УДК 004.8

Инструмент сбора и оценки общественного мнения о транспортной отрасли из открытых источников

Стрельников Никита Рафаилевич — курсант 4-го курса направления 09.05.01 «Применение и эксплуатация автоматизированных систем специального назначения». Научные интересы: обработка естественного языка, методы машинного обучения, анализ данных. E-mail: vka@mil.ru

Михайлова Светлана Александровна — преподаватель кафедры математического и программного обеспечения специальных космических комплексов. Научные интересы: распределенные системы, архитектура приложений, методы машинного обучения, анализ данных. E-mail: vka@mil.ru

Алимов Наиль Ильгизович — кандидат техн. наук, старший преподаватель кафедры математического и программного обеспечения специальных космических комплексов. Научные интересы: методы и алгоритмы обработки данных, параллельные и распределенные вычисления. E-mail: vka@mil.ru

Военно-космическая академия имени А. Ф. Можайского, Россия, 197198, Санкт-Петербург, ул. Ждановская, 13

Для цитирования: Стрельников Н. Р., Михайлова С. А., Алимов Н. И. Инструмент сбора и оценки общественного мнения о транспортной отрасли из открытых источников // Интеллектуальные технологии на транспорте. 2026. № 2 (46). С. 46–53. DOI: 10.20295/2413-2527-2026-246-46-53

Аннотация. *Представлено исследование о применении методов интеллектуального анализа текстовой информации из открытых источников для формирования управленческих решений на примере анализа отзывов населения о транспортной отрасли.* ***Цель:*** *обоснование подхода к интеллектуальному анализу текстовых отзывов пассажиров о транспортной отрасли, направленного на трансформацию неструктурированных пользовательских высказываний в структурированные данные, пригодные для формирования управленческих решений.* ***Методы:*** *использованы методы обработки естественного языка, машинное обучение и технологии интеграции данных из разнородных источников.* ***Результаты:*** *подчеркивается эффективность предложенного конвейера обработки текста, включающего классификацию по виду транспорта и тематике, анализ тональности и извлечение именованных сущностей для выявления ключевых проблем и трендов общественного мнения.* ***Практическая значимость:*** *повышение оперативности реагирования транспортных регуляторов на проблемы пассажиров, оптимизация сервиса и переход к управлению на основе данных. Исследование имеет важное значение для развития цифровых технологий в транспортной отрасли и повышения эффективности управления пассажирскими перевозками в условиях цифровой трансформации.*

Ключевые слова: *интеллектуальный анализ данных, обработка естественного языка, Python, транспорт, пассажиры, тональность, машинное обучение, медиапространство*

2.3.5 — *математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей (технические науки);* **1.2.1** — *искусственный интеллект и машинное обучение (технические науки)*

Введение

Цифровое развитие глобально меняет взаимодействие между акторами сферы услуг [1]. Транспортная отрасль не является исключением и внедряет цифровые технологии во все сферы деятельности [2]. Пользователи транспортной отрасли, в свою очередь, активно реагируют на

нововведения и буквально в режиме реального времени оценивают качество оказываемых им услуг [3]. Наиболее полными и объективными в данной ситуации являются отзывы пользователей, публикуемые в открытых источниках: в мессенджерах, на форумах, в тематических рубриках и чатах [4]. Именно в этих источниках выявляются главные системные проблемы: недовольство логикой распределения пассажиропотока, некорректная логистика транспорта и многие другие проблемы, которые активно обсуждаются пользователями между собой и являются ценным источником сведений.

Однако объем публикуемой информации становится настолько велик, что ее ручная обработка далее невозможна и необходимо применение автоматизированных систем обработки и анализа информации. Современные технологии обработки естественного языка [5] и машинного обучения [6] предлагают инструменты для трансформации неструктурированного текстового потока в структурированное знание, пригодное для аналитики и принятия решений.

Целью статьи является разработка и обоснование подхода к интеллектуальному анализу текстовых отзывов пассажиров о транспортной отрасли. В задачи исследования входит систематизация типов источников данных и методов их сбора, проектирование конвейера обработки текста, включающего этапы очистки, классификации, анализ тональности и извлечение информации, демонстрация работоспособности подхода на практических примерах, формулирование практической значимости для транспортной отрасли.

