УДК 656, 004.8, 007.5, 51-74, 510.67 **DOI** 10.15622/sp.18.2.354-389

Я.А. СЕЛИВЕРСТОВ, В.И. ЧИГУР, А.М. САЗАНОВ, С.А. СЕЛИВЕРСТОВ, А.С. СВИСТУНОВА

РАЗРАБОТКА СИСТЕМЫ ДЛЯ ТОНОВОГО АНАЛИЗА ОТЗЫВОВ ПОЛЬЗОВАТЕЛЕЙ ПОРТАЛА «AUTOSTRADA.INFO/RU»

Селиверстов Я.А., Чигур В.И., Сазанов А.М, Селиверстов С.А., Свистунова А.С. **Разработка системы для тонового анализа отзывов пользователей портала «AUTOSTRADA.INFO/RU».**

Аннотация. Социальные сети (Вконтакте, Facebook), тематические сообщества в сетях микроблогинга (Twitter), ресурсы для путешественников (TripAdvisor) и транспортные порталы (Autostrada) являются источником актуальной и оперативной информации о дорожно-транспортной обстановке, качестве предоставляемых транспортных услуг и степени удовлетворенности пассажиров уровнем транспортного обслуживания. Однако существующие системы транспортного мониторинга не содержат программных инструментов, способных осуществлять сбор и анализ дорожно-транспортной информации в среде Интернет. В настоящей работе рассматривается задача построения системы автоматического извлечения и классификации дорожно-транспортной информации с транспортных интернет-порталов и апробация разработанной системы для анализа транспортных сетей Крыма и города Севастополя. Для решения этой задачи проанализированы библиотеки с открытым исходным кодом для тематического сбора и исследования данных. Разработан алгоритм для извлечения и анализа текстов. Осуществлена разработка краулера с использованием пакета Scrapy на языке Python3 и собраны отзывы пользователей с портала http://autostrada.info/ru о состоянии транспортной системы Крыма и города Севастополя. Для лемматизации текстов и векторного преобразования текстов были рассмотрены методы tf, idf, tf-idf и их реализация в библиотеке Scikit-Learn: CountVectorizer и TF-IDF Vectorizer. Для обработки текстов были рассмотрены методы Bag-of-Words и n-gram. В ходе разработки модели классификатора рассмотрены наивный байесовский алгоритм (MultinomialNB) и модель линейного классификатора с оптимизацией стохастического градиентного спуска (SGDClassifier). В качестве обучающей выборки использовался корпус объемом 225 тысяч размеченных текстов с ресурса Twitter. Проведено обучение классификатора, в ходе которого использовалась стратегия кросс-валидации и метод ShuffleSplit. Проведено тестирование и сравнение результатов тоновой классификации. По результатам валидации лучшей оказалась линейная модель со схемой n-грамм [1, 3] и векторизатором TF-IDF. В ходе апробации разработанной системы был проведен сбор и анализ отзывов, относящихся к качеству транспортных сетей республики Крым и города Севастополя. Сделаны выводы и определены перспективы дальнейшего функционального развития разрабатываемого инструментария.

Ключевые слова: автоматический анализ текстов, краулеры, классификация текстов, интеллектуальные транспортные системы, машинное обучение, TF-IDF, наивный байесовский алгоритм, линейный классификатор, анализ тональности.

1. Введение. Стремительное развитие мобильных и облачных технологий, перевод логистической, потребительской, коммуникационной и расчетно-денежной деятельностей в

информационно-сетевое пространство открывает новые пути развития интеллектуальных транспортных систем (ИТС) и систем транспортного мониторинга.

Работа современных ИТС [1, 2] строится на данных, получаемых с систем видео-мониторинга [3], а также информации о местоположении пользователей и транспортных средств [4], которая собирается с использованием мобильных устройств, поддерживающих GPS/WiFi/Lte/WPAN стандарты передачи данных [5, 6].

Стоимость систем видео-мониторинга сравнительно высока, поэтому их размещают только на особо загруженных участках улично-дорожных сетей крупных городов и мегаполисов. Улично-дорожные сети небольших городов и поселков, а также региональные и областные транспортные сети остаются не охваченными системами видео-мониторинга, и как следствие, информация о дорожно-транспортной обстановке на них отсутствует.

Мобильные телефоны с доступом в Интернет имеются, как правило, у каждого водителя. В случае обнаружения проблемных участков на дороге или дорожно-транспортных происшествий водитель способен зафиксировать эту информацию в виде отзыва на специализированном интернет-портале.

Таким образом, одним из источников разнородной информации, относящейся к сфере транспорта, может быть web-пространство.

Транспортные данные в web-пространстве, как правило, структурированы и разбросаны по тематическим интернет-ресурсам.

К таковым относят: отраслевые сайты (http://autostrada.info/ru), тематические интернет сообщества (https://www.worldoftrucks.com/en/), группы в социальных сетях (Вконтакте, Facebook) и сетях микроблогинга (Twitter), а также чаты и форумы.

Информация на транспортных web-порталах и тематических интернет-сообществах формируется в виде отзывов непосредственно самими пользователями, поэтому для ее сбора не требуется больших затрат. Тема web-портала или интернет сообщества определяет характер размещаемой информации, например, если тематика группы «пробки», то, как правило, размещаемые пользователями отзывы содержат сведения о пробках и заторах на дорогах. Если же тематика группы «поборы на дорогах», то размещаемые пользователями отзывы содержат сведения о недобросовестной работе сотрудников весового контроля или служителей правопорядка.

Таким образом, каждой теме может быть поставлена в соответствие некоторая характеристика или фактор, оказывающий влияние на транспортный процесс и дорожные условия. Такое

структурирование информации упрощает процесс составления тематических корпусов в области транспорта, что, в свою очередь, позволяет строить более глубокие системы классификации транспортных данных и выявлять на их основе новые управляющие воздействия.

Таким образом, использование систем извлечения и анализа дорожно-транспортной информации из web-пространства в качестве систем транспортного мониторинга [7-9] открывает новые каналы поступления транспортной информации, способной повысить информированность участников дорожного движения о состоянии транспортных сетей и условий дорожного движения.

2. Анализ предметной области. Проанализируем последние публикации, в которых рассматриваются методы извлечения и анализа текстов, относящихся к транспортной сфере.

В работе [10] рассматриваются методы построения автоматического классификатора для анализа вопросов и ответов путешественников на сайте TripAdvisor в разделе Q-A форума «О поиске оптимальных маршрутов в разных городах». Целью данного исследования являлось построение вопросно-ответной системы. В работе [11] представлена методология сбора данных из социальной сети Twitter с использованием данных геотегированной службы Twitter. Собранные данные содержат информацию о моделях мобильности и поведенческих характеристиках пользователей. В работе [12] из сообщений социальной сети Twitter вычленяется информация о качестве транспортного обслуживания пассажиров городского транспорта, пробках, сбоях в расписании, минировании вокзалов и транспортных средств, авариях и ДТП — данная информация использовалась для повышения качества транспортного обслуживания в период чемпионата мира по футболу. В работе [13] разрабатывается программный модуль для ИТС на основе нечеткой онтологии, на основе правил логического вывода, семантического анализа и анализа тональностей текстов из социальных сетей (Twitter, Facebook), городских транспортных порталов, сайтов для путешественников и туристов. Данная система эффективно извлекает отзывы и сообщения, связанные с особенностями городской среды (например, автобусные и железнодорожные вокзалы, мосты, парки, рестораны, аэропорты, медицинские центры и гостиницы) и транспортными инцидентами (например, столкновениями, плохими дорогами, скоплениями и пробками). В работе [14] разрабатывается метод автоматической обработки информации о путешествиях из туристических блогов, которые

определены как туристические журналы, написанные блоггерами в дневниковой форме. В работе показано, что блоги о путешествиях — это полезный источник транспортной информации для туристов. В работе [15] разрабатываются эффективные методы сбора, извлечения и передачи данных из социальных сетей (Twitter) для информационного обеспечения интеллектуальных систем управления дорожным движением, продвинутых систем поддержки путешественников (support advanced traveler information systems) и транспортных диспетчерских центров. В работе [16] исследуется вклад семантического и семантико-синтаксического видов анализа на эффективность решения прикладных задач обработки текстов: вопросно-ответного поиска и извлечения определений из научных публикаций в области транспорта. В работе [17] рассматриваются методы, используемые для обнаружения экстремистских текстов из Интернета. В статье [18] рассматривается система для сбора, обработки и фильтрации сообщений пользователей, связанных с дорожно-транспортными происшествиями и авариями. Сообщения получены из Twitter с использованием REST API в режиме реального времени. В процессе адаптивного сбора данных формируется корпус текстов, состоящий из важных ключевых слов и их комбинаций, которые могут подразумевать дорожные происшествия. Затем сообщение преобразуется в бинарный вектор в пространстве признаков и классифицируется как относящийся к транспортному происшествию или нет. Все сообщения, относящиеся к транспортному происшествию, геокодируются для определения их местоположения и далее классифицируются в одну из пяти категорий инцидентов. Данная система была успешно протестирована в двух регионах: Питтсбурге и Филадельфии. В работе [19] разрабатывается система для автоматического сбора актуальной и ценной информации из Twitter, связанной с транспортом и транспортными услугами, во время проведения футбольных матчей. В работе успешно протестированы методы автоматического сбора и идентификации сообщений на выборках свыше 3.7 миллионов сообщений. В работе [20] анализируются общественные системы обмена велосипедами в Испании через анализ настроений в социальных сетях с учетом мнений жителей и посетителей, также выявляются положительные и отрицательные факторы и определяется их потенциальное влияние на качество туристических и транспортных услуг. В процессе анализа были обработаны данные из 46 систем обмена велосипедами за период 2010-2016 годов. Результаты исследования показывают, что

инфраструктура туризма и транспорта, включающая велосипедные дорожки, станции и обеспечение eBike, нуждается в скоординированном планировании, поскольку она неразрывно связана с уровнем туристического и транспортного обслуживания. Социальные сети в рамках данного исследования показали себя как достоверный источник транспортной информации. В статье [21] рассматривается система сбора и анализа отзывов туристов и путешественников с сайта TripAdvisor.com с учетом их географического местоположения, а также статистических и территориальных данных. Методы анализа текстов применяются для оценки восприятия туристами положительных факторов (мест, событий, достопримечательностей), которые могут использоваться в качестве инструментов поддержки планирования поездок. Исследование проведено на примере Хорватии. Результаты исследования раскрывают ценность и взаимодополняемость данных, связанных с социальными сетями, с официальной статистикой планирования транспорта и туризма.

Анализ предметной области показал, что передовые системы для извлечения и анализа тематических текстов активно внедряются в системы городского транспортного мониторинга и системы поддержки туристической и транспортной мобильности.

