DOI: 10.18721/JCSTCS.12301
УДК 004.8, 004.62, 007.5 , 51-74, 510.67, 656

ОЦЕНКА СОСТОЯНИЯ ТРАНСПОРТНЫХ МАГИСТРАЛЕЙ СЕВЕРО-ЗАПАДНОГО ФЕДЕРАЛЬНОГО ОКРУГА С ИСПОЛЬЗОВАНИЕМ АНАЛИЗА ТОНАЛЬНОСТИ ОТЗЫВОВ ПОЛЬЗОВАТЕЛЕЙ СЕТИ ИНТЕРНЕТ

Я.А. Селиверстов[1], К.В. Никитин[2], Н.В. Шаталова[1], А.А. Киселев[3]

[1] Институт проблем транспорта имени Н.С. Соломенко РАН,
Санкт-Петербург, Российская Федерация;
[2] Санкт-Петербургский политехнический университет Петра Великого,
Санкт-Петербург, Российская Федерация;
[3] Санкт-Петербургская государственная художественно-промышленная
академия имени А.Л. Штиглица,
Санкт-Петербург, Российская Федерация

В результате анализа выявлено, что социальные сети, тематические сообщества, транспортные порталы являются источником актуальной информации о дорожно-транспортной обстановке. В статье рассмотрена задача анализа состояния транспортных магистралей Северо-Западного федерального округа по отзывам, размещенным в web-пространстве. Для решения этой задачи разработана система автоматической классификации отзывов на основе тонового классификатора. Проведен анализ библиотек с открытым исходным кодом для тематического сбора и анализа данных. Осуществлена разработка краулера с использованием фреймворка Scrapy на языке Python3 и собраны отзывы с сайта http://autostrada.info/ru. Рассмотрены методы векторизации и лемматизации текстов и их реализация в библиотеке Scikit-Learn: Bag-of-Words, N-gram, CountVectorizer и TF-IDF Vectorizer. Для классификации применялся наивный байесовский алгоритм и модель линейного классификатора с оптимизацией стохастического градиентного спуска. В качестве обучающей выборки использована база размеченных отзывов с ресурса Twitter. Проведено обучение классификатора, в ходе которого использована стратегия кросс-валидации и метод ShuffleSplit. Проведено тестирование и сравнение результатов тоновой классификации на разных классификаторах. По результатам валидации лучшей оказалась линейная модель со схемой N-gram и векторизатором TF-IDF. В ходе апробации разработанной системы проведен сбор и анализ отзывов, относящихся к качеству транспортных сетей Северо-Западного федерального округа. На основе результатов произведена цветовая разметка дорог, отражающая наглядность результатов исследования. Сделаны выводы и определены перспективы дальнейшего развития данного исследования.

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

Ссылка при цитировании: Селиверстов Я.А., Никитин К.В., Шаталова Н.В., Киселев А.А. Оценка состояния транспортных магистралей Северо-Западного федерального округа с использованием анализа тональности отзывов пользователей сети Интернет // Научно-технические ведомости СПбГПУ. Информатика. Телекоммуникации. Управление. 2019. Т. 12. № 3. С. 7–24. DOI: 10.18721/JCSTCS.12301

ROAD PAVEMENT ASSESSMENT OF THE NORTH-WEST FEDERAL DISTRICT USING SENTIMENT ANALYSIS OF THE INTERNET USER REVIEWS

Ya.A. Seliverstov[1], *K.V. Nikitin*[2], *N.V. Shatalova*[1], *A.A. Kiselev*[3]

[1] Solomenko Institute of Transport Problems of the RAS, St. Petersburg, Russian Federation;
[2] Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russian Federation;
[3] Saint Petersburg Stieglitz State Academy of Art and Design, St. Petersburg, Russian Federation

As a result of the analysis, it was revealed that social networks, thematic communities, transport portals are a source of actual information about the traffic situation. The article deals with the task of analyzing the road pavement assessment of the North-West Federal District from reviews posted in the web. To solve this problem, a system for automatic classification of reviews based on the sentiment classifier has been developed. The crawler was developed using the Scrapy framework in Python3 and collected reviews from the site http://autostrada.info/ru. The methods of vectorization and lemmatization of texts and their implementation in the Scikit-Learn library are considered: Bag-of-Words, N-gram, CountVectorizer and TF-IDF Vectorizer. For the classification, a naive Bayes algorithm and a linear classifier model with optimization of stochastic gradient descent were used. As a training sample, a base of marked reviews from the Twitter resource was used. The classifier was trained, during which the cross-validation strategy and the ShuffleSplit method were used. According to the results of validation, the linear model with the N-gram scheme and the TF-IDF Vectorizer turned out to be the best. During the approbation of the developed system, the collection and analysis of feedback related to the quality of transport networks in the North-West Federal District was conducted. Based on the results, a color marking of the roads was produced, reflecting the visibility of the research results. Conclusions and prospects for the further development of this study are given.

Keywords: automatic text analysis, crowlers, texts classification, intelligent transport systems, machine learning, TF-IDF, N-gram, naive Bayes algorithm, linear classifier, sentiment analysis.

Citation: Seliverstov Ya.A., Nikitin K.V., Shatalova N.V., Kiselev A.A. Road pavement assessment of the North-West Federal District using sentiment analysis of the Internet user reviews. St. Petersburg State Polytechnical University Journal. Computer Science. Telecommunications and Control Systems, 2019, Vol. 12, No. 3, Pp. 7–24. DOI: 10.18721/JCSTCS.12301

Введение

Традиционные методы, применяемые для обнаружения дорожных событий [1, 2], в основном сосредоточены на измерении скорости, плотности и интенсивности движения с использованием различных датчиков и детекторов, которые обычно устанавливаются в фиксированных местах вдоль дорог [3]. Такие системы, как правило, имеют высокую стоимость, и поэтому их размещают лишь на особо загруженных участках городских магистралей. Также они требуют регулярного технического обслуживания и сопутствующей инфраструктуры.

Вместе с тем в последнее время актуальным источником разнородной информации, относящейся к сфере транспорта и логистики, является web-пространство [4]. Обычно такие данные находятся на тема-

тических или специализированных интернет-ресурсах, например: транспортные интернет-порталы (http://autostrada.info/ru, https://rosyama.ru/); интернет-сообщества грузоперевозчиков (https://www.worldoftrucks.com/en/, ingruz.ru); группы в социальных сетях (Вконтакте, Facebook) или в сетях микроблогинга (Twitter), а также чаты и форумы [5].

