоответствия стандартов создания китаеязычных сайтов стандартам, использующимся в России. Например, в России для получения данных с сайтов путем парсинга (автоматической выгрузки) необходимо наличие у сайтов xml-файлов sitemap, содержащих перечень страниц сайта, а также текстовых файлов robots.txt, в которых хранится информация о доступе к страницам сайта, которые у большинства ресурсов на китайском языке отсутствуют. Основными путями решения проблемы стало использование специальных китаеязычных инструментов парсинга, применение библиотеки Selenium, а также упор в исследовании на доступные для обработки сайты.

Кроме того, еще одним вызовом, с которым столкнулись авторы, стал ряд языковых особенностей организации текстов на китайском языке. К таким особенностям относятся прежде всего синтаксис китаеязычных текстов (отсутствие разделения слов на пробелы, что существенно важно при парсинге, отсутствие форм слов, син-

таксических связей типа «согласование» и «управление» и др.), иероглифическое письмо (иероглифическое написание имен собственных, небольшое количество заимствований), грамматические особенности китайского языка (отсутствие склонений и спряжений, небольшое количество грамматических маркеров времени). Данные проблемы были решены путем кастомизации под обработку китайского языка стандартных алгоритмов токенизации, частеречной разметки и распознавания именованных сущностей, обычно применяемых для языков на основе латинского и кириллического алфавитов. Для сравнения, доля наборов данных для обучения моделей на китайском языке на крупнейшей платформе по машинному обучению Hugging Face[1] составляет всего 3,2%, тогда как на английском – 37,7% [37].

Также стоит отметить, что среди авторов данной статьи нет носителей китайского языка, что являлось ограничением при поиске и верификации информации на китайском языке. Разметка тренировочной части датасета для обучения модели проводилась специалистом – одним из авторов статьи. То, что в процессе кодировки данных был задействован один человек, а не несколько, послужило ограничением при написании статьи, так как в случае работы с эмоционально окрашенными текстами могла иметь место субъективная оценка предлагаемых к разметке утверждений.

Еще одним немаловажным ограничением в рамках проведения настоящего исследования стал закрытый характер кода модели сентимент-анализа текстов на китайском языке, работа с которой стала основой для написания статьи. Код не публикуется в открытых репозиториях в связи с тем, что он защищен правом на интеллектуальную собственность (номер регистрации свидетельства РИД 2023680870). Для преодоления этого ограничения и обеспечения возможности проверки воспроизводимости результата другими исследователями был опубликован полный

[1] Hugging Face [сайт]. URL: https://huggingface.co/ (дата обращения: 30.09.2024).

датасет, содержащий 85 480 утверждений, собранных в китайских источниках, на которых проводился анализ[1]. Для обеспечения воспроизводимости полученных результатов авторы по запросу могут предоставить дополнительные технические разъяснения по алгоритму расчетов уровня сентимента на основе специально разработанной модели.

Исходя из ряда ограничений настоящего исследования, направление дальнейших работ в области изучения общественного мнения в Китае методами автоматизированного анализа текстовых данных может быть задано несколькими векторами:

– продолжение анализа в динамике с целью изучения трансформации рассматриваемых явлений, возможной на протяжении более длительного периода времени, в том числе анализа причин этих изменений;

– углубление теоретической основы исследования: например, более подробное изучение социальных сетей как «предохранительных клапанов», в том числе выявление механизмов их функционирования с учетом особенностей политической системы Китая;

– расширение охвата изучаемых источников, в том числе социальных сетей: при нахождении юридически не запрещенных путей сбора данных с китаеязычных сайтов возможно добавление пользовательских публикаций и комментариев с популярного ресурса Weibo (微博), а также других китаеязычных ресурсов. Охват более широкого спектра источников потенциально полезен наличием возможности углубить исследование за счет сегментирования типов источников на отдельные группы и подгруппы для проведения анализа в их разрезе.

Исследование обладает рядом преимуществ методологического и общенаучного характера, главным из которых является применение сентимент-анализа, базирующегося на автоматиче-

[1] Доступ к датасету осуществляется по ссылке на сайт GitHub: URL: https://github.com/issekifora/dataset_chinese_phrases (дата обращения: 30.09.2024).