3. Постановка задачи. Целью настоящей статьи является разработка и тестирование системы, способной анализировать тексты с транспортных web-порталов. В качестве интернет-ресурса выбран web-портал http://autostrada.info/ru, в качестве методов контент-анализа — анализ тональности; в качестве шкалы классификации полярности документа — бинарная шкала с двумя классами оценок: положительные и отрицательные; в качестве оцениваемого объекта выбраны транспортные сети республики Крым и города Севастополь.

Предполагается выполнить следующий перечень работ:

1) Разработать схему алгоритма для извлечения и анализа текстов.

2) Программно реализовать функционал алгоритма для сбора текстов по дорожно-транспортной проблематике.

3) Протестировать разработанную программу и собрать тексты с сайта http://autostrada.info/ru.

4) Сформировать корпуса текстов для последующего обучения классификатора.

5) Разработать тоновый классификатор.

6) Обучить классификатор и оценить его работу.

4. Анализ фреймворков для получения данных из Web. Одна из главных задач тематического веб-краулера — поиск и добавление в коллекцию документов наиболее значимых информационных источников, что обеспечивает создание коллекции высокого качества [22]. Уже существует широкий ассортимент известных библиотек. Это позволяет не писать с нуля новые поисковые боты [23].

Анализ библиотек с открытым исходным кодом из списка TOP-50 определил наиболее функциональные для нашей системы фреймворки.

На основе анализа, представленного в таблице 1, под такие характеристики подходит фреймворк Scrapy.

Scrapy — одна из наиболее популярных и производительных библиотек Python для получения данных с веб-страниц. Фреймворк Scrapy является сфокусированным, легко устанавливается, поддерживает выгрузку данных в форматах JSON, XML, CSV.

Таблица 1. Анализ библиотек (фреймворков) для получения данных из web

Название	Описание	Источник
Heritrix	Гибкий, расширяемый, надежный и масштабируемый фреймворк, написанный на Java и способный получать, архивировать и анализировать тексты. Heretrix работает в распределенной среде с помощью хеширования URL хостов.	[18, 19]
Nutch	Представляет собой инкрементный, параллельный, распределенный, кроссплатформенный модульный фреймворк для построения поисковых систем, написанный на java. Поддерживает граф связей узлов, различные фильтры и нормализацию URL.	
Scrapy	Расширяемый, сфокусированный, параллельный, кроссплатформенный и гибкий фреймворк-библиотека для Python. Легко устанавливается, поддерживает выгрузку данных в форматах JSON, XML, CSV. Широко используется для веб-скрайбинга, не имеет встроенных функций для работы в распределенной среде.	

5. Разработка алгоритма для извлечения и анализа тематических текстов. Построение системы для извлечения и анализа тематических текстов начинается с разработки обобщенного алгоритма.

Алгоритм в общем виде состоит из процедур, представленных в таблице 2, а схема алгоритма представлена на рисунке 1.

Таблица 2. Общий вид алгоритма для извлечения и анализа тематических текстов

Процедура алгоритма	Наименование процедуры
1	Формирование очереди ссылок, подаваемых на вход краулера
2	Список источников добавляются в очередь обхода краулера
3	Клаулер сканирует страницу из очереди
4	Краулер скачивает интересующий его веб-документ в базу данных
5	Проводится очистка веб-документа от «мусора»
6	Производится сохранение очищенного текста в базу данных
7	Подготовка коллекций, ручная разметка текстов и построение корпуса театических текстом
8	Запуск классификатора тональности
9	Обучение классификатора на различных корпусах текстов
10	Оценка классификатора тональности

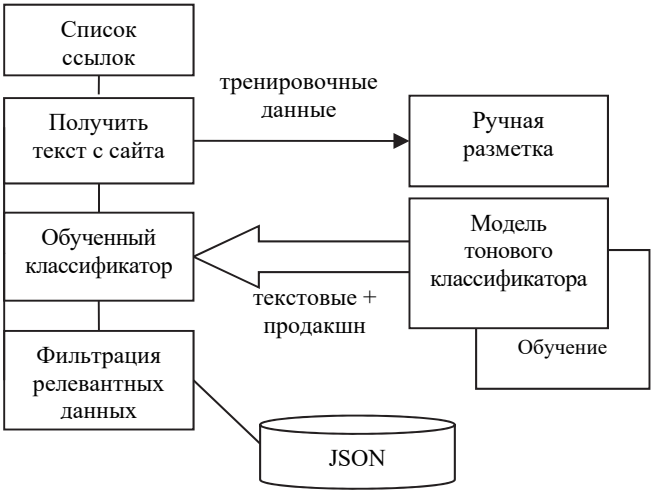

Рис. 1. Схема системы для извлечения и анализа тематических текстов

6. Разработка краулер-модуля. На первом этапе исследования разрабатывался краулер-модуль. Краулер-модуль выполняет процедуры 1-4 алгоритма (таблица 2), а именно: 1) формирует очередь ссылок; 2) добавляет список источников в очередь обхода;

3) сканирует страницу из очереди; 4) скачивает интересующий его web-документ в базу данных.

Часть листинга программы краулер-модуля представлена на листинге 1.

```
importscrapy
classRoadSpider(scrapy.Spider):
name = 'road_spider'
start_urls = [
'http://autostrada.info/ru/reviews/page/1/',
]
def parse(self, response):
for review in response.css('div.col-md-12.reviewBlock'):
        tmp = review.css('p.comment.break-word::text').extract_first()
        tmp1 = review.css('a.label.label-code::text').extract_first()
        tmp2 = review.css('a.highwayLabel::text').extract_first()
        tmp = tmp.replace('\r\n', ' ')
        tmp = tmp.replace('\n', '')
        dd = {
                'title': tmp1 + ' ' + tmp2,
                        'subtitle': review.css('div.col-sm-8.b-rate.hidden-xs
                    b::text').extract_first(),
                'date': review.css('strong.reviewDate::text').extract_first(),
                'rate': review.css('span.b-stars::attr(title)').extract_first(),
      'description': tmp,
                }
        try:
                        dd['date'] = dd['date'].replace('\t', '')
                dd['date'] = dd['date'].replace('\n', '')
                dd['date'] = dd['date'].replace('\u0433.', '')
        except:
                pass
        yielddd
```

Листинг 1. Часть программы краулер-модуля

В процессе работы краулера с сайта http://autostrada.info/ru извлекаются мнения пользователей в текстовом виде.

В результате работы краулер-модуля был собран корпус, содержащий 1130 текстов за период с 01 марта 2009 года по 1 ноября 2018 года с сайта http://autostrada.info/ru. Рассмотрим несколько примеров текстов корпуса и того, что в них содержится.

На рисунке 2 представлен пример отзыва с сайта http://autostrada.info/ru о состоянии участка трассы, пролегающий между Феодосией и Керчью.

Рис. 2. Отзыв о состоянии дорог на сайте http://autostrada.info/ru

Структура отзывов на сайте autostrada представлена на рисунке 3.

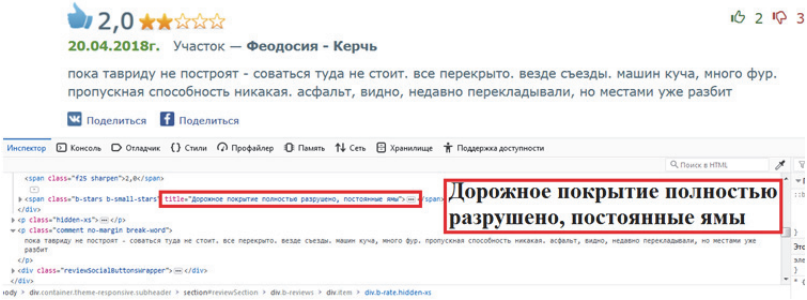

Рис. 3. Структура отзыва на сайте http://autostrada.info/ru

Извлеченный текст записывается в базу данных с указанием атрибутов: date (дата и время создания отзыва), description (описание ситуации), subtitle (наименование трасы), title (кодифицированное название трассы) и url (адрес отзыва в интернет).

Например, для отзыва, представленного на рисунке 4, атрибуты имеют вид: *date: «14.07.2018 14:18»; description: «Асфальт новый, но уже много участков с колеёй. Дорожные строители параллельно, рядом, строят магистраль «Таврида», поэтому много объездов строящихся мостов и развязок, а так-же участков с ограничением скорости. В добавок, временными ограждениями проезжая часть заужена так, что со встречной машиной разъёзжаешься «впритирку». И ограждениями-же закрыт доступ к обочинам. Поэтому, если кто остановился в потоке – сразу пробка, ведь дорога очень перегружена, в том числе и самосвалами и техникой строителей дороги»; Subtitle: «Феодосия-Керчь»; title: «М-17 Херсон – Джанкой – Керчь».*

```
{ 'date': '14.07.2018 14:18 ' ,
  'description': 'Асфальт новый, но уже много участков с колеёй. Дорожные строители  '
                 ' параллельно, рядом, строят магистраль "Таврида", поэтому много объездов  '
                 ' строящихся мостов и развязок, а так-же участков с ограничением скорости. В '
                 'добавок, временными ограждениями проезжая часть заужена так, что со '
                 'встречной машиной разъёзжаешься "впритирку". И ограждениями-же закрыт '
                 'доступ к обочинам. Поэтому, если кто остановился в потоке -- сразу пробка, ведь '
                 'дорога очень перегружена, в том числе и самосвалами и техникой строителей '
                 'дороги. ' ,
  'subtitle': 'Феодосия - Керчь',
  'title': 'М-17 Херсон – Джанкой – Керчь',
  'rate': 'Дорожное покрытие полностью разрушено, постоянные ямы' ,
  'url': http://autostrada.info/ua/highway/M-17 '}
```

Рис. 4. База данных с текстами по дорожно-транспортной проблематике

Далее осуществляются процедуры 5 и 6 алгоритма (см. таблица 2). Все собранные краулером отзывы группируются в единый текст и подвергаются процедуре предобработки: слова приводятся к нижнему регистру, затем отсеиваются все вспомогательные символы, такие как знаки препинания и стоп-слова.

Далее с помощью библиотеки pymorhy2 слова приводятся к нормальной форме. На следующем этапе осуществляется векторизация [24] и производится лексический анализ текста.

7. Лексический анализ и векторизация текста. Перед тем как использовать машинное обучение на текстовых документах необходимо перевести текстовое содержимое в числовой вектор признаков с учетом tf, idf и tf-idf [25, 26]. Векторизатор строит словарь индексов признаков.

В более сложных моделях [27] используют алгоритмы семантико-синтаксического анализа [28] и токенизации [29] с возможностью настройки тонового анализа [30].