Источники данных и методы их сбора

Источники данных, содержащие информацию о транспортной ситуации и мнениях пассажиров, можно разделить на два основных типа [7]:

1. Официальные (структурированные): нормативные документы, формирующиеся в рамках транспортного предприятия, данные системы отслеживания движения транспорта на основе ГЛОНАСС или GPS, официальные статистические данные.

2. Неофициальные (слабоструктурированные): средства массовой информации (СМИ), результаты опросов, социальные сети, мессенджеры, открытые форумы и порталы для обсуждения качества предоставляемых услуг.

Извлечение информации из подобных источников может осуществляться различными методами. В отличие от структурированных ресурсов, предоставляющих уже классифицированные данные по определенным категориям, слабоструктурированные источники представляют собой более сложную задачу для обработки.

Для организации сбора информации из слабоструктурированных источников применяются методы интеллектуальной обработки текстов и анализа неструктурированных данных [8]. Основным инструментом в этом случае выступает скрапинг — технология автоматизированного сбора данных из источников, размещенных в открытом доступе в сети Интернет. После скачивания текста необходимо выполнить его семантический анализ с применением методов обработки естественного языка, что позволяет не только извлечь из массива сообщений конкретные сущности, такие как маршруты, временные промежутки и типы транспорта, но и провести классификацию тональности высказываний (сентимент-анализ) для определения общего настроения пассажиров.

Telegram, как наиболее известный и активно используемый мессенджер, является одним из самых популярных способов обсуждения транспорта в сети Интернет. В нем существуют как официальные каналы перевозчиков, так и неофициальные сообщества, созданные пользователями. Эти каналы содержат прямые текстовые посты, комментарии, медиафайлы с подписями. Для сбора данных из Telegram используется асинхронная библиотека Telethon, предоставляющая доступ к интерфейсу прикладного программирования мессенджера.

Первоначальный отбор источников производился на основе принципа верификации [9]. Для отсеивания откровенно маргинальных или анонимных каналов с низкой степенью достоверности в выборку включались только те ресурсы, которые были официально зарегистрированы в качестве СМИ в реестре Роскомнадзора. Это позволило обеспечить минимальный порог ответственности авторов

за публикуемые сведения и гарантировать определенный уровень регулярности выхода материалов.

Для формирования первичного списка каналов использовались открытые веб-каталоги и агрегаторы Telegram, в частности tgstat.ru. С его помощью был собран датасет, включающий более 4000 открытых телеграм-каналов. В выборку включены каналы 43 различных тематических категорий, что позволяет отслеживать, как транспортные проблемы обсуждаются в контексте городской жизни, экономики, происшествий и других сфер.

Информация собиралась за период с 1 января 2022 года по 17 ноября 2025 года. Такой длительный временной срез позволяет не только оценить текущее состояние настроений пассажиров, но и проанализировать динамику их изменений в зависимости от сезона, экономической ситуации или резонансных событий.

В результате проведенных работ по сбору и первичной агрегации данных была сформирована база, в которой на текущий момент хранится 1 026 163 поста. Этот массив неструктурированных текстовых данных является основой для последующего семантического анализа и выявления закономерностей, связанных с удовлетворенностью пассажиров общественным транспортом.

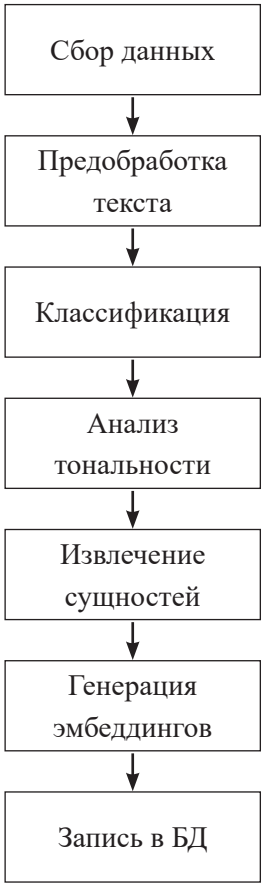

Рис. 1. Схема работы конвейера обработки и анализа текста

Конвейер обработки и анализа текста

Обработка и анализ текста выполняются за семь этапов (рис. 1).