Информация на транспортных web-порталах и в 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 на ней 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 ГИБДД [6].

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 [7].

Таким образом, использование систем извлечения и анализа дорожно-транспортной информации из web-пространства в качестве систем транспортного мониторинга [8, 9] открывает новые каналы поступления транспортной информации, способной повысить информированность участников дорожного движения о состоянии транспортных сетей и условий дорожного движения.

Анализ предметной области. Проанализируем актуальные работы, в которых рассмотрены методы анализа текстов, относящихся к транспортной сфере. В [10] изучено использование данных социальных сетей для обеспечения быстрого и более точного обнаружения и уменьшения пробок на дорогах. В [11] авторы исследовали возможность использования данных форумов с интернет-порталов для обнаружения дорожно-транспортных происшествий (ДТП) и сбора дополнительной информации об инцидентах. В [12] представлен обзор различных категорий социальных сетей, характеристик их контента и того, как эти характеристики отражаются в сообщениях, связанных с транспортом. В работе [13] проведено исследование полезности использования социальных сетей для пользователей и поставщиков транспортных услуг и потенциальной ценности социальных сетей для разработки политики в области распространения информации для населения. В [14] описана система анализа данных социальных сетей микроблогинга Twitter, которая используется для выявления транспортных заторов в реальном времени для дорог Австралии.

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.

Постановка задачи. Цель настоящего исследования – оценка состояния транспортных магистралей Северо-Западного федерального округа с использованием анализа тональности отзывов пользователей сети Интернет.

Предполагается выполнить следующий перечень работ: 1) определить тематические web-ресурсы, предоставляющие актуальную информацию о дорогах Северо-Западного федерального округа; 2) разработать схему алгоритма для извлечения и анализа текстов; 3) программно реализовать алгоритм для сбора текстов по дорожно-транспортной проблематике; 4) произвести тестирование разработанной программы и осуществить сбор текстов с выбранного интернет-ресурса; 5) сформировать корпуса текстов для последующего обучения классификатора; 6) осуществить разработ-

ку классификатора тональности; 7) провести обучение классификатора и оценить качество классификации; 8) произвести оценку состояния транспортных магистралей Северо-Западного федерального округа с использованием разработанной системы анализа текстов.

Этапы работы

Этап 1. Выполнение анализа интернет-ресурсов, содержащих актуальную информацию пользователей о состоянии дорог Северо-Западного федерального округа. Результаты анализа транспортных интернет-ресурсов представлены в табл. 1.

Таблица 1

Транспортные интернет-ресурсы

Table 1

Transport Internet Resources

Наименование	Примечание
Тематический транспортный web-ресурс «Порталы»	
«Автострада» – проект об актуальном состоянии дорожного покрытия трасс России, Украины и Беларуси	http://autostrada.info/ru
«Доринфо» – дорожные новости, репортажи, аналитика, отзывы	http://dorinfo.ru/
«Центр организации дорожного движения Правительства Москвы» – сбор данных о дорожном движении, включая параметры транспортных и пассажирских потоков, дорожных условий, действующей организации дорожного движения, параметры экологического ущерба от дорожного движения, статистику ДТП, данные по парковкам и местам временного отстоя транспорта	http://www.gucodd.ru/
«РосЯма» – проект об актуальном состоянии дорожного покрытия улично-дорожных сетей городов России	https://rosyama.ru/
«Дорожная инспекция ОНФ/Карта убитых дорог» – проект о состоянии УДС городов России	http://dorogi-onf.ru/
Портал «РосАвтодора» раздел «Автомобилистам» посвящен ситуации на дорогах	http://rosavtodor.ru/avtomobilistam
Форум «АвтоТрансИнфо» – информация и отзывы водителей грузового транспорта	https://forums.ati.su/Forum/Default.aspx
Тематический транспортный web-ресурс Твиттер	
«Dorinfo» – актуальные дорожные новости и твиты пользователей	https://twitter.com/dorinfo
«Московский транспорт» – оперативная информация о дорожной ситуации на улицах Москвы, сбоях и изменениях в работе городского транспорта, перекрытиях дорог	https://twitter.com/DtRoad
«Московское метро» – официальный твиттер-аккаунт Московского метрополитена по оперативному информированию о работе метро	https://twitter.com/nwroads
Тематический транспортный web-ресурс – группы в ВК	
«СколькоДал.РФ» – поборы и взятки на дорогах РФ: отзывы водителей, интересные сюжеты, горячие новости и скандальные расследования	https://vk.com/skolko_dal
«Автостоп Community» – путешествия автостопом и все вопросы, связанные с ним	https://vk.com/ru_autostop
«Автостоп Онлайн» – новое, бесплатное приложение для поиска водителей и пассажиров, без диспетчеров	https://vk.com/autostop.online

Таблица 2

Анализ web-краулеров

Table 2

Analysis of web crawlers

Название	Техническое описание	Примечание
Heritrix	Гибкий, расширяемый, надежный и масштабируемый фреймворк, написанный на Java, способный получать, архивировать и анализировать тексты. Heretrix работает в распределенной среде с помощью хеширования URL хостов в поисковых машинах	[17, 18]
Nutch	Представляет собой инкрементный, параллельный, распределенный, кросс-платформенный модульный фреймворк для построения поисковых систем, написанный на java. Поддерживает граф связей узлов, различные фильтры и нормализацию URL	
Scrapy	Расширяемый, сфокусированный, параллельный, кросс-платформенный и гибкий фреймворк-библиотека для Python. Легко устанавливается, поддерживает выгрузку данных в форматах JSON, XML, CSV. Широко используется для веб-скрайпинга, не имеет встроенных функций для работы в распределенной среде	

В качестве интернет-ресурса для дальнейшего исследования выберем портал http://autostrada.info/ru, так как на нем содержатся актуальные и постоянно обновляемые отзывы о состоянии дорог Северо-Западного федерального округа.