Для обработки текстов целесообразно использовать два метода: CountVectorizer и TFIDFVectorizer [31]. Ниже будет дано обоснование использования данных методов. Оба метода используют модель Bag of Words [32]. Листинг программы, выполняющей лексический анализ, векторизацию и индексирование текста, представлен на листинге 2.

```
fromsklearn.feature_extraction.text import CountVectorizer
fromsklearn.feature_extraction.text import TfidfVectorizer
fromsklearn.grid_search import GridSearchCV
fromsklearn.cross_validation import ShuffleSplit, cross_val_score
importpandasaspd
# считываемподготовленныйдатасет
        dataset = pd.read_csv('data/cleaned_data.csv', index_col=0).dropna()
```

```
# массив n-граммных схем, которые будут использоваться в работе
# например, (1, 3) означает униграммы + биграммы + триграммы
ngram_schemes = [(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6)]
forngram_scheme in ngram_schemes:
print('N-gram Scheme:', ngram_scheme)
count_vectorizer = CountVectorizer(analyzer = "word", ngram_range=ngram_scheme)
tfidf_vectorizer = TfidfVectorizer(analyzer = "word", ngram_range=ngram_scheme)
vectorizers = [count_vectorizer, tfidf_vectorizer]
vectorizers_names = ['Count Vectorizer', 'TF-IDF Vectorizer']
for i in range(len(vectorizers)):
        print(vectorizers_names[i])
        vectorizer = vectorizers[i]
        X = vectorizer.fit_transform(dataset['text'])
        y = dataset['label']
        cv = ShuffleSplit(len(y), n_iter=5, test_size=0.3, random_state=0)
```

Листинг 2. Лексический анализ и векторизация текста

Рассмотрим более подробно используемые выше методы и обоснуем их выбор.

7.1. Метод Bag of Words. Математическая модель Bag of Words (перевод с англ. — мешок слов) — это модель обработки текста, при котором слова выбираются в случайном порядке.

Модель Bag of Words [33] позволяет перейти к компактному представлению документа, в котором любое слово $w_t \in V$ словаря V в документе d_i имеет количество вхождений равное n_t, следовательно, любой документ d_i может быть представлен вектором в виде [32]:

$$\overline{d}_i = \left(n_1\left(w_1\right) + n_{\ t}\left(w_t\right) + \cdots + n_m\left(w_m\right)\right), \tag{1}$$

где m — количество слов в документе d_i.

Как правило, выделяют два основных типа атрибутов:

1. Частотные атрибуты — когда каждое значение в векторе $\overline{d}$ соответствует количеству вхождений признаков (слов) в документ d; тогда $n_i(d) \in (0;+\infty)$;

2. Бинарные (наличия/отсутствия) атрибуты — когда каждое значение в векторе $\overline{d}$ бинарное (true/false или 0/1) и отражает факт присутствия признака w_i в документе d, тогда $n_i(d) \in \{0;1\}$.

Алгоритм построения модели следующий: 1) составляется словарь терминов из всех слов, встречающихся в тексте, при этом из

текста предварительно исключаются все знаки препинания, числа и «стоп-слова»; 2) для каждого документа определяется вектор, каждая компонента которого соответствует термину из словаря, а ее значение определяется числом, характеризующим сколько раз это слово встретилось в тексте. Размерность вектора соответствует мощности словаря.

Такой подход довольно распространен и прост в реализации, но он не избавлен от недостатков. Например, отзыв «трасса не очень хорошая» имеет негативную тональность, однако, если рассматривать каждое слово по отдельности, невозможно будет это определить. Кроме того, модель, вероятно, «выучит», что слово «хороший» имеет положительную тональность, но в данном случае это не то, что требуется.

Проблема определения смыслового окраса текста может быть решена с помощью метода n-gram. Обычно для таких задач используют схемы с униграмными, биграмными или тригамными признаками и их совместные комбинации независимо друг от друга.

7.2. Метод *n*-gram. Математическая модель n-gram — это модель представления текстов в виде набора последовательностей, состоящих из N слов. Различают следующие модели n-gram: 1 слово — униграммы, при которой определяется вероятность $P(w_i)$ появления i-го слова (w_i) в тексте; 2 слова — биграммы, при которой определяется вероятность появления пар слов $P(w_i|w_{i-1})$ в тексте, 3 слова — триграммы, при которой определяется вероятность появления троек слов $P(w_i|w_{i-2}, w_{i-1})$ в тексте [34, 35].

Таким образом, задача сводится к определению вероятности появления цепочки слов $V_m = (w_1, w_2, \ldots, w_t)$ в некотором тексте d_m.

Вероятность $P(w_1, w_2, \ldots, w_t)$ можно представить в виде произведения условных вероятностей входящих в нее n-gram [27]:

$$P(w_1, w_2, \ldots, w_t) = \prod_{i=1}^{t} P(w_i | w_1, w_2, \ldots, w_{i-1}), \quad (2)$$

или аппроксимируя $P(w)$ при ограниченном контексте длинной (n-l), согласно [34]:

$$P(w_1, w_2, \ldots, w_t) \cong \prod_{i=1}^{t} P(w_i | w_{i-n+1}, w_{i-n+2}, \ldots, w_{i-1}). \quad (3)$$

Вероятность появления n-gram вычисляется согласно [34]:

$$P\left(w_i \middle| w_{i-n+1},\ldots,w_{i-1}\right) = \frac{C\left(w_{i-n+1},\ldots,w_i\right)}{C\left(w_{i-n+1},\ldots,w_{i-1}\right)}, \quad (4)$$

где C — количество появлений последовательности слов в обучающем корпусе.

В процессе подбора лучших параметров для модели рассматриваются схемы n-gram от 1 до 5. Как правило, в длинных документах среднее количество словоупотреблений будет выше, чем в коротких, даже если они посвящены одной теме. Чтобы избежать этих потенциальных несоответствий, достаточно разделить количество употреблений каждого слова в документе на общее количество слов. Этот признак называется «частота термина» или Term Frequency [29].

7.3. Мера Term Frequency. Частота термина — отношение числа вхождений некоторого слова к общему числу слов документа, при которой оценивается важность слова w_i в пределах отдельного документа [25]:

$$tf\left(w_i;d\right) = \frac{n_t}{\sum_i n_i}. \quad (5)$$

Следующим уточняющим параметром словоупотреблений является мера обратной частоты документа (Inverse Document Frequency).

7.4. Мера Inverse Document Frequency. Обратная частота документа **idf** — инверсия частоты, с которой некоторое слово встречается в документах коллекции [25]:

$$idf\left(w_i;D\right) = \log\frac{D}{\left|\left\{d_i \in D \middle| w_i \in d_i\right\}\right|}, \quad (6)$$

где $\left|D\right|$ — число документов в коллекции; $\left|\left\{d_i \in D \middle| w_i \in d_i\right\}\right|$ — число документов из коллекции D в которой встречается слово w_i.

Использование меры idf позволяет снизить вес широкоупотребительных слов, которые являются менее информативным, чем те, которые используются только в небольшой части.

Примером низко информативных слов могут служить служебные слова, артикли, предлоги, союзы.

На последнем этапе вычисляется ключевая характеристика tf-idf, определяющая перечень уникальных слов однозначно определяющих данный документ.

7.5. Мера Term Frequency-Inverse Document Frequency. Частота термина-обратная частота документа $tf-idf$ — статистическая мера, которая используется для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса [32]:

$$tf-idf = tf_{i,j} \ln\left(\frac{N}{df_i}\right), \tag{7}$$

где $tf_{i,j}$ — отношение количества вхождений слова к общему числу терминов документа, df_i — число документов из коллекции, в которых встречается слово, N — число документов в коллекции.

Таким образом, вес некоторого слова пропорционален количеству употреблений этого слова в документе и обратно пропорционален частоте употребления слова в других документах коллекции. Следовательно, если слово часто встречается в каком-либо документе и при этом встречается редко в других документах, то это слово имеет большую значимость для анализируемого документа.

Следовательно, именно меру $tf-idf$ целесообразно использовать в разработке системы извлечения и анализа тематических текстов. В программном исполнении $tf-idf$ реализован в библиотеке Scikit-Learn в виде стандартного метода векторизатора TF-IDF Vectorizer [31].

Рассмотрев основные методы анализа текстов, перейдем к построению классификатора тональности.

8. Разработка тонового классификатора. Для построения модели тонового классификатора рассмотрим и сравним две наиболее используемые модели классификации: наивный байесовский классификатор и линейный классификатор на основе стахостического градиента. Как известно, существуют и другие часто используемые методы, которые являются эффективными для различных задач классификации, например метод опорных векторов (SVM) или сверточные нейронные сети (CNN). Данные методы будут рассмотрены в рамках следующих исследований.

В программном исполнении наивный байесовский классификатор реализован в библиотеке Scikit-Learn в виде

стандартного метода MultinomialNB, а линейный классификатор на основе стахостического градиента в виде — SGDClassifier [31].

Листинг программы тонового классификатора на основе стандартных методов MultinomialNB и SGDClassifier классификаторов представлен на листинге 3.

```
# Наивный байес
clf = MultinomialNB()
NB_result = cross_val_score(clf, X, y, cv=cv).mean()
# Линейный классификатор
clf = SGDClassifier()
parameters = {
'loss': ('log', 'hinge'),
'penalty': ['none', 'l1', 'l2', 'elasticnet'],
'alpha': [0.001, 0.0001, 0.00001, 0.000001]
}
gs_clf = GridSearchCV(clf, parameters, cv=cv, n_jobs=-1)
gs_clf = gs_clf.fit(X, y)
L_result = gs_clf.best_score_
```

Листинг 3. Листинг программы тонового классификатора

Рассмотрим более подробно модели отобранных классификаторов и обоснуем их выбор.

8.1. Наивный байесовский классификатор. Существуют два подхода к наивному байесовскому классификатору — мультиномиальный и многомерный, которые дают разные результаты.

Определим, какой из подходов лучше использовать для классификации текстов в данном случае [29].

Многомерная модель (https://docplayer.ru/45424867-Naivnyy-bayesovskiy-klassifikator.html): пусть $V = \{w_t\}_{t=1}^{|V|}$ — словарь; тогда документ d_i — это вектор длины $|V|$, состоящий из битов B_{it}; $B_{it} = 1$, если слово w_t встречается в документе d_i.

Правдоподобие принадлежности d_i классу c_j рассчитывается согласно [36]:

$$p\left(d_i \middle| c_j\right) = \prod_{t=1}^{|V|}\left(B_{it}\, p\left(w_t \middle| c_j\right) + \left(1 - B_{it}\right)\left(1 - p\left(w_t \middle| c_j\right)\right)\right). \quad (8)$$

Для обучения такого классификатора нужно получить вероятности $p\left(w_t \middle| c_j\right)$. Рассмотрим процесс обучения.

Обучение: пусть дан набор документов $D = \{d_i\}_{i=1}^{|V|}$, которые распределены по классам c_j, дан словарь $V = \{w_t\}_{t=1}^{|V|}$ и заданы биты документов B_{it}.