ской обработке больших текстовых данных. Разработанный метод позволяет, во-первых, сформировать репрезентативную выборку публикаций за счет большого объема исследуемых текстов, а во-вторых – оценить эмоциональную окраску отдельных мнений (публикаций в социальных сетях), а также новостных сообщений в СМИ в измеримом, количественном выражении. Таким образом, предложенная методология позволяет сделать выводы, подкрепленные объективными результатами. Методология и результаты исследования полезны для повышения объективности, оперативности и эффективности принятия стратегических решений в области планирования инфраструктуры, а также могут служить методологической основой проведения аналогичных исследований других ключевых сфер экономики.

ЛИТЕРАТУРА

1. *Hu Y.S.* The impact of increasing returns on knowledge and big data: from Adam Smith and Allyn Young to the age of machine learning and digital platforms // Prometheus. 2020. Vol. 36, No. 1. P. 10–29. DOI: 10.13169/prometheus.36.1.0010.

2. *Henke N., Libarikian A., Wiseman B.* Straight talk about big data // McKinsey Quarterly: [сайт]. 28.10.2016. URL: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/straight-talk-about-big-data (дата обращения: 16.01.2023).

3. 中华人民共和国国家互联网信息办公室。第47次《中国互联网络发展状况统计报告》（全文）[*Государственная канцелярия интернет-информации КНР*. Сорок седьмой статистический отчет о состоянии развития Интернета в Китае (полный текст)]. 03.02.2021. URL: http://www.cac.gov.cn/2021-02/03/c_1613923423079314.htm (дата обращения: 16.01.2023).

4. Individuals using the Internet (% of population) // World Bank: [сайт]. 2023. URL: https://data.worldbank.org/indicator/IT.NET.USER.ZS (дата обращения: 22.03.2023).

5. *Liu B.* Sentiment analysis: mining opinions, sentiments, and emotions. Cambridge: Cambridge university press, 2015. 381 p. DOI: 10.1017/CBO9781139084789.

6. *Pang B., Lee L.* Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales // Proceedings of ACL. 2005. P. 115–124. DOI: 10.3115/1219840.1219855.

7. *Taboada M.* Sentiment Analysis: An Overview from Linguistics // Annual Review of Linguistics. 2016. Vol 2. P. 325–347. DOI: 10.1146/annurev-linguistics-011415-040518. EDN: YAKIFR.

8. *Ohman E.* The validity of lexicon-based emotion analysis in interdisciplinary research // Proceedings of the Workshop on Natural Language Processing for Digital Humanities. December 16–19, 2021 / NLP Association of India. Silchar, India, 2021. P. 7–12.

9. *Колмогорова А.В., Калинин АА., Маликова А.В.* Лингвистические принципы и методы компьютерной лингвистики для решения задач сентимент-анализа русскоязычных текстов // Актуальные проблемы филологии и педагогической лингвистики. 2018. № 1 (29). С. 139–148. DOI: 10.29025/2079-6021-2018-1(29)-139-148. EDN: YRHARM.

10. Construct validity of six sentiment analysis methods in the text of encounter notes of patients with critical illness / G.E. Weissman, L.H. Ungar, M.O. Harhay [et al.] // Journal of biomedical informatics. 2019. No. 89. P. 114–121. DOI: 10.1016/j.jbi.2018.12.001.

11. Medical sentiment analysis using social media: towards building a patient assisted system / S. Yadav, A. Ekbal, S. Saha, P. Bhattacharyya // Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, May 07–12, 2018. P. 2790–2797. EDN: YZXBDW.

12. *Luis M.D., Juan C.M., Glen M.* Social media as a resource for sentiment analysis of Airport Service Quality (ASQ) // Journal of Air Transport Management. 2019. No. 78. P. 106–115. DOI: 10.1016/j.jairtraman.2019.01.004.

13. *Islam M.R., Zibran M.F.* Sentiment analysis of software bug related commit messages // Network. 2018. Vol. 740. P. 740.

14. Twitter sentiment analysis applied to finance: A case study in the retail industry / T.T.P. Souza, O. Kolchyna, P.C. Treleaven, T. Aste // ArXiv. Submitted on 2 Jul 2015 (v. 1), last revised 11 Jul 2015. URL: arXiv preprint arXiv:1507.00784 (дата обращения: 30.09.2024).

15. Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods / F.N. Ribeiro, M. Araújo, P. Gonçalves [et al.] // EPJ Data Science. 2016. Vol. 5, No. 1. P. 1–29. DOI: 10.1140/epjds/s13688-016-0085-1. EDN: RMUGIO.

16. *Van Atteveldt W., Van der Velden M.A., Boukes M.* The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms // Communication Methods and Measures. 2021. No. 15 (2). P. 121–140. DOI: 10.1080/19312458.2020.1869198.

17. *Kasper K.N.* Assessing the Validity of Sentiment Analysis Measures through Polychoric Correlation // University of New Mexico. Digital Repository. 2020. URL: https://digitalrepository.unm.edu/math_etds/174 (дата обращения: 30.09.2024).

18. *Pandian A.P.* Performance evaluation and comparison using deep learning techniques in sentiment analysis // Journal of Soft Computing Paradigm. 2021. No. 3 (2). P. 123–134. DOI: 10.36548/jscp.2021.2.006.

19. Happy parents' tweets: An exploration of Italian Twitter data using sentiment analysis / L. Mencarini, D.I.H. Farias, M. Lai [et al.] // Demographic Research. 2019. Vol. 40. P. 693–724. DOI: 10.4054/DemRes.2019.40.25.

20. *Stieglitz S., Dang-Xuan L.* Emotions and information diffusion in social media–sentiment of microblogs and sharing behavior // Journal of management information systems. 2013. Vol. 29, No. 4. P. 217–248. DOI: 10.2753/MIS0742-1222290408.

21. COVID-19 vaccine sentiment analysis using public opinions on Twitter / P. Chinnasamy, V. Suresh, K. Ramprathap [et al.] // Materials Today: Proceedings. 2022. Vol. 64. P. 448–451. DOI: 10.1016/j.matpr.2022.04.809.

22. 陈凌，宋衍欣。基于公众情绪上下文的LSTM情感分析研究——以台风“利奇马”为例//现代情报 [*Чэнь Лин, Сун Яньсинь.* Сентимент-анализ публичных настроений с помощью LSTM на примере тайфуна «Лекима» // Сяньдай Цинбао]. 2020. Т. 40, №6. С. 98–105. DOI: 10.3969/j.issn.1008-0821.2020.06.010.

23. 杨洸。社交媒体网络情感传染及线索影响机制的实证分析//深圳大学学报（人文社科版）[*Ян Гуан.* Эмпирический анализ эмоционального заражения и механизмов воздействия подсказок в социальных сетях // Вестник Шэньчжэньского университета (гуманитарные и социальные науки)]. 2020. Т. 37, № 6. С. 115–126.

24. 岳宗朴，刘彩，李莹，陆文静。基于微博数据挖掘的“新冠疫情”评论文本分析/天津中医药大学管理学院 [*Юэ Цзунпу, Лю Цай, Ли Ин, Лу Вэньцзин*. Анализ текстовых комментариев по тематике «Новая коронавирусная инфекция» на основе анализа данных Weibo / Факультет менеджмента Тяньцзиньского университета традиционной китайской медицины]. 2020 (12). С. 48–50.

25. 姚天昉。娄德成。汉语语句主题语义倾向分析方法的研究//中文信息学报 [*Яо Тяньфан, Лоу Дэчэн*. Исследование метода анализа тематико-семантической структуры текстов на китайском языке // Китайский журнал о науках об информации]. 2007. № 5. С. 73–79. ISBN: 1003–0077 (2007) 05–0000–00.

26. 知乎第一季度营收同比增长55.4%，月活用户1.016亿//IT之家 [Доход компании Zhihu в первом квартале вырос на 55,4% по сравнению с аналогичным периодом прошлого года, при 101,6 млн ежемесячных активных пользователей // IT Чжи Цзя]. 2022. URL: https://baijiahao.baidu.com/s?id=1733794653563608924 (дата обращения: 01.03.2023).

27. 中央人民政府。粤港澳大湾区：完善联动机制加快跨境医疗合作 [Госсовет КНР. Гуандун, Гонконг, Макао и зона Большого залива: совершенствование механизма связи для ускорения трансграничного медицинского сотрудничества]. 2021. URL: http://www.zlb.gov.cn/2021-08/23/c_1211341836.htm (дата обращения: 09.03.2023).