Этап 2. Выполнение анализа фреймворков, предназначенных для парсинга[1] и краулинга[2]. На сегодняшний момент уже существует широкий спектр известных библиотек [15, 16], которые позволяют не писать с нуля поисковые роботы. На основе анализа технических описаний фреймворков для сбора текстов, представленного в табл. 2, целесообразно выбрать фреймворк Scrapy.

Scrapy используется для получения данных с различных интернет-ресурсов, является популярным и производительным фреймворком, написанным на Python.

Этап 3. Разработка алгоритма работы системы для извлечения и анализа текстов. Алгоритм состоит из процедур, представленных в табл. 3, а схема алгоритма представлена на рис. 1.

Этап 4. Построение краулер-модуля. Краулер-модуль выполняет процедуры 1–4 алгоритма: 1) формирует очередь ссылок; 2) добавляет список источников в очередь обхода; 3) сканирует страницу из очереди; 4) скачивает интересующий его веб-документ в базу данных.

В результате работы краулер-модуль формирует базу данных с отзывами пользователей.

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.

Этап 5. Векторизация [19] и лексический анализ отзывов. Все слова необходимо перевести в числовой вектор признаков с помощью одного из методов TF, IDF и TF-IDF [20]. Для этой процедуры используются векторизаторы. Векторизатор строит словарь индексов признаков. Мы будем использовать два векторизатора: CountVectorizer и TF-IDF Vectorizer [21]. Оба метода используют модель Bag of Words [22] и модель N-gram [23].

[1] Парсинг (parsing – разбор) – автоматизированный сбор неструктурированной информации, ее преобразование и выдача в структурированном виде.

[2] Краулинг (crawling – сканирование) – процесс сбора данных в Интернет, состоящий из навигации на веб-страницах, анализа их ссылок и содержимого.

Таблица 3

Общий вид алгоритма для извлечения и анализа тематических текстов

Table 3

General view of the algorithm for extracting and analyzing thematic texts

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

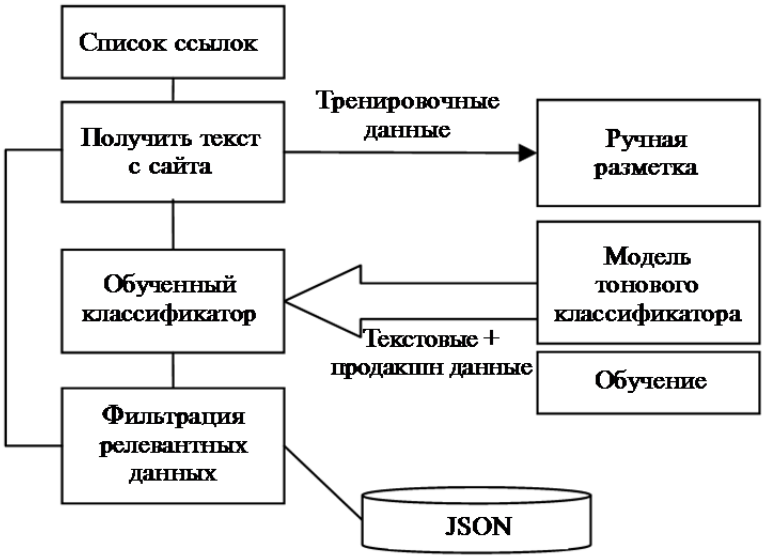

Рис. 1. Схема алгоритма работы системы для извлечения и анализа тематических текстов
Fig. 1. Diagram of the algorithm of the system for extracting and analyzing thematic texts

Этап 6. Разработка классификатора тональности. Для построения модели тонового классификатора рассмотрим две наиболее используемые модели классификации: наивный байесовский классификатор и линейный классификатор на основе стохастического градиента.

Существуют два подхода к наивному байесовскому классификатору: мультиномиальный и многомерный, которые дают разные результаты [24].

Недостатком многомерного байесовского классификатора является то, что он не учитывает количество вхождений слова в документ.

Эта проблема решена в мультиномиальной модели байесовского классификатора, где документ представляет собой последовательность слов.

Многомерная модель дает лучшую оценку предсказания на текстах с небольшим объемом слов. Мультиномиальная

модель – в случае, когда размер текстов составляет несколько тысяч слов [24].

Таким образом, в ходе разработки классификатора целесообразнее использовать мультиномиальную модель байесовского классификатора.

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.

В ряде случаев задачи текстовой классификации, включающие в себя более одного класса, сводятся к нескольким задачам бинарной классификации [25, 26].

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

В программном исполнении наивный байесовский классификатор реализован в библиотеке Scikit-Learn в виде стандартного метода MultinomialNB, а линейный классификатор на основе стохастического градиента – в виде SGDClassifier.

В связи с тем, что объем текстов на транспортном интернет-портале http://autostrada.info/ru составляет десятки тысяч слов, для разработки модели классификатора тональности целесообразно использовать мультиномиальную модель байесовского классификатора (MultinomialNB) и линейную модель классификатора на основе стохастического градиента (SGDClassifier).

Рассмотренные этапы (1–6) в общем виде представляют собой методику построения системы анализа отзывов на основе классификатора тональности.

Программная реализация

Краулер-модуль будем разрабатывать на основе выбранного фреймворка Scrapy. В качестве языка программирования выберем Python 3. Часть программы краулер-модуля представлена в листинге 1.

```
Import scrapy
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
        yield dd
```

Листинг 1. Часть программы краулер-модуля
Listing 1. Part of the crawler module program

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 по 1 ноября 2018 года с сайта http://autostrada.info/ru. Рассмотрим несколько примеров текстов корпуса и того, что в них содержится.

На рис. 2 представлен пример структуры отзыва с сайта http://autostrada.info/ru о состоянии участка трассы, пролегающего между Лугой и Невелем.

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

Например, для отзыва, представленного на рис. 3, атрибуты имеют вид: date: «05.02.2018 15:21»; description: «Участок дороги от Пскова до Луги...»; Subtitle: «Луга – Невель»; title: «Санкт-Петербург – Псков – Невель».

Таким образом, результатом работы краулер-модуля является база данных «dd» с отзывами пользователей.

На следующем этапе осуществляется векторизация и лексический анализ текста.

Для векторизации и лексического анализа текста будем использовать два метода из библиотеки sklearn: CountVectorizer и TF-IDF Vectorizer с мерой TF-IDF. Данные методы используют модели Bag of Words и модель N-gram.