Предобработка и очистка данных выполняется в несколько этапов. Текст приводится к единому нижнему регистру, удаляются символы пунктуации и ссылки. Затем текст токенизируется и приводится к морфологической норме с помощью лемматизации для сохранения смысловой нагрузки и группирования различных форм слова (например, слова «опоздали» и «опаздывает» приводятся к общей лемме «опаздывать»), что позволяет выполнять более точный семантический и частотный анализ. В заключение удаляются стоп-слова на базе расширенного списка, адаптированного для транспортной отрасли.

Классификация по виду транспорта определяет, к какой сфере относится сообщение (общественный, железнодорожный или воздушный транспорт), а классификация по тематике/проблеме соотносит текст с категорией проблемы (расписание, комфорт, безопасность, цена, работа персонала и т. д.). Это позволяет автоматически тегировать обращения и агрегировать статистику по типам проблем. Данная задача эффективно решается с помощью предварительно обученной модели XLM-RoBERTa. Особенность этой модели в том, что она в первую очередь предназначена для задач, в которых для принятия решений используется все предложение, что в поставленной задаче дает максимальное преимущество.

Анализ тональности определяет эмоциональную окраску высказываний: позитивная, нейтральная или негативная. Для русского языка используется предобученная модель RuBERT-tiny2-russian-sentiment, дополнительно обученная на датасете, размеченном экспертами в транспортной области с использованием фреймворка Transformers.

Извлечение именованных сущностей выделяет в тексте значимые объекты, такие как названия перевозчиков и подрядчиков, станции метро, вокзалы, аэропорты, улицы, города, номера поездов, номера маршрутов автобусов, авиарейсы и многие другие. Для данной задачи на русском языке используется модель RuBERT-base-cased.

Заключительным этапом конвейера сбора и анализа текстов является векторизация датасета с целью дальнейшего выполнения задач анализа и принятия решения. Библиотека Sentence-Transformers преобразует текст в эмбеддинги. Эмбеддинги вместе с исходным текстом, метаданными (источник, время) и результатами классификации сохраняются в базе данных. Для хранения используется система управления базами данных (СУБД) PostgreSQL с расширением pgvector. Это позволяет в рамках одной СУБД выполнять как обычные запросы по метаданным, так и семантический поиск по векторной близости.

Программная архитектура и реализация

Система реализуется в виде набора микросервисов, управляемых Docker Compose, что обеспечивает модульность и масштабируемость (рис. 2).

Сервис сбора данных — набор скриптов на Python, использующих Telethon для парсинга каналов и размещения в очередь для обработки.

Сервис предобработки принимает сырой текст, выполняет очистку и лемматизацию.

Сервис машинного обучения — это отдельные сервисы, обернутые в FastAPI, который предоставляет REST API для классификации, анализа тональности и извлечения именованных сущностей. Внутри используют загруженные модели из Hugging Face: для бинарной классификации —