28. 吉林一医院门口车祸无人救治？院方：医生不能脱岗，护士保安一人一岗//北晚在线 [Автокатастрофа у входа в больницу в Цзилине и никто не пришел на помощь? Комментарий со стороны больницы: Врачи не могут покидать свои рабочие места, медсестры и охранники также находятся на своем посту // Бэйвань цзай сянь]. 2020. URL: https://baijiahao.baidu.com/s?id=1680050105222894394&wfr=spider&for=pc (дата обращения: 09.03.2023).

29. 2021年黑龙江省计划完成交通运输投资600亿元//人民网 [Провинция Хэйлунцзян планирует реализовать 60 млрд юаней в виде инвестиций в развитие транспорта в 2021 году // Жэньминван]. 2021. URL: http://hlj.people.com.cn/n2/2021/0319/c220024-34631394.html (дата обращения: 09.03.2023).

30. 内蒙古4名“厅官”被开除党籍或公职//新华网 [Четверо «официальных лиц» во Внутренней Монголии были исключены из партии и лишены права занимать государственные должности // Синьхуа]. 2021. URL: http://www.xinhuanet.com/2021-09/10/c_1127850249.htm (дата обращения: 09.03.2023).

31. 我国支持民营和境外资本参与新型基础设施投资运营//新华网 [Китай поддерживает привлечение частного и иностранного капитала для инвестирования в новую инфраструктуру // Синьхуа]. 2021. URL: http://www.xinhuanet.com/2021-09/10/c_1127850249.htm (дата обращения: 09.03.2023).

32. 济南718事件情况是怎样的？//知乎 [Каковы обстоятельства инцидента 718 в Цзинане? // Чжиху]. 2022. URL: https://www.zhihu.com/question/282692759 (дата обращения: 09.03.2023).

33. *Duan Y., Liu L., Wang Z.* COVID-19 sentiment and the Chinese stock market: evidence from the official news media and Sina Weibo // Research in International Business and Finance. 2021. Vol. 58. DOI: 10.1016/j.ribaf.2021.101432.

34. *Peng W., Tang L.* Health content in Chinese newspapers // Journal of health communication. 2010. Vol. 15, No. 7. P. 695–711. DOI: 10.1080/10810730.2010.514028.

35. *Hassid J.* Safety valve or pressure cooker? Blogs in Chinese political life // Journal of Communication. 2012. Vol. 62, No. 2. P. 212–230. DOI: 10.1111/j.1460-2466.2012.01634.x.

36. *Chen D.* Review essay: The safety valve analogy in Chinese politics // Journal of East Asian Studies. 2016. Vol. 16, No. 2. P. 281–294. DOI:10.1017/jea.2016.4.

37. AI Language Models: Technological, Socio-Economic and Policy Considerations // OECD. 2023. Vol. 352. P. 1.

Сведения об авторах

Мария Сергеевна Анташева

Эксперт отдела информационно-аналитических систем Центра стратегической аналитики и больших данных Института статистических исследований и экономики знаний НИУ ВШЭ

ResearcherID: HTN-3351-2023

Полина Александровна Лобанова
Заведующая отделом информационно-аналитических систем Центра стратегической аналитики и больших данных Института статистических исследований и экономики знаний НИУ ВШЭ
ResearcherID: W-4562-2017

Юлия Камаловна Исаева
Ведущий программист отдела разработки интеллектуальных систем Центра стратегической аналитики и больших данных Института статистических исследований и экономики знаний НИУ ВШЭ
SPIN-код: 6151-8711
ResearcherID: O-4549-2018

Елизавета Алексеевна Сабидаева
Ведущий эксперт отдела информационно-аналитических систем Центра стратегической аналитики и больших данных Института статистических исследований и экономики знаний НИУ ВШЭ

Анна Сергеевна Пиекалнитс
Ведущий эксперт отдела исследований больших данных Центра стратегической аналитики и больших данных Института статистических исследований и экономики знаний НИУ ВШЭ

Ирина Владимировна Логинова
Заведующая отделом исследований больших данных Центра стратегической аналитики и больших данных Института статистических исследований и экономики знаний НИУ ВШЭ
SPIN-код: 2221-7707
ResearcherID: J-6034-2015