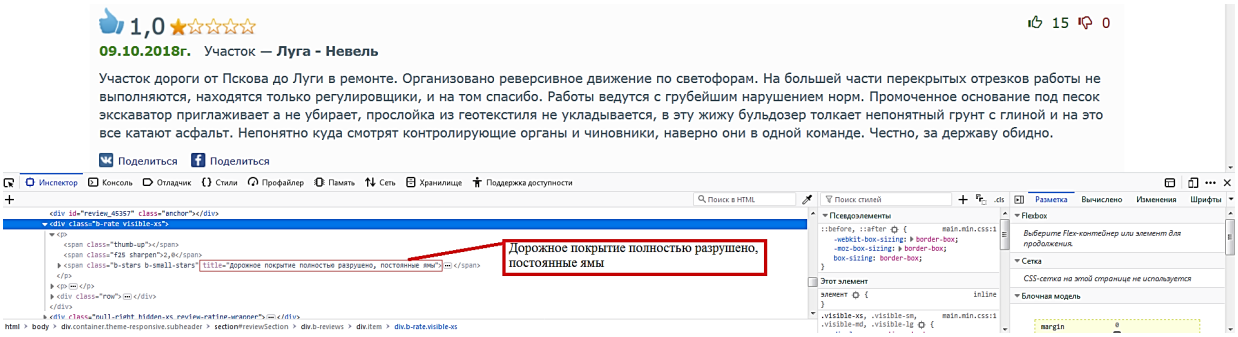

Рис. 2. Структура отзыва на сайте http://autostrada.info/ru

Fig. 2. Review structure on the site http://autostrada.info/ru

```
{'date': '09.10.2018',
 'description': ' Участок дороги от Пскова до Луги в ремонте. Организовано'
                ' реверсивное движение по светофорам. На большей части'
                ' перекрытых отрезков работы не выполняются, находятся только'
                ' регулировщики, и на том спасибо. Работы ведутся с грубейшим'
                ' нарушением норм. Промоченное основание под песок экскаватор'
                ' приглаживает а не убирает, прослойка из геотекстиля не'
                ' укладывается, в эту жижу бульдозер толкает непонятный грунт'
                ' с глиной и на это все катают асфальт. Непонятно куда смотрят'
                ' контролирующие органы и чиновники, наверно они в одной'
                'команде. Честно, за державу обидно'
'subtitle': 'Луга - Невель',
'title': 'Санкт-Петербург – Псков – Невель',
'rate': 'Дорога со значительными разрушениями дорожного полотна',
'url': https://autostrada.info/ru/highway/M-20'}
```

Рис. 3. Структура отзыва в базе данных

Fig. 3. Recall structure in the database

```
# 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 2. Программа тонового классификатора
Listing 2. Listing the tone classifier program

Программную разработку тонового классификатора будем вести на языке Python 3, используя рассмотренные выше модели классификации.

Программа тонового классификатора на основе стандартных методов MultinomialNB и SGDClassifier классификаторов представлена в листинге 2.

Обучение разработанного классификатора будем проводить с использованием готовой выборки[3], состоящей приблизительно из 225 тысяч размеченных твитов, имеющих положительные и отрицательные окрасы.

В ходе тестирования качество классификации было максимизировано путем перебора выбранных различных сочетаний классификатора, метода векторизации, схемы N-gram и других параметров. 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 пять итераций и в тестовую выборку отсекалось 30 % данных. Результаты последних пяти итераций представлены на рис. 4. По результатам валидации лучшей оказалась линейная модель со схемой N-gram: (1, 3) (униграммы + биграммы + триграммы), векторизатором TF-IDF и параметрами: penalty – l2[4], alpha – 0.000001[5], loss – log[6]. Ее результат ≈ 0.72.

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

Результаты

В результате работы тестовой эксплуатации системы для извлечения и анализа дорожно-транспортной информации с сайта http://autostrada.info/ru удалось получить информацию о проблемных участках улично-дорожной сети и неблагоприятных дорожных ситуациях на дорогах Северо-Западного региона России.

Проанализировав классификацию отзывов, получили две выборки: положительные отзывы и отрицательные. Результаты анализа сведены в табл. 4.

Примеры положительных и отрицательных отзывов, полученных при классификации, представлены в табл. 5.

[3] http://study.mokoron.com/

[4] Функция штрафа L2-регуляризация, которая штрафует весовые значения добавлением суммы их квадратов к ошибке.

[5] Константа, которая умножает член регуляризации.

[6] Функция потерь в виде логистической регрессии.

```
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: 0.690909090909
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: 0.719154646092
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. 4. Оценка работы тонового классификатора
Fig. 4. Evaluation of the tone classifier

Таблица 4

Результаты автоматической классификации трасс по отзывам

Table 4

Results of automatic classification of tracks according to reviews

Номер трассы	Наименование трассы	Участок трассы	Протяженность участка, км	Тип отзыва
Отрицательные отзывы			3385	отрицательный
М-20	Санкт-Петербург – Псков – Невель	Луга – Невель	384,0	отрицательный
Р-41	Павлово – Луга	–	190,0	отрицательный
Р-52	Шимск – Феофилова Пустынь	–	79,0	отрицательный
М-10	Новая Ладога	М–10 – Кириши	44,0	отрицательный
А-121	Санкт-Петербург – Сортавала	Санкт-Петербург – Приозерск	154,0	отрицательный
Р-8	Устюжна – Валдай	Боровичи – Лыкошино	267,0	отрицательный
Р-85	Вышний Волочек – Сонково	Вышний Волочек – Максатиха	96,0	отрицательный
М-11	Москва – Санкт-Петербург	обход Вышнего Волочка	76,0	отрицательный
Р-104	Сергиев Посад – Череповец	Рыбинск – Череповец	180,0	отрицательный
Р-37	Лодейное Поле – Вытегра	–	189,0	отрицательный
Р-19	Петрозаводск – Ошта	–	168,0	отрицательный
Р-20	Спасская Губа – А-132	–	197,0	отрицательный
А-132	Суоярви – Юстозеро	–	140,0	отрицательный
Р-18	Беломорск – М-18 "Кола"	–	36,0	отрицательный
А-135	Кемь – Лонка	Калевала – Финлян-дия	89,0	отрицательный
А-137	Тунгозеро – Калевала	–	70,0	отрицательный
А-136	Лоухи – Суоперя	–	170,0	отрицательный
Р-100	Судиславль – Солигалич	–	167,0	отрицательный
Р-100	Судиславль – Солигалич	Судиславль – Галич	76,0	отрицательный
Р-7	Чекшино – Никольск	Тотьма – Никольск	210,0	отрицательный
Р-157	Урень – Котлас	Никольск – Великий Устюг	170,0	отрицательный
Р-87	Ржев – Осташков	–	125,0	отрицательный
А-112	Тверь – Ржев	–	108,0	отрицательный
Положительные отзывы			9874	положительный
А-130	Олонец – Вяртсиля	–	249,0	положительный
А-121	Санкт-Петербург – Сортавала	Приозерск – Сортавала	132,0	положительный
Р-21	Пряжа – Леметти	–	177,0	положительный
А-133	Петрозаводск – Суоярви	–	134,0	положительный
Р-15	Шуйская – Гирвас	–	78,0	положительный
Р-21	Санкт-Петербург – Мурманск	Санкт-Петербург – Медвежьегорск	566,0	положительный