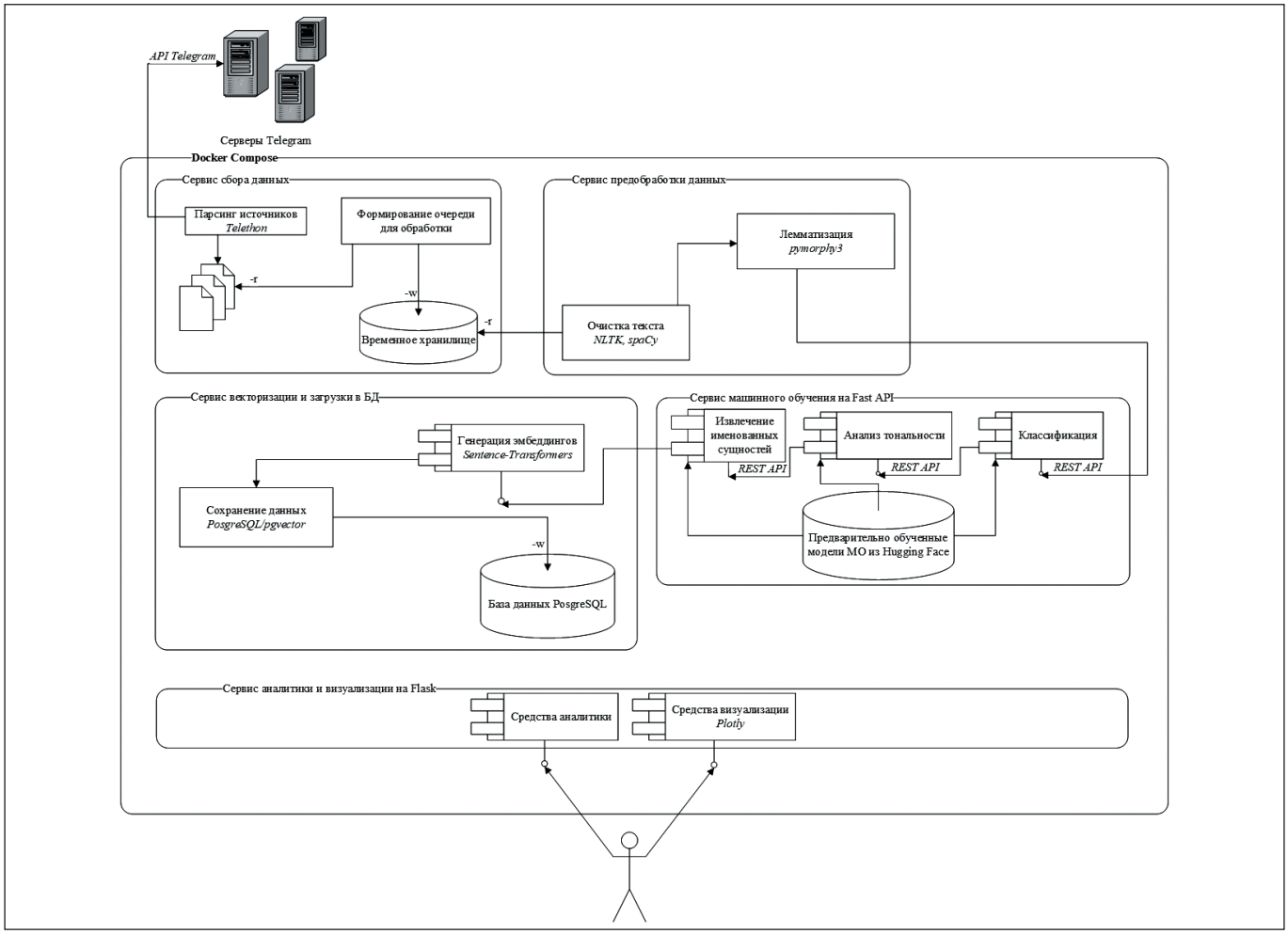

Рис. 2. Схема работы системы обработки данных и взаимодействия с пользователями

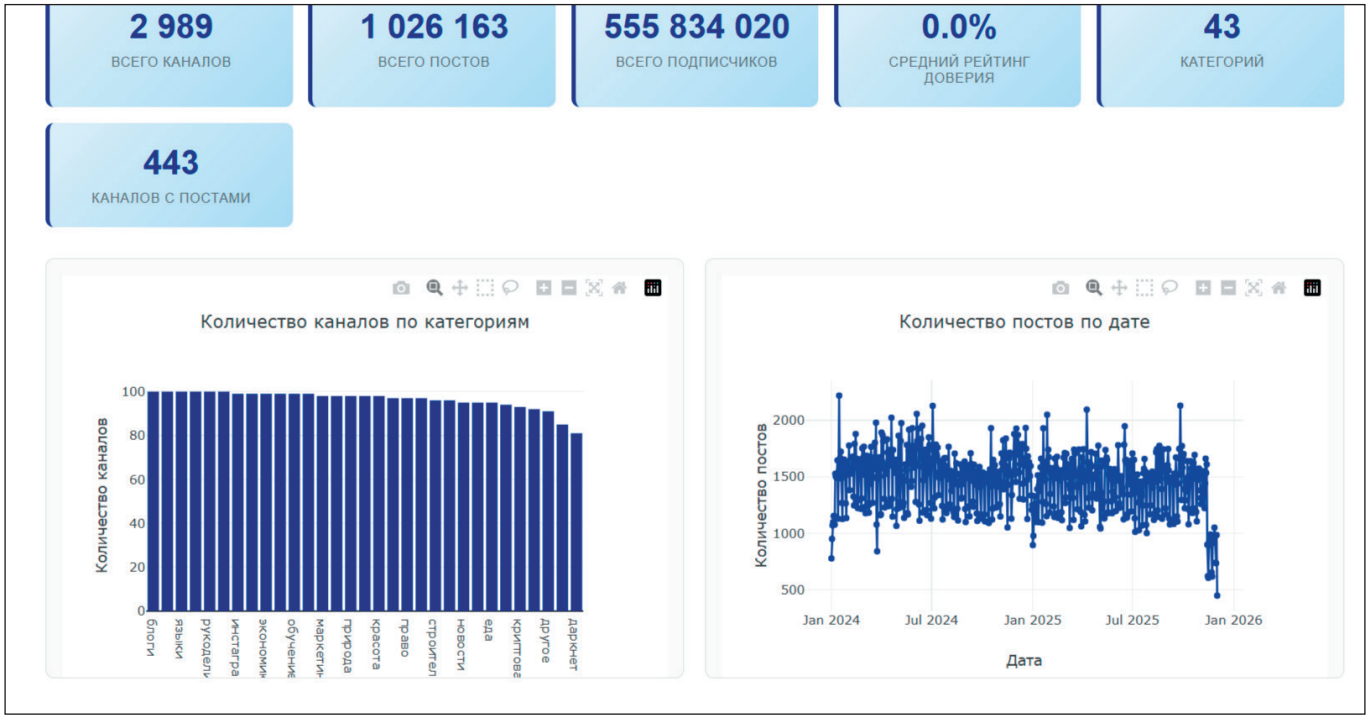

Рис. 3. Страница веб-приложения, предназначенная для отображения инструмента визуализации данных

XLM-RoBERTa, для определения тональности — RuBERT-tiny2-russian-sentiment, для извлечения сущностей — RuBERT-base-cased.

Сервис векторизации и загрузки в БД получает очищенный текст и результаты анализа, генерирует эмбеддинг с помощью Sentence-Transformers и сохраняет полную запись в ячейку с расширением pgvector для задачи бинарной классификации «относится к транспорту / не относится к транспорту».

Сервис аналитики и визуализации — веб-приложение на Flask, позволяющее анализировать собранную информацию с помощью инструментов визуализации — дашбордов на основе библиотеки Plotly (рис. 3).

Практическая значимость и внедрение

На рис. 4 представлен пример сообщения одного из открытых телеграм-каналов, показывающего актуальность и практическую значимость реализуемого подхода к анализу информации в транспортной отрасли.

Разработанный подход и прототип системы имеют непосредственное практическое применение для различных субъектов транспортной отрасли.

Транспортным операторам и перевозчикам система обеспечивает оперативное выявление локальных инцидентов и системных проблем, анализ эффективности службы поддержки, мониторинг репутации бренда, оценку реакции пассажиров на новые тарифы или услуги. Данные служат основанием для превентивного ремонта, корректировки расписания, улучшения информирования.

Для городских и государственных органов управления транспортом возможны макроанализ удовлетворенности пассажиров работой всей транспортной системы города или региона, выявление наиболее проблемных узлов и маршрутов, оценка социального эффекта от инфраструктурных проектов. Инструмент позволяет перейти от реактивного к проактивному управлению.

Внедрение подобной системы позволяет перейти от управления интуитивного или основанного на ограниченных опросах к управлению на основе данных.