DOI: 10.19181/4m.2023.32.2.1

SENTIMENT ANALYSIS AS AN INFORMATION AGENDA AND PUBLIC OPINION RESEARCH METHOD (ON THE EXAMPLE OF CHINESE MASS MEDIA AND SOCIAL NETWORKS)

Antasheva Mariia S.
HSE University, Moscow, Russia,
msantasheva@hse.ru
ORCID: 0000-0002-5255-8773

Lobanova Polina A.
HSE University, Moscow, Russia,
plobanova@hse.ru
ORCID: 0000-0002-9878-9390

Isaeva Iuliia K.,
HSE University, Moscow, Russia,
yisaeva@hse.ru
ORCID: 0000-0002-7974-8294

Sabidaeva Elizaveta A.,
HSE University, Moscow, Russia,
esabidaeva@hse.ru
ORCID: 0000-0001-9115-2285

Piekalnits Anna S.,
HSE University, Moscow, Russia,
apiekalnits@hse.ru
ORCID: 0000-0003-0585-5350

Loginova Irina V.,
HSE University, Moscow, Russia,
iloginova@hse.ru
ORCID: 0000-0002-3376-2728

For citation: Antasheva M., Lobanova P., Isaeva I., Sabidaeva E., Piekalnits A., Loginova I. Sentiment analysis as an information agenda and public opinion research method (on the example of Chinese mass media and social networks). S*otsiologiya 4M (Sociology: methodology, methods, mathematical modeling)*, 2023, no. 57, p. 7–41. DOI: 10.19181/4m.2023.32.2.1

Abstract. The information agenda broadcast by Chinese media resources is a source of up-to-date data on public opinion on key issues of social welfare. Due to the technical peculiarities of the organization of Chinese websites and the need to attract additional resources for automatic processing (parsing) of texts in Chinese, this topic is not widely represented in domestic and foreign studies. The purpose of this paper is to demonstrate the methodology and results of public opinion estimation on the example of data collected from Chinese media and social networks based on a trained sentiment analysis model of Chinese text data. The ML model was used to comparatively analyze Chinese language content on urban infrastructure development issues for the period 2020–2022. The results are presented in the format of sentiment distribution charts based on media and social media data by month over a 2-year period. It is revealed that the level of sentiment differs significantly depending on the type of data source. A steady prevalence of positive sentiment in mass media and negative sentiment in social networks was determined, which can be explained by differences in the composition of text authors, restrictions imposed on the content published in the sources, as well as different purposes of resource use by users.

Keywords: sentiment analysis, emotional colouring of texts, urban infrastructure development, public opinion, Chinese language, machine learning, data mining, social networking websites

Acknowledgments: The paper was prepared in the framework of a research grant funded by the Ministry of Science and Higher Education of the Russian Federation (grant ID: 075-15-2022-325).

References

1. Hu Y.S. The impact of increasing returns on knowledge and big data: from Adam Smith and Allyn Young to the age of machine learning and digital platforms, *Prometheus,* 2020, vol. 36 (1), p. 10–29.
2. Henke N., Libarikian A., Wiseman B. Straight talk about big data. *McKinsey Quarterly*. 2016. URL: https://www.mckinsey.com/capabilities/