Окончание табл. 4

Номер трассы	Наименование трассы	Участок трассы	Протяженность участка, км	Тип отзыва
Р-17	Медвежьегорск – Великая Губа	–	120,0	положительный
Р-5	Вологда – Медвежьегорск	–	636,0	положительный
Р-2	Долматово – Каргополь	–	223,0	положительный
Р-1	Брин-Наволок – Прокшино	–	517,0	положительный
М-8	Москва – Ярославль – Архангельск	Вологда – Архангельск	770,0	положительный
Р-176	Чебоксары – Йошкар-Ола – Сыктывкар	–	872,0	положительный
Р-56	Великий Новгород – Псков	–	251,0	положительный
М-9	Москва – Волоколамск – Латвия	Москва – Великие Луки	470,0	положительный
М-20	Санкт-Петербург – Псков – Невель	–	500,0	положительный
М-10	Москва – Санкт-Петербург	Москва – Валдай	390,0	положительный
А-111	Торжок – Осташков	–	126,0	положительный
Р-84	Тверь – Устюжна	–	283,0	положительный
А-181	Санкт-Петербург – Выборг – Торфяновка	–	147,0	положительный
А-180	Санкт-Петербург – Нарва	–	160,0	положительный
А-127	Зверево – Малиновка	–	85	положительный
А-123	Зеленогорск – Выборг	–	92,0	положительный
А-125	Молодежное –Черкасово	–	70	положительный
А-120	Молодежное – Большая Ижора	Черемыкино – Кировск	123,0	положительный
А-118	КАД	–	142	положительный
М-18	Санкт-Петербург – Мурманск	Санкт-Петербург – Медвежьегорск	566	положительный
А-114	Вологда – Новая Ладога	Тихвин – Новая Ладога	100	положительный
Р-7	Чекшино – Никольск	Чекшино – Тотьма	150	положительный
Р-157	Урень – Котлас	Урень – Никольск	263	положительный
Р104	Сергиев Посад – Череповец	Калязин – Рыбинск	128	положительный
Р-25	Сыктывкар – Ухта	–	321	положительный
Р-243	Кострома – Киров – Пермь	Киров – Пермь	490	положительный
Р-168	Киров – Адышево – Верхошижемье – Советск	–	137	положительный
Р-176	Чебоксары – Йошкар-Ола – Сыктывкар	Яранск – Киров	250,0	положительный
Р-600	Ярославль – Кострома – Иваново	Кострома – Иваново	103	положительный
Р-101	Островское – Заволжск – Кинешма	–	53	положительный

Таблица 5

Примеры классификации отзывов на трассы Северо-Западного региона

Table 5

Examples of classification of reviews on the tracks of the North-West region

Номер трассы	Трасса	Положительные	Отрицательные
М-10	Москва – Санкт-Петербург (Великий Новгород – Чудово)	Санкт-Петербург – Новгород замечательная дорога	Ужас!!! Ехали 25.08.2018 со стороны Демянска. Нет слов просто, пробки и все из за светофоров, а самое главное из-за ремонта моста. Хотя бы объездную бы придумали самый нужный участок дороги и вот такая Ж...((((((Я не против ремонта это хорошо но варианты объезда тоже нужно было продумывать. С ребёнком 2,5 часа в пробке .(((
М-18	Санкт-Петербург – Мурманск (Санкт-Петербург – Медвежьегорск)	Все хорошо, ехать можно	'До Петрозаводска-ОК! В Карелии много камер
А-114	Вологда – Новая Ладога	'Участок от поворота на Устюжну до Новой Ладоги. Почти везде отличное покрытие, местами прямо немецкий автобан! Трафик небольшой. Про камеры в Вологодской области уже говорили, но там, где их нет, абсолютно спокойно едется 120 и выше! Начиная от Пикалёво (аккуратно едем по объездной, есть один незаметный очень поганый кусок) трафик постепенно начинает расти	Дорога хорошая, но такого количества камер вы не встретите нигде. По-моему это называется – беспредел со стороны ГИБДД. Получила штраф, что называется, на ровном месте. Едешь на разрешенной по трассе скорости, вдруг знак – резкое снижение скорости до 40 км. При этом никакого населенного пункта. Через триста метров, ограничение снимается и ты можешь опять ехать 90-100. Как это называть иначе, как ловушка для того, чтобы стричь штрафы? И таких выдумок на трассе достаточно

В Северо-Западном федеральном округе протяженность положительно оцененных дорог по отзывам пользователей портала Autostrada.info/ru составила 9874 км или 75 %, а отрицательно оцененных дорог – 3385 км или 25 %.

Диаграмма оценки состояния дорог Северо-Западного федерального округа по отзывам пользователей Autostrada.info/ru, с учетом протяженности, представлена на рис. 5.

Для наглядности результатов исследования приведем размеченную карту дорог Северо-Западного федерального округа, соответствующую положительным и отрицательным отзывам (см. рис. 6).