При расширении источников информации, например, добавлении сообщений для анализа из каналов других мессенджеров, новостных лент (СМИ и городских пабликов) или групп в социальных сетях, достаточно добавить соответствующие

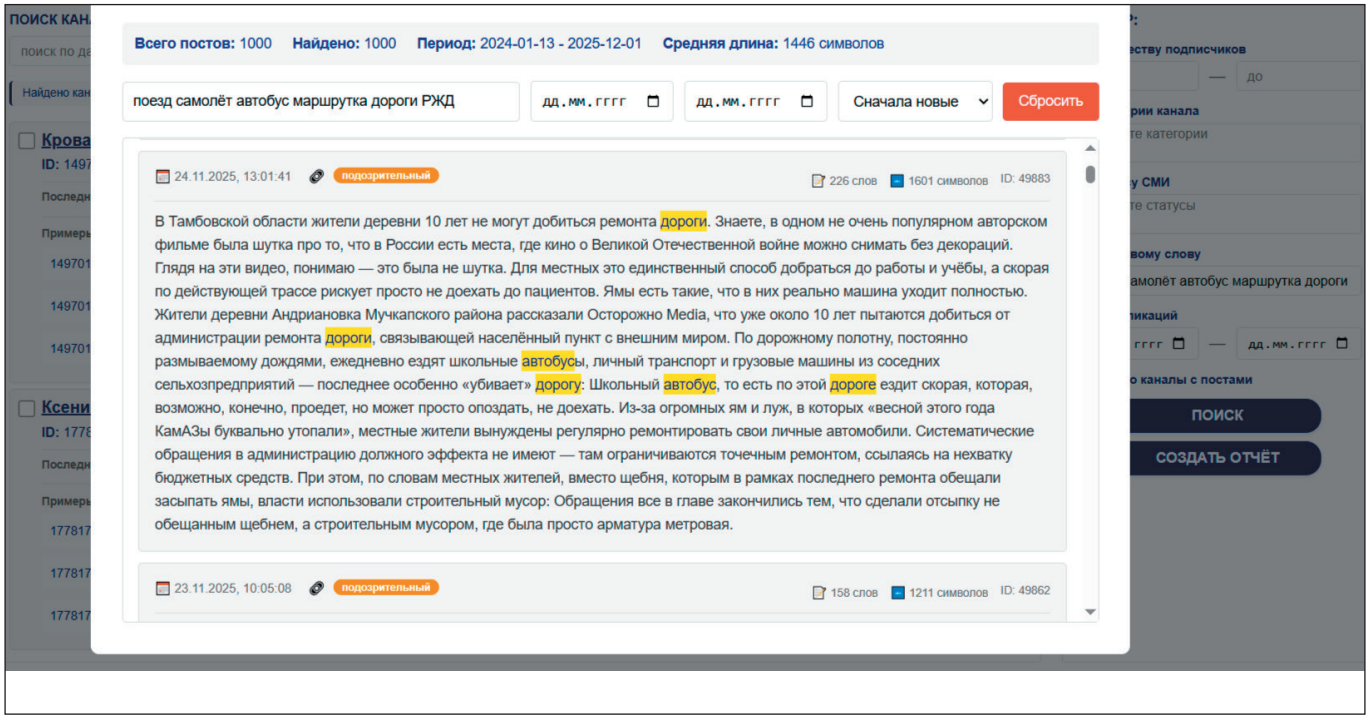

Рис. 4. Страница веб-приложения, предназначенная для отображения результатов поиска по ключевым словам

скрипты в сервис сбора данных. Таким образом, возможно выполнение анализа на большей выборке данных, что способствует формированию статистически корректной картины.

Заключение

Представленный инструмент, построенный на основе конвейера обработки текста, включающего классификацию по виду транспорта и тематике, анализ тональности и извлечение именованных сущностей, позволяет продемонстрировать перспективы применения анализа информации из открытых источников с целью оценки общественного мнения о транспортной отрасли.