Тогда можно подсчитать оценки вероятностей того, что-то или иное слово встречается в том или ином классе [36]:

$$p(w_t|c_j) = \frac{1 + \sum_{i=1}^{|D|} B_{it} p(c_j|d_i)}{2 + \sum_{i=1}^{|D|} p(c_j|d_i)}. \quad (9)$$

Априорные вероятности классов рассчитываются в соответствии с [36]:

$$p(c_j) = \frac{1}{|D|} \sum_{i=1}^{|D|} p(c_j|d_j). \quad (10)$$

Тогда классификация будет происходить в соответствии с [36]:

$$c = \arg\max_j p(c_j) p(d_j|c_j) =$$

$$= \arg\max_j \left(\frac{1}{|D|} \sum_{i=1}^{|D|} p(c_j|d_i) \right) \prod_{t=1}^{|V|} \left(B_{it} p(w_t|c_j) + (1 - B_{it})(1 - p(w_t|c_j)) \right) = \quad (11)$$

$$= \arg\max_j \left(\log \left(\sum_{i=1}^{|D|} p(c_j|d_i) \right) + \sum_{t=1}^{|V|} \log \left(B_{it} p(w_t|c_j) + (1 - B_{it})(1 - p(w_t|c_j)) \right) \right).$$

В мультиномиальной модели документ — это последовательность слов, отобранных методом «Bag of Words» [37]. Данный метод был рассмотрен выше. Для подсчета правдоподобия документа требуется перемножить вероятности того, что из «мешка» были «вытащены» те самые слова, которые встретились в документе.

Наивное предположение заключается в том, что из «мешка» «вытаскиваются» разные слова независимо друг от друга.

Мультиномиальная модель: пусть $V = \{w_t\}_{t=1}^{|V|}$ — словарь, тогда документ d_i — это вектор длины $|d_i|$, состоящий из слов, каждое из которых «вынуто» из словаря с вероятностью $p(w_t|c_j)$.

Правдоподобие принадлежности d_i классу c_j имеет вид [36]:

$$p\left(d_i\middle|c_j\right)=p\left(\left|d_i\right|\right)\left|d_i\right|!\prod_{t=1}^{|V|}\frac{1}{N_{it}!}p\left(w_t\middle|c_j\right)^{N_{it}}, \tag{12}$$

где N_{it} — количество вхождений w_t в d_i.

Для обучения такого классификатора требуется определить вероятности $p\left(w_t\middle|c_j\right)$. Далее рассматривается процесс обучения.

Обучение: пусть дан набор документов $D=\left\{d_i\right\}_{i=1}^{|D|}$, которые уже распределены по классам c_j (возможно, даже вероятностно распределены), дан словарь $V=\left\{w_t\right\}_{t=1}^{|V|}$, и значение вхождения N_{it} известно.

Тогда, в соответствии с (13), можно подсчитать оптимальные оценки вероятностей, что то или иное слово встречается в том или ином классе [36]:

$$p\left(w_t\middle|c_j\right)=\frac{1+\sum_{i=1}^{|D|}N_{it}p\left(c_j\middle|d_i\right)}{|V|+\sum_{s=1}^{|V|}\sum_{i=1}^{|D|}N_{is}p\left(c_j\middle|d_i\right)}. \tag{13}$$

Априорные вероятности классов рассчитываются согласно [36]:

$$p\left(c_j\right)=\frac{1}{|D|}\sum_{i=1}^{|D|}p\left(c_j\middle|d_i\right). \tag{14}$$

Тогда классификация будет происходить в соответствии с [36]:

$$\begin{aligned} c&=\arg\max_j p\left(c_j\right)p\left(d_j\middle|c_j\right)= \\ &=\arg\max_j\left(\frac{1}{|D|}\sum_{i=1}^{|D|}p\left(c_j\middle|d_i\right)\right)\prod_{t=1}^{|V|}\frac{1}{N_{it}!}p\left(w_t\middle|c_j\right)^{N_{it}}= \\ &=\arg\max_j\left(\log\left(\sum_{i=1}^{|D|}p\left(c_j\middle|d_i\right)\right)+\sum_{t=1}^{|V|}N_{it}\log p\left(w_t\middle|c_j\right)\right). \end{aligned} \tag{15}$$

Многомерная модель дает лучшую оценку предсказания на текстах с объемом, не превышающем 100 слов. Когда размер текстов составляет несколько тысяч слов, то лучшие результаты дает мультиномиальная модель. Таким образом, выбор MultinomialNB классификатора является обоснованным. Далее приведено его сравнение с линейным SGDClassifier классификатором.

8.2. Линейный классификатор с обучением стохастического градиентного спуска. Основная идея линейного классификатора заключается в том, что признаковое пространство может быть разделено гиперплоскостью на две полуплоскости, в каждой из которых прогнозируется одно из двух значений целевого класса.

Пусть вектор $\vec{x}$ представляет собой входные данные, а на выходе классификатора вычисляется показатель $\vec{y}$ по формуле [38]:

$$y = f\left(\vec{w} \cdot \vec{x}\right) = f\left(\sum_i w_i x_i\right), \tag{16}$$

где $\vec{x}$ — нормализованный вектор из частот слов в документе; $\vec{w}$ — действительный вектор весов той же размерности, что и признаковое пространство; f — функция преобразования скалярного произведения.

В ряде случаев задачи текстовой классификации, включающие в себя более одного класса, сводятся к нескольким задачам бинарной классификации [37]. В этом случае метки целевого класса $D_C \subset D$ обозначаются «1» (положительные примеры), а нецелевого «-1» (отрицательные примеры), а функция принадлежности $\mathrm{y}:\mathrm{D}\to\{\mathrm{l};-\mathrm{l}\}$ может быть представлена в виде [39]:

$$y = \begin{cases} +1, & x \in D_C \\ -1, & x \notin D_C \end{cases}. \tag{17}$$

Естественной интерпретацией для y в дискретном случае будет разделяющая гиперплоскость между различными классами. Для ускорения этого метода используется метод стохастического градиентного спуска: на каждой итерации спуск осуществляется с учетом одного случайно выбранного документа $d \in D$.

Значения весов вектора $\vec{w}$ определяются в ходе обучения на тестовых выборках [39].

Обучение: пусть $y^*: X \to Y$ — целевая зависимость, известная только на объектах обучающей выборки $X^l = \left(x_i, y_i\right)_{i=1}^l, y_i = y^*\left(x_i\right)$.

Требуется найти вектор весов $\boldsymbol{w}$, при котором алгоритм $a(x,w)$ аппроксимирует целевую зависимость $y^*(x_i)$.

Подобная задача сводится к поиску вектора $\boldsymbol{w}$ и доставляющего минимум функционалу [38]:

$$Q(w)=\sum_{i=1} L\left(a(x_i,w),y_i\right)\to \min_w, \tag{18}$$

где $L(a,y)$ — заданная функция потерь, характеризующая величину ошибки ответа $a(x,w)$ при правильном ответе y.

Применение для минимизации $Q(w)$ метод градиентного спуска. В этом методе выбирается некоторое начальное приближение для вектора весов w, затем запускается итерационный процесс, на каждом шаге которого вектор w изменяется в направлении и наиболее быстрого убывания функционала Q.

Это направление противоположно вектору градиента [38]:

$$\nabla Q(w)=\left(\frac{\partial Q(w)}{\partial w_j}\right)_{j=1}^{n};\ w:=w-\eta\nabla Q(w), \tag{19}$$

где $\eta>0$ — величина шага в направлении антиградиента, называемая также темпом обучения (learning rate).

Предполагая, что функция потерь L и функция активации f дифференцируемы, распишем градиент в виде [38]:

$$w:=w-\eta\sum_{i=1} L'_a\left(a(x_i,w),y_i\right)f'\left(\langle w,x_i\rangle\right)x_i. \tag{20}$$

Каждый прецедент (x_i,y_i) вносит аддитивный вклад в изменение вектора w, но вектор w изменяется только после перебора всех l объектов.

Для повышения скорости сходимости данного процесса прецеденты выбираются случайным образом (x_i,y_i), для каждого делается градиентный шаг и сразу обновляется вектор весов [38]:

$$w:=w-\eta L'_a\left(a(x_i,w),y_i\right)f'\left(\langle w,x_i\rangle\right)x_i. \tag{21}$$

Инициализация весов может производиться различными способами. Необходимо брать небольшие случайные значения, например: $w_j := random\left(-\frac{1}{2n}, \frac{1}{2n}\right)$.

Критерий останова такого алгоритма основан на приблизительной оценке функционала Q методом экспоненциальной скользящей средней. Точное значение потребовало бы вычисления l скалярных произведений $\langle w, x_i \rangle$, что довольно накладно. Когда градиентный метод подходит к окрестности минимума, оценка скользящего среднего стабилизируется и приближается к точному значению функционала Q.

Метод стохастического градиента хорошо приспособлен для динамического обучения, когда обучающие объекты поступают потоком, и надо быстро обновлять вектор весов при появлении каждого нового объекта.

Метод позволяет настраивать веса на избыточно больших выборках за счет того, что случайной подвыборки может оказаться достаточно для обучения. Допускаются различные стратегии обучения. В случае большой выборки или динамического потока можно вообще не сохранять обучающие объекты. В случае малой выборки можно повторно предъявлять для обучения одни и те же объекты, что способствует повышению качества классификации.

9. Обучение и тестирование тонового классификатора. Для обучения классификатора использовалась готовая выборка (http://study.mokoron.com/), состоящая приблизительно из 225 тысяч размеченных текстов, имеющих положительный и отрицательная окрас. В ходе тестирования качество классификации было максимизировано посредством перебора различных сочетаний классификаторов, методов векторизации, схем n-грамм и других параметров. Байесовский классификатор не нуждается в подборе параметров, а параметры линейной модели специально подбирались по сетке. В качестве сетки был использован словарь, в котором ключи представляли собой названия параметров, а значения состояли из наборов, которым требуется проверка.

Далее строилось декартово произведение на этих параметрах и по полученным точкам со всеми возможными наборами измерялось качество классификации. Этот процесс называется поиском по сетке.

В ходе тестирования были рассмотрены: вид функции потерь, вид регуляризации и множитель альфа перед регуляризацией.