Рис. 5. Оценка дорог Северо-Западного федерального округа по отзывам пользователей Autostrada.info/ru

Fig. 5. Road rating of the North-West Federal District according to user reviews Autostrada.info/ru

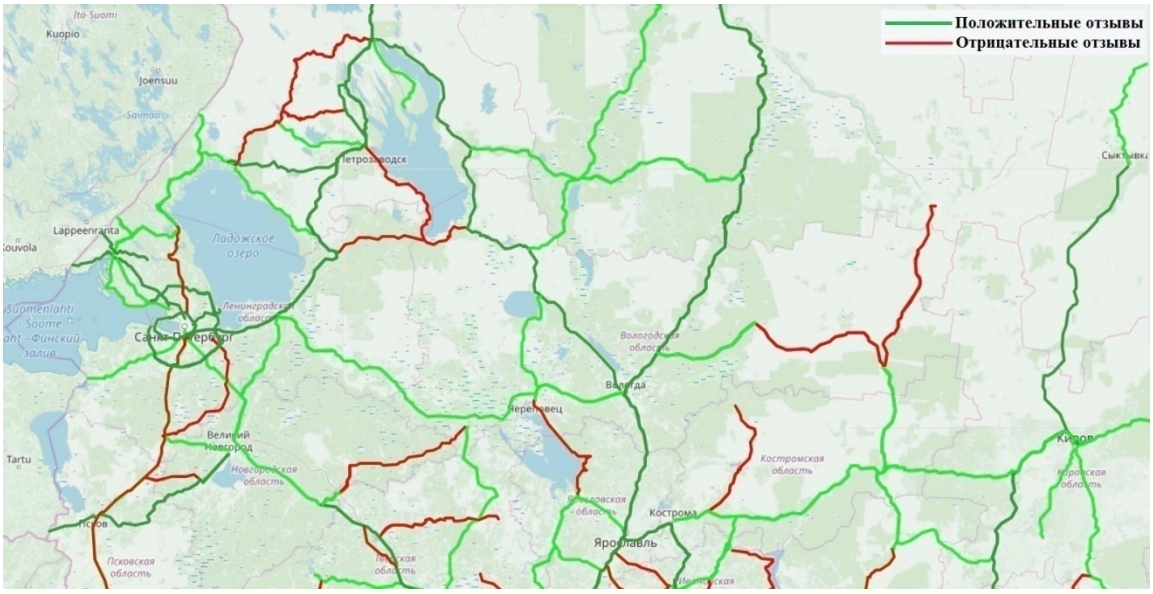

Рис. 6. Размеченная в соответствии с отзывами карта дорог Северо-Западного федерального округа

Fig. 6. Road map of the North-West Federal District marked according to reviews

Выводы

Разработана система для извлечения и тонового анализа отзывов с портала Autostrada.info/ru о дорогах Северо-Западного федерального округа. В результате классификации отзывы были разбиты на две выборки – положительные и отрицательные. Точность классификации составила 71,94 %. В соответствии с отзывами была произведена цветовая разметка карты дорог Северо-Западного федерального округа. Размеченная карта позволила визуально отобразить проблемные участки улично-дорожной сети, а база с отрицательными отзывами – содержательно охарактеризовать указанные в отзыве проблемы.

Данная информация может использоваться дорожными службами в качестве первичной информации для выявления проблемных участков улично-дорожной сети транспортных магистралей, на которых отсутствуют дорожные видеокамеры.

В дальнейшем планируется реализовать глубокую классификацию отзывов по тематическим группам, таким как пробки, ДТП, ремонт, гололед, снежные заторы, штрафы и др. В рамках следующего этапа планируется сравнить методы BAG-of-Words и TF-IDF с методом векторного представления слов Word2Vec, который показал лучшие результаты [27]. Также планируется рассмотреть новые методы тематической классификации

текстов, такие как сверточные нейронные сети, метод опорных векторов и др.

Подобные технологии позволят расширить существующие системы транспортного мониторинга в части учета новых показателей [28, 29] и дадут толчок к развитию новых систем управления дорожным движением [30] и транспортной мобильностью населения [31, 32].

Исследование выполнено в рамках государственного задания Минобрнауки РФ НИОКТР «Разработка теоретических основ организации сложных когнитивных транспортных систем» № АААА-А19-119032590097-6

СПИСОК ЛИТЕРАТУРЫ

1. **Coifman B., Dhoorjaty S.** Event data-based traffic detector validation tests // J. of Transportation Engineering. 2004. Vol. 130(3). Pp. 313–321 // URL: https://doi.org/10.1061/(ASCE)0733-947X(2004)130:3(313)

2. **Maghrour Z.M., Torok A.** Single loop detector data validation and imputation of missing data. Measurement // URL: https://doi. org/10.1016/j.measurement.2017.10.066 (Дата обращения: 01.11.2017)

3. **Laña I., Olabarrieta I. (Iñaki), Vélez M., Del Ser J.** On the imputation of missing data for road traffic forecasting: New insights and novel techniques // Transportation Research Part C: Emerging Technologies. 2018. Vol. 90. Pp. 18–33 // URL: https://doi.org/10.1016/j.trc.2018.02.021

4. **Semwal D., Patil S., Galhotra S., Arora A., Unny N.** (2015). STAR: Real-time spatiotemporal analysis and prediction of traffic insights using social media // Proc. of the 2nd IKDD Conf. on Data Sciences. Bangalore, India: ACM, 2015. P. 7 // URL: http://dx.doi.org/10.1145/2778865.2778872

5. **Gutiérrez C., Figuerias P., Oliveira P., Costa R., Jardim-Goncalves R.** Twitter mining for traffic events detection. 2015 Science and Information Conf. London, UK: IEEE, 2015. Pp. 371–378 // URL: http://dx.doi.org/10.1109/SAI.2015.7237170.