За счет реализации системы в виде набора микросервисов, управляемых Docker Compose, обеспечиваются модульность и масштабируемость, что позволяет рассматривать данный программный комплекс в качестве основы инструмента анализа информации, в том числе для ассистирования оператора при решении проблемных вопросов, связанных с качеством обслуживания пассажиров. Направления развития программного комплекса могут быть связаны с внедрением дополнительного функционала, такого как прогнозирование негативной тональности отзывов, анализ поликодового текста, подключение интеллектуальных агентов для формирования рекомендаций по устранению выявленных проблем.

СПИСОК ИСТОЧНИКОВ

1. Цифровые пассажирские сервисы: будущее транспорта обсудили на ЦИПРе // Ассоциация «Цифровой транспорт и логистика». URL: http://www.dtla.ru/news/tsifrovye-passazhirskie-servisy-budushchee-transporta-obsudili-na-tsipre/ (дата обращения: 15.01.2026).

2. 80 % компаний транспортной отрасли используют цифровые технологии, но потенциал роста есть // Цифровая индустрия промышленной России. URL: http://cipr.ru/news/80-kompanij-transportnoj-otrasli-ispolzuyut-czifrovye-tehnologii-no-potenczial-rosta-est (дата обращения: 15.01.2026).

3. Анализ комментариев в социальных сетях и мессенджерах как метод оценки социальной результативности цифровых городских сервисов / О. Г. Филатова [и др.] // International Journal of Open Information Technologies. 2024. Т. 12, № 11. С. 103–110.

4. What do Riders Say and Where? The Detection and Analysis of Eyewitness Transit Tweets / O. Kabbani [et al.] // Journal of Intelligent Transportation Systems. 2023. Vol. 27, iss. 3. Pp. 347–363. DOI: 10.1080/15472450.2022.2026773

5. Lui Y., Li Y., Li W. A Natural Language Processing Approach for Appraisal of Passenger Satisfaction and Service Quality of Public Transportation // IET Intelligent Transport Systems. 2019. Vol. 13, iss. 11. Pp. 1701–1707. DOI: 10.1049/iet-its.2019.0054

6. Максютин П. А., Шульженко С. Н. Обзор методов классификации текстов с помощью машинного обучения // Инженерный вестник Дона. 2022. № 12.

7. Zannat K. E., Choudhury C. F. Emerging Big Data Sources for Public Transport Planning: A Systematic Review on Current State of Art and Future Research Directions // Journal of the Indian Institute of Science. 2019. Vol. 99, iss. 4. Pp. 601–619. DOI: 10.1007/s41745-019-00125-9

8. Chowdhury S., Alzarrad A. Applications of Text Mining in the Transportation Infrastructure Sector: A Review // Information. 2023. Vol. 14, iss. 4. Art. no. 201. DOI: 10.3390/info14040201

9. Коновалова М. В. Когнитивные аспекты верификации в интернет-медиадискурсе // Лингвокультурология. 2019. № 13. С. 125–131.

Дата поступления: 06.04.2026
Решение о публикации: 22.05.2026

A Tool for Collecting and Assessing Public Opinion on the Transport Industry from Open Sources

Nikita R. Strelnikov — 4th year Cadet in 09.05.01 Application and Operation of Automated Systems for Special Purposes. Research interests: natural language processing, machine learning methods, data analysis. E-mail: vka@mil.ru

Svetlana A. Mikhaylova — Lecturer at the Department of mathematical and software support for special space systems. Research interests: natural language processing, machine learning methods, data analysis. E-mail: vka@mil.ru

Nail I. Alimov — PhD in Engineering, Senior Lecturer at the Department of mathematical and software support for special space systems. Research interests: natural language processing, machine learning methods, data analysis. E-mail: vka@mil.ru

Mozhaisky Military Aerospace Academy, 13, Zhdanovskaya str., Saint Petersburg, 197198, Russia

For citation: Strelnikov N. R., Mikhaylova S. A., Alimov N. I. A Tool for Collecting and Assessing Public Opinion on the Transport Industry from Open Sources, *Intellectual Technologies on Transport*, 2026, no. 2 (46), pp. 46–53. DOI: 10.20295/2413-2527-2026-246-46-53 (In Russian)