mckinsey-digital/our-insights/straight-talk-about-big-data (date of access: 16 January 2023).
3. 中华人民共和国国家互联网信息办公室。第47次《中国互联网络发展状况统计报告》（全文）. [Cyberspace Administration of China. (2021) *Forty-seventh statistical report on the Internet development in China.*] URL: http://www.cac.gov.cn/2021-02/03/c_1613923423079314.htm (date of access: January 16 2023) (in Chinese).
4. World Bank. *Individuals using the Internet (% of population)*. 2023. URL: https://data.worldbank.org/indicator/IT.NET.USER.ZS?end=2021&start=2021&view=bar (date of access: March 22 2023).
5. Liu B. *Sentiment analysis: mining opinions, sentiments, and emotions*. Cambridge: Cambridge university press, 2015. 381 p.
6. Pang B., Lee L. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales, *Proceedings of ACL,* 2005, p. 115–124.
7. Taboada M. Sentiment Analysis: An Overview from Linguistics, *Annual Review of Linguistics*, 2016, vol. 2, p. 325–349.
8. Ohman E. The validity of lexicon-based emotion analysis in interdisciplinary research, *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*. NLP Association of India, 2021. P. 7-12.
9. Kolmogorova A. V., Kalinin A. A., & Malikova A. V. Linguistic principles and computational linguistics methods for the purposes of sentiment analysis of Russian texts (in Russian), *Current Issues in Philology and Pedagogical Linguistics*, 2018, vol. 1 (29), p. 139–148.
10. Weissman G.E., Ungar L.H., Harhay M.O., Courtright K.R., Halpern S.D. Construct validity of six sentiment analysis methods in the text of encounter notes of patients with critical illness, *Journal of biomedical informatics*, 2019, vol. 89, p. 114–121.
11. Yadav S., Ekbal A., Saha S., Bhattacharyya P. Medical sentiment analysis using social media: towards building a patient assisted system, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 2018 (LREC 2018).
12. Luis M.D., Juan C.M., Glen M. Social media as a resource for sentiment analysis of Airport Service Quality (ASQ), *Journal of Air Transport Management*, 2019, vol. 78, p. 106–115.

13. Islam M.R., Zibran M.F. Sentiment analysis of software bug related commit messages, *Network,* 2018, vol. 740, p. 740.
14. Souza T.T.P., Kolchyna O., Treleaven P.C., Aste T. *Twitter sentiment analysis applied to finance: A case study in the retail industry,* 2015, arXiv preprint arXiv:1507.00784.
15. Ribeiro F.N., Araújo M., Gonçalves P., André Gonçalves M., Benevenuto F. Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods, *EPJ Data Science,* 2016, vol. 5, p. 1–29.
16. Van Atteveldt W., Van der Velden M.A., & Boukes M. The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms, *Communication Methods and Measures*, 2021, vol. 15 (2), p. 121–140.
17. Kasper K.N. Assessing the Validity of Sentiment Analysis Measures through Polychoric Correlation, *University of New Mexico Digital Repository*. 2020. URL: https://digitalrepository.unm.edu/math_etds/174 (date of access: March 1 2023)
18. Pandian A.P. Performance evaluation and comparison using deep learning techniques in sentiment analysis, *Journal of Soft Computing Paradigm,* 2021, vol. 3 (2), p. 123–134.
19. Mencarini L. et al. Happy parents' tweets. *Demographic Research*, 2019, vol. 40, p. 693–724.
20. Stieglitz S., Dang-Xuan L. Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior, *Journal of management information systems,* 2013, vol. 29 (4), p. 217–248.
21. Chinnasamy P. et al. COVID-19 vaccine sentiment analysis using public opinions on Twitter. *Materials Today: Proceedings,* 2022, vol. 64, p. 448–451.
22. 陈凌，宋衍欣。基于公众情绪上下文的LSTM情感分析研究——以台风“利奇马”为例//现代情报 [Chen Ling, Song Yanxin. (2020) Public sentiment analysis using LSTM on the example of Typhoon Lekima. *Xiandai qingbao*] Vol. 40 (6), p. 98–105. (in Chinese).
23. 杨洸。社交媒体网络情感传染及线索影响机制的实证分析//深圳大学学报（人文社科版）。[Yang Guang. (2020) Empirical Analysis of Emotional Contagion and Mechanisms of Impact of Prompts in Social Networks. *Shenzhen University Newsletter (Humanities & Social Sciences)*] Vol. 37 (6), p. 115–126. (in Chinese).
24. 岳宗朴，刘彩，李莹，陆文静。基于微博数据挖掘的“新冠疫情”评论文本分析//天津中医药大学管理学院。[Yue Zongpu, Liu Cai,