6. **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» // Труды СПИИРАН. 2019. Т. 18. № 2. С. 354–389 // URL: http://dx.doi.org/10.15622/sp.18.2.354-389

7. **Rashidi T.H., Abbasi A., Maghrebi M., Hasan S., Waller T.S.** Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges // Transportation Research Part C: Emerging Technologies. 2017. Vol. 75. Pp. 197–211 // URL: http://dx.doi.org/10.1016/j.trc.2016.12.008

8. **Seliverstov Ya.A., Seliverstov Sv.A., Komashinskiy V.I., Tarantsev A.A., Shatalova N.V., Grigoriev V.A.** Intelligent systems preventing road traffic accidents in megalopolises in order to evaluate // Proc. of 2017 20th IEEE Internat. Conf. on Soft Computing and Measurements. 2017. Pp. 489–92 // URL: https://doi.org/10.1109/CTSYS.2017.8109528

9. **Seliverstov Y.A., Seliverstov S.A., Malygin I.G., Tarantsev A.A., Shatalova N.V., Lukomskaya O.Y., Tishchenko I.P., Elyashevich A.M.** Development of management principles of urban traffic under conditions of information uncertainty // Communications in Computer and Information Science. 2017. Vol. 754. Pp. 399–41 // URL: https://doi.org/10.1007/978-3-319-65551-2_29

10. **Djahel S., Doolan R., Muntean G.-M., Murphy J.** A communications-oriented perspective on traffic management systems for smart cities: Challenges and innovative approaches // IEEE Communication Surveys and Tutorials, 2015. Vol. 17(1). Pp. 125–151 // URL: http://dx.doi.org/10.1109/COMST.2014.2339817

11. **Fu K., Nune R., Tao J.X.** Social media data analysis for traffic incident detection and management // Transportation Research Board 94th Annual Meeting. 2015. No. 15-4022 // URL: https://trid.trb.org/view/1338383

12. **Gal-Tzur A., Grant-Muller S.M., Kuflik T., Minkov E., Nocera S., Shoor I.** The potential of social media in delivering transport policy goals // Transport Policy. 2014. No. 32. Pp. 115–123 // URL: http://dx.doi.org/10.1016/j.tranpol.2014.01.007

13. **Gal-Tzur A., Grant-Muller S.M., Minkov E., Nocera S.** The impact of social media usage on transport policy: Issues, challenges and recommendations // Procedia – Social and Behavioral Sciences. 2014. Vol. 111. Pp. 937–946 // URL: http://dx.doi.org/10.1016/j.sbspro.2014.01.128

14. **Gong Y., Deng F., Sinnott R.O.** Identification of (near) real-time traffic congestion in the cities of Australia through Twitter // Proc. of the ACM 1st Internat. Workshop on Understanding the City with Urban Informatics. Melbourne, Australia: ACM. 2015. Pp. 7–12 // URL: http://dx.doi.org/10.1145/2811271.2811276

15. **Zhenhua Zhang, Ming Ni, Qing He, Jing Gao.** Final report. Mining transportation information from social media for planned and unplanned events. University at Buffalo, SUNY & Transportation Informatics Tier I University Transportation Center, 2016. 68 p.

16. **Zhang Z., He Q., Gao J., Ni M.** A deep learning approach for detecting traffic accidents

from social media data // Transportation Research Part C: Emerging Technologies. 2018. Vol. 86. Pp. 580–596. DOI:10.1016/j.trc.2017.11.027

17. **Шелманов А.О., Каменская М.А., Ананьева М.И., Смирнов И.В.** Семантико-синтаксический анализ текстов в задачах вопросно-ответного поиска и извлечения определений // Искусственный интеллект и принятие решений. 2016. № 4. С. 47–61.

18. **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.

19. **Chen K., Zhang Z., Long J., Zhang H.** Turning from TF-IDF to TF-IGM for term weighting in text classification // Expert Systems with Applications, 2016. Vol. 66. Pp. 245–260. DOI: 10.1016/j.eswa.2016.09.009

20. **Bissan Ghaddar, Joe Naoum-Sawaya.** High dimensional data classification and feature selection using support vector machines // European J. of Operational Research. 2018. Vol. 265. Iss. 3. Pp. 993–1004.

21. **Jimenez-Marquez J. L., Gonzalez-Carrasco I., Lopez-Cuadrado J.L., Ruiz-Mezcua B.** Towards a big data framework for analyzing social media content // Internat. J. of Information Management. 2019. Vol. 44. Pp. 1–12. DOI: 10.1016/j.ijinfomgt.2018.09.003

22. **Bissan Ghaddar, Joe Naoum-Sawaya.** High dimensional data classification and feature selection using support vector machines // European J. of Operational Research. 2018. Vol. 265. Iss. 3. Pp. 993–1004.

23. **Dey A., Jenamani M., Thakkar J.J.** Senti-N-Gram : An N-gram lexicon for sentiment analysis // Expert Systems with Applications. 2018. Vol. 103. Pp. 92–105. DOI: 10.1016/j.eswa.2018.03.004

24. **Сизов А.А., Николенко С.И.** Наивный байесовский классификатор. DOCPLAYER // URL: https://docplayer.ru/45424867-Naivnyy-bayesovskiy-klassifikator.html (Дата обращения: 25.01.2019).

25. **Воронцов К.В.** Вероятностное тематическое моделирование // URL: http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf

26. **Воронцов К.В.** Лекции по линейным алгоритмам классификации // URL: http://www.machinelearning.ru/wiki/images/6/68/voron-ML-Lin.pdf

27. **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.

28. **Селиверстов С.А., Селиверстов Я.А.** Обзор показателей транспортной обеспеченности мегаполиса // Вестник гражданских инженеров. 2015. № 5 (52). С. 237–247.

29. **Селиверстов С.А., Селиверстов Я.А.** О методе оценки эффективности организации процесса дорожного движения мегаполиса // Вестник транспорта Поволжья. 2015. № 2 (50). С. 91–96.

30. **Селиверстов С.А., Селиверстов Я.А.** Метод построения пути субъективного предпочтительного следования // Известия СПбГЭТУ ЛЭТИ. 2016. Т. 4. С. 31–37.

31. **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 СПбГПУ. Информатика. Телекоммуникации. Управление. 2016. № 1 (236). С. 65–81. DOI: 10.5862/JCSTCS.236.7

32. **Шаталова Н.В.** Развитие транспортной отрасли как основополагающей при решении проблем стратегического и экономического характера // Модернизация и научные исследования в транспортном комплексе. 2017. Т. 1. С. 230–233.

Статья поступила в редакцию 30.06.2019.

REFERENCES

1. **Coifman B., Dhoorjaty S.** Event data-based traffic detector validation tests. *J. of Transportation Engineering*, 2004, Vol. 130(3), Pp. 313–321. Available: *https://doi.org/10.1061/(ASCE)0733-947X(2004)130:3(313)*

2. **Maghrour Zefreh M., Torok A.** Single loop detector data validation and imputation of missing data, Measurement. Available: *https://doi.org/10.1016/j.measurement.2017.10.066* (Accessed: 01.11.2017).