Abstract. *The study presents the development of a tool for intelligent analysis of passenger feedback on transport.* ***Purpose:*** *is to create an automated monitoring and semantic analysis system aimed at transforming unstructured user statements into structured data for managerial decision-making.* ***Methods:*** *modern information technologies were used, including natural language processing methods, machine learning, and integration of data from heterogeneous sources.* ***Results:*** *emphasize the effectiveness of the proposed text processing pipeline, which includes classification by transport mode and topic, sentiment analysis, and named entity recognition, for identifying key issues and public opinion trends.* ***Practical significance:*** *lies in increasing the responsiveness of transport regulators to passenger problems, optimizing service, and transitioning to data-driven management.*

The research is important for the development of digital technologies in the transport industry and improving the efficiency of passenger transportation management in the context of digital transformation.

Keywords: *data mining, natural language processing, Python, transport, passengers, sentiment, machine learning, media space*

REFERENCES

1. Tsifrovye passazhirskie servisy: budushchee transporta obsudili na TsIPRe [Digital Passenger Services: the Future of Transport Discussed at CIPR], *Sayt Assotsiatsii "Tsifrovoy transport i logistika" [Website of the Digital Transport and Logistics Association]*. Published online on June 02, 2025. Available at: http://www.dtla.ru/news/tsifrovye-passazhirskie-servisy-budushchee-transporta-obsudili-na-tsipre/ (accessed: January 15, 2026). (In Russian)

2. 80 % kompaniy transportnoy otrasli ispolzuyut tsifrovye tekhnologii, no potentsial rosta est [80% of Companies in the Transportation Industry Use Digital Technologies, but There is Growth Potential], *Tsifrovaya industriya promyshlennoy Rossii [Digitalization of Industrial Russia]*. Published online on June 30, 2025. Available at: http://cipr.ru/news/80-kompanij-transportnoj-otrasli-ispolzuyut-czifrovye-tehnologii-no-potenczial-rosta-est/ (accessed: January 15, 2026). (In Russian)

3. Filatova O. G., et al. Analiz kommentariev v sotsialnykh setyakh i messendzherakh kak metod otsenki sotsialnoy rezultativnosti tsifrovykh gorodskikh servisov [Analysis of Comments on Social Networks and Messengers as a Method of Evaluating the Social Effectiveness of Digital Urban Services], *International Journal of Open Information Technologies*, 2024, vol. 12, no. 11, pp. 103–110. (In Russian)

4. Kabbani O., et al. What do Riders Say and Where? The Detection and Analysis of Eyewitness Transit Tweets, *Journal of Intelligent Transportation Systems*, 2023, vol. 27, iss. 3, pp. 347–363. DOI: 10.1080/15472450.2022.2026773

5. Liu Y., Li Y., Li W. A Natural Language Processing Approach for Appraisal of Passenger Satisfaction and Service Quality of Public Transportation, *IET Intelligent Transport Systems*, 2019, vol. 13, iss. 11, pp. 1701–1707. DOI: 10.1049/iet-its.2019.0054

6. Maksyutin P. A., Shuljenko S. N. Obzor metodov klassifikatsii tekstov s pomoshchyu mashinnogo obucheniya [An Overview of Text Classification Methods Using Machine Learning], *Inzhenernyy vestnik Dona [Engineering Journal of Don]*, 2022, no. 12. (In Russian)

7. Zannat K. E., Choudhury C. F. Emerging Big Data Sources for Public Transport Planning: A Systematic Review on Current State of Art and Future Research Directions, *Journal of the Indian Institute of Science*, 2019, vol. 99, iss. 4, pp. 601–619. DOI: 10.1007/s41745-019-00125-9

8. Chowdhury S., Alzarrad A. Applications of Text Mining in the Transportation Infrastructure Sector: A Review, *Information*, 2023, vol. 14, iss. 4, art. no. 201, 24 p. DOI: 10.3390/info14040201

9. Konovalova M. V. Kognitivnye aspekty verifikatsii v Internet-mediadiskurse [Cognitive Aspects of Verification on the Internet Media Discourse], *Lingvokulturologiya*, 2019, no. 13, pp. 125–131. (In Russian)

Received: April 06, 2026
Accepted: May 22, 2026