В качестве стратегии кросс-валидации применялся метод ShuffleSplit из библиотеки scikit-learn, производилось 5 итераций и в тестовую выборку отсекалось 30 процентов данных. Результаты последних трех итераций представлены на рисунке 5.

```
N-gram Scheme: (1, 1)
Count Vectorizer
NB: 0.636833277424
Linear: 0.667829587387
Linear Parameters: {'alpha': 0.0001, 'penalty': 'elasticnet', 'loss': 'log'}

TF-IDF Vectorizer
NB: 0.583092921838
Linear: 0.698989898989
Linear Parameters: {'alpha': 1e-05, 'penalty': 'elasticnet', 'loss': 'log'}

N-gram Scheme: (1, 2)
Count Vectorizer
NB: 0.681784636028
Linear: 0.705333780611
Linear Parameters: {'alpha': 0.001, 'penalty': 'l2', 'loss': 'hinge'}

TF-IDF Vectorizer
NB: 0.608587722241
Linear: 0.717343173432
Linear Parameters: {'alpha': 1e-05, 'penalty': 'elasticnet', 'loss': 'log'}

N-gram Scheme: (1, 3)
Count Vectorizer
NB: 0.692787655149
Linear: 0.714793693391
Linear Parameters: {'alpha': 0.001, 'penalty': 'l2', 'loss': 'log'}

TF-IDF Vectorizer
NB: 0.633143240523
Linear: 0.719490103992
Linear Parameters: {'alpha': 0.0001, 'penalty': 'l2', 'loss': 'hinge'}

N-gram Scheme: (1, 4)
Count Vectorizer
NB: 0.69533713519
Linear: 0.719154646892
Linear Parameters: {'alpha': 0.001, 'penalty': 'l2', 'loss': 'hinge'}

TF-IDF Vectorizer
NB: 0.650788326065
Linear: 0.719490103992
Linear Parameters: {'alpha': 0.0001, 'penalty': 'l2', 'loss': 'hinge'}

N-gram Scheme: (1, 5)
Count Vectorizer
NB: 0.690640724589
Linear: 0.715062059712
Linear Parameters: {'alpha': 0.001, 'penalty': 'l2', 'loss': 'hinge'}

TF-IDF Vectorizer
NB: 0.660986246226
Linear: 0.718483730292
Linear Parameters: {'alpha': 0.0001, 'penalty': 'l2', 'loss': 'hinge'}
```

Рис. 5. Выбор модели классификатора

По результатам валидации лучшей оказалась линейная модель с триграммной схемой (1, 3), векторизатором TF-IDF и параметрами: penalty — l2 (функция штрафа L2-регуляризация, которая штрафует весовые значения добавлением суммы их квадратов к ошибке);

alpha — 0.000001(константа, которая умножает член регуляризации); loss — log (функция потерь в виде логистической регрессии). Ее результат составил ≈ 0.72.

Качество классификации превышает 70%, что говорит о правильном подборе релевантных обучающих выборок.

Теперь перейдем к практической реализации и тестовой эксплуатации системы для извлечения и анализа дорожно-транспортной информации с сайта http://autostrada.info/ru о состоянии транспортных сетей республики Крым и города Севастополь.

10. Практическая реализация. С использованием построенного тонового классификатора был проведен анализ отзывов пользователей портала http://autostrada.info/ru и осуществлена оценка качества транспортных сетей республики Крым и города Севастополь за период с 2011–2018 годы.

В результате анализа классификации 1130 отзывов были получены две выборки: 493 положительных отзыва и 637 отрицательных. Результаты анализа в таблице 3.

Таблица 3. Результаты автоматической классификации трасс по отзывам

Номер трассы	Наименование трассы	Участок трассы	Количество положительных отзывов	Количество отрицательных отзывов
			493	*637*
Н-05	Красноперекопск – Симферополь	-	25	8
Н-06	Симферополь – Севастополь	-	27	4
Н-19	Ялта – Севастополь	-	19	1
М-18	Харьков – Симферополь – Ялта	Новомосковск – Запорожье	49	199
М-18	Харьков – Симферополь – Ялта	Харьков – Новомосковск	39	111
М-18	Харьков – Симферополь – Ялта	Запорожье – Мелитополь	112	145
М-18	Харьков – Симферополь – Ялта	Симферополь – Ялта	14	2
М-18	Харьков – Симферополь – Ялта	Джанкой – Симферополь	13	6

Продолжение таблицы 3.

Номер трассы	Наименование трассы	Участок трассы	Количество положительных отзывов	Количество отрицательных отзывов
			493	*637*
М-18	Харьков – Симферополь – Ялта	Мелитополь – Джанкой	30	18
М-18	Харьков – Симферополь – Ялта	-	38	57
М-17	Херсон – Джанкой – Керчь	Херсон – Армянск	32	24
М-17	Херсон – Джанкой – Керчь	Джанкой – Феодосия	9	15
М-17	Херсон – Джанкой – Керчь	Красноперекопск – Феодосия	1	2
М-17	Херсон – Джанкой – Керчь	Феодосия – Керчь	5	5
М-17	Херсон – Джанкой – Керчь	Херсон – Джанкой	4	3
М-17	Херсон – Джанкой – Керчь	Красноперекопск – Джанкой	4	3
М-17	Херсон – Джанкой – Керчь	-	7	10
Р-23	Симферополь – Феодосия	-	18	6
Р-25	Симферополь – Евпатория	-	11	10
Р-27	Севастополь – Инкерман	-	10	-
Р-29	Алушта – Феодосия	Коктебель – Феодосия	5	-
Р-29	Алушта – Феодосия	-	11	2
Р-34	Ялта – Алушта	-	2	-
Р-35	Грушевка – Судак	-	5	4
Р-58	Окружная Севастополя	-	3	2

Наличие положительных и отрицательных отзывов пользователей в границах одного и того же участка трассы характеризуют оценки различных важных для водителей параметров.

Для наглядности результатов исследования приведем размеченную карту дорог Крыма и города Севастополь, соответствующую положительным (зеленый цвет) и отрицательным (желтый цвет) отзывам (см. рисунок 6).

Рис. 6. Размеченная в соответствии с отзывами карта дорог Крыма и города Севастополь

Примеры положительных и отрицательных отзывов, полученных при классификации, представлены в таблице 4.

Таблица 4. Примеры классифицированных отзывов на положительные и отрицательные (трассы Крыма и города Севастополя)

Номер трассы	Трасса	Положительные	Отрицательные
М-17	Херсон – Джанкой – Керчь (Феодосия – Керчь)	Отличная дорога, хотя на карте написано очень плохая).	Ехать только днем, местами очень плохо.
М-18	'Харьков – Симферополь – Ялта' (Запорожье – Мелитополь)	Дорога хорошая. Средняя скорость 80-100 км/ч на обычной легковушке. Ям нет, все залатаны. Ехать можно не напрягаясь. Учтите, что даже ночью трасса загружена, много грузовиков и людей едущий на отдых.	Крайне не рекомендую ехать! Состояние ужасное! Ямы, наплывы, латки, колейность, стиральная доска. Объезжайте!

Продолжение таблицы 4.

Номер трассы	Трасса	Положительные	Отрицательные
А-146	Краснодар – Верхнебаканский (Абинск – Верхнебаканский)	Дорога отличная, есть незначительные пробои на мостах.	Просто ужас, ехать затруднительно и опасно. Через каждые 3-5 км на обочине памятники, напоминающие о ДТП.
Н-05	Красноперекопск – Симферополь	Трасса хорошая.	Дорога до границы не очень. Наплывы. Ехать днём. После границы ещё хуже.
Н-06	Симферополь – Севастополь	Всегда была в хорошем состоянии. Никаких проблем.	Последнее время много аварий, и качество дороги играет не последнюю роль
Н-19	Ялта – Севастополь	Отличная дорога, ям нет.	Из за жары регулярно куча ДТП.
Р-23	Симферополь – Феодосия	Хорошая дорога.	Не очень.
Р-25	Симферополь – Евпатория	Дорога хорошая, ям нет, трещин мало.	Дороги нет. Она отсутствует вообще. Только по территории города Саки остался кусок который сделали. Все остальное просто яма. Любой Фома пусть сам проедет по ней 100 метров с закрытыми глазами на скорости 80 км и останется вообще без колес.
Р-27	Севастополь – Инкерман	Отличная двухполосная дорога, без намеков на ямы.	-

Продолжение таблицы 4.

Номер трассы	Трасса	Положительные	Отрицательные
Р-29	Алушта – Феодосия (Коктебель - Феодосия)	Горная дорога. Ехать комфортно. Ямы иногда на повороте, а так хорошее покрытие. Ехать не быстро, но машин не много. Очень красивые виды.	Между Судаком и Коктебелем ужасная горная дорога.
Р-34	Ялта – Алушта	Хорошая дорога.	-
Р-35	Грушевка – Судак	Нормальная трасса, с нормальным покрытием, есть небольшой участок через горы возле Судака, дальше ровная дорога.	Ужасная по состоянию дорога.

Каждый отзыв содержал следующие атрибуты: номер и наименование трассы, наименование уточненного участка трассы, дату и время регистрации отзыва и отзыв о качестве дороги. Примеры положительного и отрицательного отзывов представлены на рисунках 7 и 8.

```
841 =>
array (
  'title' => 'Н-19 Ялта – Севастополь',
  'subtitle' => NULL,
  'date' => '02.11.2015 20:18',
  'rate' => 'Дорога с идеальным или близким к идеальному покрытием',
  'description' => 'Дорожное покрытие хорошее, трещины есть, но мало. ',
```

Рис. 7. Пример положительного отзыва

```
217 =>
array (
  'title' => 'М-18 Харьков – Симферополь – Ялта',
  'subtitle' => 'Харьков – Новомосковск',
  'date' => '19.05.2018 22:12',
  'rate' => 'Дорожное покрытие полностью разрушено, постоянные ямы',
  'description' => 'Дорога в ужасном состоянии на отрезке Харьков – Днепр,
                    старайтесь строить маршрут через М-29 ',
```

Рис. 8. Пример отрицательного отзыва

По результатам анализа отзывов можно сделать вывод о том, что качество транспортных сетей Крыма с 2016 года постепенно

улучшается. На сентябрь 2018 около 60% дорожно-транспортной инфраструктуры республики Крым и города Севастополь все еще требуют ремонта. Оставшиеся 40 % находятся в удовлетворительном состоянии. Так, например, требуют ремонта и расширения дороги регионального значения: Р-29 Алушта – Феодосия, Р-23 Симферополь – Феодосия, Р-35 Грушевка – Судак, Р-25 Симферополь – Евпатория, Р-260 Таврида, Р-58 окружная дорога Севастополя и другие. Также необходимо увеличение снегоуборочной и ремонтной техники и повышение качества работы дорожных служб, так как в период с января 2017 – март 2017 негативные отзывы пользователей содержали информацию о заторах и ДТП по причине ухудшения качества дорожного полотна.

В дальнейшем планируется реализовать глубокую классификацию отзывов по тематическим группам, таким как: пробки, ДТП, ремонт, гололед и снежные заторы, ямы и выбоины, пропажа людей, штрафы и другое.

В рамках следующего этапа планируется также сравнить методы bag-of-words и tf-idf с методом векторного представления слов word2vec, который в ряде работ [41, 42] показал лучшие результаты. Также планируется рассмотреть методы тематической классификации текстов, такие как свёрточные нейронные сети (CNN) [43, 44], метод опорных векторов (SVM) [45, 46] и др.

11. Заключение. В ходе тестирования работы классификатора проведен анализ отзывов пользователей, относящихся к качеству транспортных сетей республики Крым и города Севастополь. Классификатор позволил разделить отзывы на положительные и отрицательные, и выявить проблемные участки транспортных сетей Крыма.