Li Ying, Lu Wenjing. (2020) Text commentaries analysis on «The new Coronavirus infection» in Weibo. *Tianjin University of Traditional Chinese Medicine, School of Management*.] Vol. 12, p. 48–50. (in Chinese).
25. 姚天昉。娄德成。汉语语句主题语义倾向分析方法的研究//中文信息学报。[Yao Tianfang, Lou Decheng. (2007) Research on the analysis method of topic semantic tendency of Chinese sentences // *Chinese Journal in Information Science*.] No. 5, p. 73–79. (in Chinese).
26. 知乎第一季度营收同比增长55.4%，月活用户1.016亿//IT之家。[Zhihu's first quarter revenue up 55.4% year over year, with 101.6 million monthly active users. (2022). *IT Zhi Jia.*]. URL: https://baijiahao.baidu.com/s?id=1733794653563608924 (date of access: March 1 2023) (in Chinese).
27. 中央人民政府。粤港澳大湾区：完善联动机制加快跨境医疗合作。[Central People's Government. (2021) *Guangdong, Hong Kong, Macao and the Greater Bay Area: Improving the linkage mechanism to speed up cross-border medical cooperation*.] URL: http://www.zlb.gov.cn/2021-08/23/c_1211341836.htm (date of access: March 9 2023) (in Chinese).
28. 吉林一医院门口车祸无人救治？院方：医生不能脱岗，护士保安一人一岗//北晚在线。[A car accident at the entrance of a hospital in Jilin and no one came for help? Hospital: Doctors cannot leave their posts, and nurses and security guards are on one post. (2020). *Beiwan zai xian*.] URL: https://baijiahao.baidu.com/s?id=1680050105222894394&wfr=spider&for=pc (date of access: March 9 2023) (in Chinese).
29. 2021年黑龙江省计划完成交通运输投资600亿元//人民网。[In 2021, Heilongjiang Province plans to complete an investment of 60 billion yuan in transportation. (2021). *Renmin wang*.]. URL: http://hlj.people.com.cn/n2/2021/0319/c220024-34631394.html (date of access: March 9 2023) (in Chinese).
30. 内蒙古4名“厅官”被开除党籍或公职//新华社。[Four «department officials» in Inner Mongolia were expelled from the party and public office. (2021). *Xinhua*.]. URL: http://www.xinhuanet.com/2021-09/10/c_1127850249.htm (date of access: March 9 2023) (in Chinese).
31. 我国支持民营和境外资本参与新型基础设施投资运营//新华社 [China supports the participation of private and foreign capital in the investment and operation of new infrastructure. (2021). *Xinhua*.]. URL: https://www.gov.cn/zhengce/2021-09/22/content_5638771.htm (date of access: March 9 2023) (in Chinese).

32. 济南718事件情况是怎样的？知乎//知乎。[What is the situation of the 718 incident in Jinan? (2022). *Zhihu.*]. URL: https://www.zhihu.com/question/282692759 (date of access: March 9 2023) (in Chinese).
33. Duan Y., Liu L., Wang Z. COVID-19 sentiment and the Chinese stock market: evidence from the official news media and Sina Weibo, *Research in International Business and Finance.* 2021, vol. 58.
34. Peng W., Tang L. Health content in Chinese newspapers, *Journal of health communication,* 2010, vol. 15 (7), p. 695–711.
35. Hassid J. Safety valve or pressure cooker? Blogs in Chinese political life, *Journal of Communication*, 2012, vol. 62 (2), p. 212–230.
36. Chen D. Review essay: The safety valve analogy in Chinese politics, *Journal of East Asian Studies*, 2016, vol. 16 (2), p. 281–294.
37. OECD. AI Language Models: Technological, Socio-Economic and Policy Considerations, *OECD*. 2023, vol. 352, p. 1.

Information about the authors

Mariia S. Antasheva
Expert of Data Analysis Unit, Institute of Statistical Studies and Economics of Knowledge, HSE University
ResearcherID: HTN-3351-2023

Polina A. Lobanova
Head of Data Analysis Unit, Institute of Statistical Studies and Economics of Knowledge, HSE University
ResearcherID: W-4562-2017

Iuliia K. Isaeva
Leading programmer of Data Analysis Unit, Institute of Statistical Studies and Economics of Knowledge, HSE University
SPIN-code: 6151-8711
ResearcherID: O-4549-2018

Elizaveta A. Sabidaeva
Leading expert of Big Data Research Unit, Institute of Statistical Studies and Economics of Knowledge, HSE University

Anna S. Piekalnits
Leading expert of Data Analysis Unit, Institute of Statistical Studies and Economics of Knowledge, HSE University

Irina V. Loginova
Head of Big Data Research Unit, Institute of Statistical Studies and Economics of Knowledge, HSE University
SPIN-code: 2221-7707
ResearcherID: J-6034-2015