Подобные системы, основанные на анализе отзывов пользователей, позволят устанавливать причинно-следственные связи транспортно-логистической активности населения [40], формировать кодифицированные библиотеки шаблонов транспортного поведения [47], выполнять среднесрочное и долгосрочное прогнозирование процессов транспортной мобильности, формировать новые [48] и расширять существующие критерии и параметры управления транспортными потоками [49], выходя за рамки циклов светофорного регулирования и типовых схем прокладки маршрутов.

Использование систем оперативного анализа разнородных данных web-контента в составе интеллектуальных транспортных систем является эффективной и обоснованной технологией сегодняшнего времени.

Авторский коллектив благодарит администрацию сайта autostrada.info/ru за предоставленное разрешение на обработку и анализ текстовой информации.

Литература

1. *Seliverstov Y.A. et al.* Development of management principles of urban traffic under conditions of information uncertainty // Conference on Creativity in Intelligent Technologies and Data Science. 2017. pp. 399–418.
2. *Искандеров Ю.М.* Интеллектуальные транспортные системы: возможности и особенности применения // Мир дорог. 2013. № 68. С. 38–39.
3. *Искандеров Ю.М.* Использование инструментария семантических графов с оболочками при создании интеллектуальных транспортных систем // Международная научно-практическая конференция «Интеллектуальные системы на транспорте». 2011. С. 75–82.
4. *Искандеров Ю.М.* Построение модели интегрированной информационной системы транспортной логистики на основе мультиагентных технологий // Сборник статей Международной научно-практической конференции «Новая экономика и основные направления ее формирования». 2016. С. 62–69.
5. *Искандеров Ю.М., Ласкин М.Б., Лебедев И.С.* Особенности моделирования транспортно-технологических процессов в цепях поставок // Восьмая Всероссийская научно-практическая конференция «Имитационное моделирование. Теория и практика» (ИММОД-2017). 2017. С. 110–113.
6. *Свистунова А.С., Чумак А.С.* Интеллектуализация информационного обеспечения процесса перевозки негабаритных грузов // XVII Международная научно-практическая конференция «Логистика: современные тенденции развития». 2018. С. 76–79.
7. *Seliverstov Y.A. et al.* The method of selecting a preferred route based on subjective criteria // 2017 IEEE II International Conference on Control in Technical Systems (CTS). 2017. pp. 126–130.
8. *Seliverstov Ya.A. et al.* Intelligent systems preventing road traffic accidents in megalopolises in order to evaluate // 2017 20th IEEE International Conference on Soft Computing and Measurements (SCM). 2017. pp. 489–492.
9. *Малыгин И.Г., Комашинский В.И., Афонин П.Н.* Системный подход к построению когнитивных транспортных систем и сетей // Научно-аналитический журнал «Вестник Санкт-Петербургского университета Государственной противопожарной службы МЧС России». 2015. № 4. С. 68–73.
10. *Gal-Tzur A., Rechavi A., Beimel D., Freund S.* An improved methodology for extracting information required for transport related decisions from Q & A forums: A case study of TripAdvisor // Travel Behaviour and Society. 2018. vol. 10. pp. 1–9.
11. *Chaniotakis E., Antoniou C.* Use of Geotagged Social Media in Urban Settings: Empirical Evidence on its Potential from Twitter // 2015 IEEE 18th International Conference on Intelligent Transportation Systems. 2015. pp. 214–219.
12. *Kuflik T. et al.* Automating a framework to extract and analyse transport related social media content: The potential and the challenges // Transportation Research Part C: Emerging Technologies. 2017. vol. 77. pp. 275–291.
13. *Ali F. et al.* Fuzzy Ontology-based Sentiment Analysis of Transportation and City Feature Reviews for Safe Traveling // Transportation Research Part C: Emerging Technologies. 2017. vol. 77. pp. 33–48.
14. *Nanba H. et al.* Automatic compilation of travel information from automatically identified travel blogs // Proceedings of the ACL-IJCNLP 2009 Conference Short Papers. 2009. pp. 205–208.

15. *Zhang Z. et al.* Final Report. Mining Transportation Information from Social Media for Planned and Unplanned Events // Transportation Informatics, University Transportation Center. 2016. 68 p.
16. *Тихомиров И.А. и др.* Инструменты анализа научно-технологических заделов России // Труды Института системного анализа Российской академии наук. 2016. Т. 66. № 3. С. 98–104.
17. *Ананьева М.И.* О проблеме выявления экстремистской направленности в текстах // Вестник Новосибирского государственного университета. Серия: Информационные технологии. 2016. Т. 14. № 4. С. 5–13.
18. *Gu Y., Qian Z.S., Chen F.* From Twitter to detector: Real-time traffic incident detection using Social Media data // Transportation research part C: emerging technologies. 2016. vol. 67. pp. 321–342.
19. *Kuflik T. et al.* Automating a framework to extract and analyse transport related Social Media content: The potential and the challenges // Transportation Research Part C: Emerging Technologies. 2017. vol. 77. pp. 275–291.
20. *Serna A., Gerrikagoitia J.K., Bernabe U., Ruiz T.* A Method to Assess Sustainable Mobility for Sustainable Tourism: The Case of the Public Bike Systems // Information and Communication Technologies in Tourism. 2017. pp. 727–739.
21. *Serna A., Gasparovic S.* Transport analysis approach based on big data and text mining analysis from social media // Transportation Research Procedia. 2018. vol. 33. pp. 291–298.
22. *Блеканов И.С., Бондаренко Д.С.* Оценка эффективности методов поиска тематических сообществ в веб-пространстве // Научно-технические ведомости Санкт-Петербургского государственного политехнического университета. Информатика. Телекоммуникации. Управление. 2010. № 5(108). С. 18–24.
23. *Печников А.А., Сотенко Е.М.* Программы-краулеры для сбора данных о представительских сайтах заданной предметной области – аналитический обзор // Современные наукоемкие технологии. 2017. № 2. С. 58–62.
24. *Отраднов К.К., Раев В.К.* Экспериментальное исследование эффективности методик векторизации текстовых документов и алгоритмов их кластеризации // Вестник Рязанского государственного радиотехнического университета. 2018. № 64. С. 73–84.
25. *Михайлов Д.В., Козлов А.П., Емельянов Г.М.* Выделение знаний и языковых форм их выражения на множестве тематических текстов: подход на основе меры TF-IDF // Компьютерная оптика. 2015. Т. 39. № 3. С. 429–438.
26. *Ghaddar B., Naoum-Sawaya J.* High dimensional data classification and feature selection using support vector machines // European Journal of Operational Research. 2018. vol. 265. no. 3. pp. 993–1004.
27. *Rabiner L., Juang B.* Fundamentals of Speech Recognition // Prentice Hall. 1993. 507 p.
28. *Шелманов А.О. и др.* Семантико-синтаксический анализ текстов в задачах вопросно-ответного поиска и извлечения определений // Искусственный интеллект и принятие решений. 2016. № 4. С. 47–61.
29. *Кузнецов А.Н., Вышемирский Д.А.* Об одном подходе к решению задачи токенизации при анализе больших массивов пользовательских паролей // Безопасность информационных технологий. 2017. № 2. С. 50–60.
30. *Рубцова Ю.В.* Построение корпуса текстов для настройки тонового классификатора // Программные продукты и системы. 2015. № 1. С. 72–78.
31. *Мюллер А., Гвидо С.* Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными // Альфа-книга. 2017. 393 с.

32. Карякина А.А., Ботов Д.С. Анализ текстов для прогнозирования оттока клиентов Интернет-Провайдера // Челябинский физико-математический журнал. 2018. Т. 3. № 2. С. 227–236.
33. *Нугуманова А.Б., Бессмертный И.А., Пецина П., Байбурин Е.М.* Обогащение модели Bag of Words семантическими связями для повышения качества классификации текстов предметной области // Программные продукты и системы. 2016. № 2. С. 89–99.
34. *Кипяткова И.С.* Программно-алгоритмическое обеспечение создания синтаксическо-статистической модели русского языка по текстовому корпусу // Труды СПИИРАН. 2013. № 1(24). С. 332–348.
35. *Петровский М.И., Глазкова В.В.* Алгоритмы машинного обучения для задачи анализа и рубрикации электронных документов // Вычислительные методы и программирование: новые вычислительные технологии. 2007. Т. 8. № 2. С. 57–69.
36. *Сизов А.А., Николенко С.И.* Наивный Байесовский классификатор. DOCPLAYER. URL: https://docplayer.ru/45424867-Naivnyy-bayesovskiy-klassifikator.html. (дата обращения: 25.01.2019).
37. *Воронцов К.В.* Вероятностное тематическое моделированиею URL: http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf. (дата обращения: 25.01.2019).
38. *Воронцов К.В.* Лекции по линейным алгоритмам классификации. URL: http://www.machinelearning.ru/wiki/images/6/68/voron-ML-Lin.pdf. (дата обращения: 25.01.2019).
39. *Шаграев А.Г., Фальк В.Н.* Линейные классификаторы в задаче классификации текстов // Вестник Московского энергетического института. 2013. № 4. С. 204–208.
40. *Селиверстов Я.А., Селиверстов С.А.* Использование систем класса ГАТЛОСЭМИ для упреждения причин возникновения ДТП и неблагоприятных социальных исходов в «умном городе» // Научно-технические ведомости Санкт-Петербургского государственного политехнического университета. Информатика. Телекоммуникации. Управление. 2016. № 1(236). С. 65–81.
41. *Ботов Д.С., Кленин Ю.Д., Николаев И.Е.* Извлечение информации с использованием нейросетевых моделей языка на примере анализа вакансий в системах онлайн-рекрутмента // Вестник Югорского государственного университета. 2018. № 3(50). С. 37–48.
42. *Kim D., Seo D., Cho S., Kang P.* Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec // Information Sciences. 2019. vol. 477. pp. 15–29.
43. *Liao S. et al.* CNN for situations understanding based on sentiment analysis of twitter data // Procedia Computer Science. 2017. vol. 111. pp. 376–381.
44. *Lee G. et al.* Sentiment classification with word localization based on weakly supervised learning with a convolutional neural network // Knowledge-Based Systems. 2018. vol. 152. pp. 70–82.
45. *Deng Y., Sander A., Faulstich L., Denecke K.* Towards automatic encoding of medical procedures using convolutional neural networks and autoencoders // Artificial Intelligence in Medicine. 2019. vol. 93. pp. 29–42.
46. *Alimova I.S., Tutubalina E.V.* Entity-level classification of adverse drug reactions: a comparison of neural network models // Proceedings of the Institute for System Programming of the RAS. 2018. vol. 30. no. 5. pp. 177–196.
47. *Селиверстов Я.А., Селиверстов С.А.* Формальное построение цепочек транспортной активности городского населения // Научно-технические ведомости СПбГПУ. Информатика. Телекоммуникации. Управление. 2015. № 4(224). С. 91–104.

48. *Селиверстов С.А., Селиверстов Я.А.* Обзор показателей транспортной обеспеченности мегаполиса // Вестник гражданских инженеров. 2015. № 5(52). С. 237–247.
49. *Селиверстов С.А., Селиверстов Я.А.* О методе оценки эффективности организации процесса дорожного движения мегаполиса // Вестник транспорта Поволжья. 2015. № 2(50). С. 91–96.

Селиверстов Ярослав Александрович — канд. техн. наук, старший научный сотрудник, лаборатория интеллектуальных транспортных систем, Федеральное государственное бюджетное учреждение науки Институт проблем транспорта им. Н.С. Соломенко Российской академии наук (ИПТ РАН); магистрант, кафедра компьютерных систем и программных технологий института компьютерных наук и технологий, Санкт-Петербургский политехнический университет Петра Великого. Область научных интересов: интеллектуальные транспортные системы, машинное обучение, интеллектуальный анализ данных, компьютерное моделирование транспортных систем. Число научных публикаций — 69. seliverstov-yr@mail.ru; 12-я линия В.О., 13, 199178, Санкт-Петербург, Российская Федерация; р.т.: +7(812) 321-95-68.

Чигур Виктория Игоревна — студентка, факультет прикладной математики, Санкт-Петербургский государственный университет (СПбГУ). Область научных интересов: интеллектуальный анализ данных, машинное обучение, большие данные, компьютерная безопасность, системное программирование. Число научных публикаций — 3. v.chigur67@gmail.com; Университетская набережная, 7–9, 199034, Санкт-Петербург, Российская Федерация; р.т.: +7(812)328–20–00.

Сазанов Арсений Михайлович — аспирант, Институт компьютерных наук и технологий, Санкт-Петербургский политехнический университет Петра Великого. Область научных интересов: интеллектуальные системы, нейронные сети, искусственный интеллект. Число научных публикаций — 2. arseny.sazanov@gmail.com; Политехническая, 21, 195251, Санкт-Петербург, Российская Федерация; р.т.: +7(812)297-16-28.

Селиверстов Святослав Александрович — канд. техн. наук, старший научный сотрудник, лаборатория интеллектуальных транспортных систем, Федеральное государственное бюджетное учреждение науки Институт проблем транспорта им. Н.С. Соломенко Российской академии наук (ИПТ РАН); магистрант, Санкт-Петербургский политехнический университет Петра Великого. Область научных интересов: интеллектуальные транспортные системы, машинное обучение, интеллектуальный анализ данных, компьютерное моделирование транспортных систем. Число научных публикаций — 67. seliverstov_s_a@mail.ru; 12-я линия В.О., 13, 199178, Санкт-Петербург, Российская Федерация; р.т.: +7(812)321-95-68.

Свистунова Александра Сергеевна — программист, лаборатория информационных технологий на транспорте, Федеральное государственное бюджетное учреждение науки Санкт-Петербургского института информатики и автоматизации Российской академии наук (СПИИРАН). Область научных интересов: интеллектуальные транспортные системы, интеллектуальный анализ данных. Число научных публикаций — 2. svistunova_alexandra@bk.ru; 39, 14 линия В.О., 199178, Санкт-Петербург, Российская Федерация; р.т.: +7(812) 328-34-11.

Поддержка исследований. Работа выполнена при поддержке гранта РФФИ № 18-410-920016 в рамках инициативного проекта, проводимого совместно с Правительством Севастополя на тему: «Исследование социально-экономических и экологических процессов города Севастополя с ростом индустриального, транспортно-транзитного и туристского потенциалов».

DOI 10.15622/sp.18.2.354-389

Y.A SELIVERSTOV, V.I. CHIGUR, A.M. SAZANOV, S.A. SELIVERSTOV, A.S. SVISTUNOVA

SENTIMENT ANALYSIS OF "AUTOSTRADA.INFO/RU" USERS' COMMENTS

Seliverstov Y.A, Chigur V.I., Sazanov A.M., Seliverstov S.A., Svistunova A.S. **Sentiment Analysis of "AUTOSTRADA.INFO/RU" Users' Comments.**

Abstract. As a result of the analysis, it was revealed that social networks (Vkontakte, Facebook), thematic communities in microblogging networks (Twitter), resources for travelers (TripAdvisor), transport portals (Autostrada) are a source of up-to-date and operational information about the traffic situation, the quality of transport services and passenger satisfaction with the quality of levels of transport services. However, the existing transport monitoring systems do not contain software tools capable of collecting and analyzing traffic information located in the Internet environment. This paper discusses the task of building a system for automatically retrieving and classifying road traffic information from transport Internet portals and testing the developed system for analyzing the transport networks of Crimea and the city of Sevastopol. To solve this problem, an analysis of open source libraries for thematic data collection and analysis was carried out. An algorithm for extracting and analyzing texts has been developed. A crawler was developed using the Scrapy package in Python3, and user feedback from the portal http://autostrada.info/ru was collected on the state of the transport system of Crimea and the city of Sevastopol. For texts lemmatization and vector text transformation, the tf, idf, tf-idf methods and their implementation in the Scikit-Learn library were considered: CountVectorizer and TF-IDF Vectorizer. For word processing, Bag-of-Words and n-gram methods were considered. During the development of the classifier model, the naive Bayes algorithm (MultinomialNB) and the linear classifier model with optimization of the stochastic gradient descent (SGDClassifier) were used. As a training sample, a corpus of 225,000 labeled texts from the Twitter resource was used. The classifier was trained, during which the cross-validation strategy and the ShuffleSplit method were used. Testing and comparison of the results of the pitch classification were carried out. According to the results of validation, the linear model with the n-gram scheme [1, 3] and the vectorizer TF-IDF turned out to be the best. During the approbation of the developed system, the collection and analysis of reviews related to the quality of transport networks of the Republic of Crimea and the city of Sevastopol were conducted. Conclusions are drawn and prospects for further functional development of the developed tools are defined.

Keywords: Automatic Text Analysis, Crowlers, Classification of Texts, Intelligent Transport Systems, Machine Training, TF-IDF, Naive Bayes Algorithm, Linear Classifier, Sentiment Analysis.

Seliverstov Yaroslav Aleksandrovich — Ph.D., Senior Researcher, Laboratory of Intelligent Transport Systems, Solomenko Institute of Transport Problems of the Russian academy of sciences; Master Student, Department of Computer Systems and Software Technologies of Institute of Computer Science and Technology, Peter the Great St.Petersburg Polytechnic University. Research interests: intelligent transport systems, machine learning, data mining, computer simulation of transport systems. The number of publications — 69. seliverstov-yr@mail.ru; 13, 12-th Line V.O., 199178, St. Petersburg, Russian Federation; office phone: +7(812) 321-95-68.

Chigur Viktoriya Igorevna — bachelor's student, Faculty of Applied Mathematics, St. Petersburg State University. Research interests: data mining, machine learning, big data, computer security, system programming. The number of publications — 3. v.chigur67@gmail.com; 7–9, University Embankment, 199034, , Russian Federation; office phone: +7(812)328–20–00.

Sazanov Arseniy Mikhailovich — Ph.D. Student, Institute of Computer Science and Technology, Peter the Great St.Petersburg Polytechnic University. Research interests: intellectual systems, neural networks, artificial intelligence. The number of publications — 2. arseny.sazanov@gmail.com; 21, Polytechnicheskaya, 195251, St. Petersburg, Russian Federation; office phone: +7(812)297-16-28.

Seliverstov Svyatoslav Aleksandrovich — Ph.D., Laboratory of Intelligent Transport Systems, Laboratory of Intelligent Transport Systems, Solomenko Institute of Transport Problems of the Russian academy of sciences; Master Student, Peter the Great St.Petersburg Polytechnic University. Research interests: intelligent transport systems, machine learning, data mining, computer simulation of transport systems. The number of publications — 67. seliverstov_s_a@mail.ru; 13, 12-th Line V.O., 199178, St. Petersburg, Russian Federation; office phone: +7(812)321-95-68.

Svistunova Aliaksandra Sergeevna — programmer, Transport Information Technologies Laboratory, St. Petersburg Institute of Informatics and Automation of the Russian Academy of Sciences (SPIIRAS). Research interests: intelligent transport systems, data mining. The number of publications — 2. svistunova_alexandra@bk.ru; 39, 14-th Line V.O., 199178, St. Petersburg, Russian Federation; office phone: +7(812) 328-34-11.

Acknowledgements. The scientific research was supported by the Russian Foundation for Basic Research within the framework of the project № 18-410-920016 p_a "Research of socio-economic and ecological processes of Sevastopol with the growth of industrial, traffic, transit and tourist potentials".

References

1. Seliverstov Y.A. et al. Development of management principles of urban traffic under conditions of information uncertainty. Conference on Creativity in Intelligent Technologies and Data Science. 2017. pp. 399–418.
2. Iskanderov Yu.M. [Intellectual transport systems: possibilities and features of application]. *Mir dorog – World of roads.* 2013. vol. 68. pp. 38–39. (In Russ.).
3. Iskanderov Yu.M. [Using the toolkit of semantic graphs with shells when creating intelligent transport systems]. *Mezhdunarodnaya nauchno-prakticheskaya konferenciya "Intellektual'nye sistemy na transporte"* [Proceedings of International Scientific-Practical Conference on Intelligent Transport Systems]. 2011. pp. 75–82. (In Russ.).
4. Iskanderov Yu.M. [Building a model of an integrated transport logistics information system based on multi-agent technologies]. *Sbornik statej Mezhdunarodnoj nauchno-prakticheskoj konferencii "Novaya ehkonomika i osnovnye napravleniya ee formirovaniya"* [Proceedings of International Scientific-Practical Conference on New Economy and the Main Directions of its Formation]. 2016. pp. 62–69. (In Russ.).
5. Iskanderov Yu.M., Laskin M.B., Lebedev I.S. [Features of modeling of transport and technological processes in supply chains]. *Vos'maya Vserossijskaya nauchno-prakticheskaya konferenciya "Imitacionnoe modelirovanie. Teoriya i praktika"* [8th All-Russian Scientific-Practical Conference "Simulation. Theory and Practice" (IMMOD-2017)]. 2017. pp. 110–113. (In Russ.).

6. Svistunova A.S., Chumak A.S. [Intellectualization of information support of the process of transportation of oversized cargo]. *XVII Mezhdunarodnaya nauchno-prakticheskaya konferenciya "Logistika: sovremennye tendencii razvitiya"* [XVII International Scientific-Practical Conference "Logistics: Modern Development Trends"]. 2018. pp. 76–79. (In Russ.).
7. Seliverstov Y.A. et al. The method of selecting a preferred route based on subjective criteria. 2017 IEEE II International Conference on Control in Technical Systems (CTS). 2017. pp. 126–130.
8. Seliverstov Ya.A. et al. Intelligent systems preventing road traffic accidents in megalopolises in order to evaluate. 2017 20th IEEE International Conference on Soft Computing and Measurements (SCM). 2017. pp. 489–492.
9. Malygin I.G., Komashinskiy V.I., Afonin P.N. [System approach to the construction of cognitive transport systems and networks]. *Nauchno-analiticheskij zhurnal "Vestnik Sankt-Peterburgskogo universiteta Gosudarstvennoj protivopozharnoj sluzhby MCHS Rossii" – Scientific and analytical journal Bulletin of the St. Petersburg University of the State Fire Service EMERCOM of Russia.* 2015. vol. 4. pp. 68–73. (In Russ.).
10. Gal-Tzur A., Rechavi A., Beimel D., Freund S. An improved methodology for extracting information required for transport related decisions from Q & A forums: A case study of TripAdvisor. *Travel Behaviour and Society.* 2018. vol. 10. pp. 1–9.
11. *Chaniotakis E., Antoniou C.* Use of Geotagged Social Media in Urban Settings: Empirical Evidence on its Potential from Twitter. 2015 IEEE 18th International Conference on Intelligent Transportation Systems. 2015. pp. 214-219.
12. Kuflik T. et al. Automating a framework to extract and analyse transport related social media content: The potential and the challenges. *Transportation Research Part C: Emerging Technologies.* 2017. vol. 77. pp. 275–291.
13. Ali F. et al. Fuzzy Ontology-based Sentiment Analysis of Transportation and City Feature Reviews for Safe Traveling. *Transportation Research Part C: Emerging Technologies.* 2017. vol. 77. pp. 33–48.
14. Nanba H. et al. Automatic compilation of travel information from automatically identified travel blogs. Proceedings of the ACL-IJCNLP 2009 Conference Short Papers. 2009. pp. 205–208.
15. Zhang Z. et al. Final Report. Mining Transportation Information from Social Media for Planned and Unplanned Events. Transportation Informatics, University Transportation Center. 2016. 68 p.
16. Tikhomirov I.A. et al. [Analysis tools for scientific and technological groundwork in Russia]. *Trudy Instituta sistemnogo analiza Rossijskoj akademii nauk – Proceedings of the Institute for System Analysis of the Russian Academy of Sciences*. 2016. vol. 66. no. 3. pp. 98–104. (In Russ.).
17. Anan'yeva M.I. [On the problem of identifying extremist orientation in the texts]. *Vestnik Novosibirskogo gosudarstvennogo universiteta. Seriya: Informacionnye tekhnologii – Bulletin of the Novosibirsk State University. Series: Information Technology.* 2016. vol. 14. no. 4. pp. 5–13. (In Russ.).
18. Gu Y., Qian Z.S., Chen F. From Twitter to detector: Real-time traffic incident detection using Social Media data. *Transportation research part C: emerging technologies*. 2016. vol. 67. pp. 321–342.
19. Kuflik T. et al. Automating a framework to extract and analyse transport related Social Media content: The potential and the challenges. *Transportation Research Part C: Emerging Technologies.* 2017. vol. 77. pp. 275–291.
20. Serna A., Gerrikagoitia J.K., Bernabe U., Ruiz T. A Method to Assess Sustainable Mobility for Sustainable Tourism: The Case of the Public Bike Systems. Information and Communication Technologies in Tourism. 2017. pp. 727–739.

21. Serna A., Gasparovic S. Transport analysis approach based on big data and text mining analysis from social media. *Transportation Research Procedia.* 2018. vol. 33. pp. 291–298.
22. Blekanov I.S., Bondarenko D.S. [Evaluation of the effectiveness of search methods for thematic communities in the web space]. *Nauchno-tekhnicheskie vedomosti Sankt-Peterburgskogo gosudarstvennogo politekhnicheskogo universiteta. Informatika. Telekommunikacii. Upravlenie – Scientific and Technical Gazette of St. Petersburg State Polytechnic University. Computer science. Telecommunications. Control.* 2010. vol. 5(108). pp. 18–24. (In Russ.).
23. Pechnikov A.A., Sotenko Ye.M. [Crawler programs for collecting data on representative sites of a given subject area – analytical review]. *Sovremennye naukoemkie tekhnologii – Modern high technologies.* 2017. vol. 2. pp. 58–62. (In Russ.).
24. Otradnov K.K., Rayev V.K. [Experimental study of the effectiveness of methods for vectorization of text documents and algorithms for their clustering]. *Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo universiteta – Vestnik of Ryazan State Radio Engineering University.* 2018. vol. 64. pp. 73–84. (In Russ.).
25. Mikhaylov D.V., Kozlov A.P., Yemel'yanov G.M. [The selection of knowledge and linguistic forms of their expression on the set of thematic texts: an approach based on the measure TF-IDF]. *Komp'yuternaya optika – Computer Oprics.* 2015. vol. 39. no. 3. pp. 429–438. (In Russ.).
26. Ghaddar B., Naoum-Sawaya J. High dimensional data classification and feature selection using support vector machines. *European Journal of Operational Research.* 2018. vol. 265. no. 3. pp. 993–1004.
27. Rabiner L., Juang B. Fundamentals of Speech Recognition.Prentice Hall. 1993. 507 p.
28. Shelmanov A.O. et al. [Semantic-syntactic analysis of texts in the tasks of question-answer search and extraction of definitions]. *Iskusstvennyj intellekt i prinyatie reshenij – Artificial Intelligence and Decision Making.* 2016. vol. 4. pp. 47–61. (In Russ.).
29. Kuznetsov A.N., Vyshemirskiy D.A. [On one approach to solving the tokenization problem when analyzing large arrays of user passwords]. *Bezopasnost' informacionnyh tekhnologij – Information Technology Security.* 2017. vol. 2. pp. 50–60. (In Russ.).
30. Rubtsova YU.V. [Building a corpus of texts to adjust the tone classifier]. *Programmnye produkty i sistemy – Software products and systems.* 2015. vol. 1. pp. 72–78. (In Russ.).
31. Myuller A., Gvido S. *Vvedeniye v mashinnoye obucheniye s pomoshch'yu Python. Rukovodstvo dlya spetsialistov po rabote s dannymi* [Introduction to machine learning using Python. A Guide for Data Specialists]. Al'fa-kniga 2017. 393 p. (In Russ.).
32. Karyakina A.A., Botov D.S. [Analysis of texts for forecasting the outflow of clients of the Internet provider]. *Chelyabinskij fiziko-matematicheskij zhurnal – Chelyabinsk Physics and Mathematics Journal.* 2018. vol. 3. no. 2. pp. 227–236. (In Russ.).
33. Nugumanova A.B., Bessmertnyy I.A., Petsina P., Bayburin Ye.M. [Enrichment of the Bag of Words model with semantic links to improve the quality of domain text classification]. *Programmnye produkty i sistemy – Software products and systems.* 2016. vol. 2. pp. 89–99. (In Russ.).
34. Kipyatkova I.S. [Software and algorithmic support for the creation of a syntactic-statistical model of the Russian language by text corpus]. *Trudy SPIIRAN – Proceedings of SPIIRAS*. 2013. vol. 1(24). pp. 332–348. (In Russ.).
35. Petrovskiy M.I., Glazkova V.V. [Machine learning algorithms for the task of analyzing and categorizing electronic documents]. *Vychislitel'nye metody i programmirovanie: novye vychislitel'nye tekhnologii – Numerical methods and programming: new computing technologies.* 2007. Issue 8. vol. 2. pp. 57–69. (In Russ.).

36. Sizov A.A., Nikolenko S.I. Naivnyy Bayyesovskiy klassifikator [Naive Bayes Classifier]. Available at: https://docplayer.ru/45424867-Naivnyy-bayesovskiy-klassifikator.html (accessed: 25.01.2019). (In Russ.).
37. K.V. Vorontsov. Veroyatnostnoye tematicheskoye modelirovaniye [Probabilistic thematic modeling]. Available at: http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf (accessed: 25.01.2019). (In Russ.).
38. Vorontsov K.V. Lektsii po lineynym algoritmam klassifikatsii [Lectures on linear classification algorithms]. Available at: http://www.machinelearning.ru/wiki/images/6/68/voron-ML-Lin.pdf (accessed: 25.01.2019). (In Russ.).
39. Shagrayev A.G., Fal'k V.N. [Linear classifiers in the task of classifying texts]. *Vestnik Moskovskogo ehnergeticheskogo instituta – Bulletin of the Moscow Power Engineering Institute.* 2013. vol. 4. pp. 204–208. (In Russ.).
40. Seliverstov Ya.A., Seliverstov S.A. [The use of GATLOSEMI class systems to pre-empt the causes of accidents and adverse social outcomes in the "smart city"]. *Nauchno-tekhnicheskie vedomosti Sankt-Peterburgskogo gosudarstvennogo politekhnicheskogo universiteta. Informatika. Telekommunikacii. Upravlenie – Scientific and technical statements of the St. Petersburg State Polytechnic University. Computer science. Telecommunications. Control.* 2016. № 1(236). pp. 65–81. (In Russ.).
41. Botov D.S., Klenin Yu.D., Nikolayev I.Ye. [Extraction of information using neural network language models on the example of the analysis of vacancies in online recruitment systems]. *Vestnik Yugorskogo gosudarstvennogo universiteta – Bulletin of the Yugra State University.* 2018. № 3(50). pp. 37–48. (In Russ.).
42. Kim D., Seo D., Cho S., Kang P. Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec. *Information Sciences.* 2019. vol. 477. pp. 15–29.
43. Liao S. et al. CNN for situations understanding based on sentiment analysis of twitter data. *Procedia Computer Science*. 2017. vol. 111. pp. 376–381.
44. Lee G. et al. Sentiment classification with word localization based on weakly supervised learning with a convolutional neural network. *Knowledge-Based Systems*. 2018. vol. 152. pp. 70–82.
45. Deng Y., Sander A., Faulstich L., Denecke K. Towards automatic encoding of medical procedures using convolutional neural networks and autoencoders. *Artificial Intelligence in Medicine*. 2019. vol. 93. pp. 29–42.
46. Alimova I.S., Tutubalina E.V. Entity-level classification of adverse drug reactions: a comparison of neural network models. *Proceedings of the Institute for System Programming of the RAS.* 2018. vol. 30. no. 5. pp. 177–196.
47. Seliverstov Ya.A., Seliverstov S.A. [Formal construction of transport activity chains of the urban population]. *Nauchno-tekhnicheskie vedomosti SPbGPU. Informatika. Telekommunikacii. Upravlenie – St. Petersburg State Polytechnic University Journal. Computer science. Telecommunications. Control*. 2015. vol. 4(224). pp. 91–104. (In Russ.).
48. Seliverstov S.A., Seliverstov Ya.A. [Review of metropolitan transport security indicators]. *Vestnik grazhdanskikh inzhenerov – Bulletin of Civil Engineers.* 2015. vol. 5(52). pp. 237–247. (In Russ.).
49. Seliverstov S.A., Seliverstov Ya.A. [On the method of evaluating the effectiveness of the organization of the traffic process of a megacity]. *Vestnik transporta Povolzh'ya – Bulletin of transport of the Volga region*. 2015. vol. 2(50). pp. 91–96. (In Russ.).