3. **Laña I., Olabarrieta I. (Iñaki), Vélez M., Del Ser J.** On the imputation of missing data for road traffic forecasting: New insights and novel techniques. *Transportation Research Part C: Emerging Technologies*, 2018, Vol. 90, Pp. 18–33. Available: *https://doi.org/10.1016/j.trc.2018.02.021*

4. **Semwal D., Patil S., Galhotra S., Arora A., Unny N.** STAR: Real-time spatiotemporal analysis and prediction of traffic insights using social media. *Proceedings of the 2nd IKDD Conference on Data Sciences.* Bangalore, India: ACM, 2015, P. 7 Available: *http://dx.doi.org/10.1145/2778865.2778872*

5. **Gutiérrez C., Figuerias P., Oliveira P., Costa R., Jardim-Goncalves R.** Twitter mining for traffic events detection. *2015 Science and Information Conference*. London, UK: IEEE, 2015, Pp. 371–378. Available: *http://dx.doi.org/10.1109/SAI.2015.7237170*

6. **Seliverstov Y.A., Chigur V.I., Sazanov A.M., Seliverstov S.A., Svistunova A.S.** Sentiment Analysis of «AUTOSTRADA.INFO/RU" Users» Comments. *SPIIRAS Proceedings*, 2019, Vol. 18(2), Pp. 354–389. Available: *https://doi.org/10.15622/sp.18.2.354-389*

7. **Rashidi T.H., Abbasi A., Maghrebi M., Hasan S., Waller T.S.** Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges. *Transportation Research Part C: Emerging Technologies*, 2017, Vol. 75, Pp. 197–211. Available: *http://dx.doi.org/10.1016/j.trc.2016.12.008*

8. **Seliverstov Ya.A., Seliverstov Sv.A., Komashinskiy V.I., Tarantsev A.A., Shatalova N.V., Grigoriev V.A.** Intelligent systems preventing road traffic accidents in megalopolises in order to evaluate. *Proceedings of 2017 20th IEEE International Conference on Soft Computing and Measurements*, 2017, Pp. 489–92. Available: *https://doi.org/10.1109/CTSYS.2017.8109528*

9. **Seliverstov Y.A., Seliverstov S.A., Malygin I.G., Tarantsev A.A., Shatalova N.V., Lukomskaya O.Y., Tishchenko I.P., Elyashevich A.M.** Development of management principles of urban traffic under conditions of information uncertainty. *Communications in Computer and Information Science*, 2017, Vol. 754, Pp. 399–41. Available: *https://doi.org/10.1007/978-3-319-65551-2_29*

10. **Djahel S., Doolan R., Muntean G.-M., Murphy J.** A communications-oriented perspective on traffic Management Systems for Smart Cities: Challenges and innovative approaches. *IEEE Communication Surveys and Tutorials*, 2015, Vol. 17(1), Pp. 125–151. Available: *http://dx.doi.org/10.1109/COMST.2014.2339817*

11. **Fu K., Nune R., Tao J.X.** Social media data analysis for traffic incident detection and management. *Transportation Research Board 94th Annual Meeting*, 2015, No. 15-4022. Available: *https://trid.trb.org/view/1338383*

12. **Gal-Tzur A., Grant-Muller S.M., Kuflik T., Minkov E., Nocera S., Shoor I.** The potential of social media in delivering transport policy goals. *Transport Policy*, 2014, No. 32, Pp. 115–123. Available: *http://dx.doi.org/10.1016/j.tranpol.2014.01.007*

13. **Gal-Tzur A., Grant-Muller S.M., Minkov E., Nocera S.** The impact of social media usage on transport policy: Issues, challenges and recommendations. *Procedia – Social and Behavioral Sciences*, 2014, Vol. 111, Pp. 937–946. Available: *http://dx.doi.org/10.1016/j.sbspro.2014.01.128*

14. **Gong Y., Deng F., Sinnott R.O.** Identification of (near) real-time traffic congestion in the cities of Australia through Twitter. *Proceedings of the ACM first international workshop on understanding the city with urban informatics. Melbourne, Australia*: *ACM*, 2015, Pp. 7–12. Available: *http://dx.doi.org/10.1145/2811271.2811276*

15. **Zhenhua Zhang, Ming Ni, Qing He, Jing Gao.** *Final report. Mining transportation information from social media for planned and unplanned events.* University at Buffalo, SUNY & Transportation Informatics Tier I University Transportation Center, 2016, 68 p.

16. **Zhang Z., He Q., Gao J., Ni M.** A deep learning approach for detecting traffic accidents from social media data. *Transportation Research Part C: Emerging Technologies*, 2018, Vol. 86, Pp. 580–596. DOI: 10.1016/j.trc.2017.11.027

17. **Shelmanov A.O., Kamenskaya M.A., Ananyeva M.I., Smirnov I.V.** Semantic-syntactic analysis for question answering and definition extraction. *Artificial Intelligence and Decision Making*, 2016, No. 4, Pp. 47–61. (rus)

18. **Kuznetsov A.N., Vyshemirskiy D.A.** One approach to solving tokenization problem for analysis of large-scale collections of user-defined passwords. *Bezopasnost informatsionnykh tekhnologiy* [*IT Security*], 2017, No. 2, Pp. 50–60. (rus)

19. **Chen K., Zhang Z., Long J., Zhang H.** Turning from TF-IDF to TF-IGM for term weighting in text classification. *Expert Systems with Applications*, 2016, Vol. 66, Pp. 245–260. DOI: 10.1016/j.eswa.2016.09.009

20. **Bissan Ghaddar, Joe Naoum-Sawaya.** High dimensional data classification and feature selection using support vector machines. *European Journal of Operational Research*, 2018, Vol. 265, Iss. 3, Pp. 993–1004.

21. **Jimenez-Marquez J.L., Gonzalez-Carrasco I., Lopez-Cuadrado J.L., Ruiz-Mezcua B.** Towards a big data framework for analyzing social media content. *International Journal of Information Management*, 2019, Vol. 44, Pp. 1–12. DOI: 10.1016/j.ijinfomgt.2018.09.003

22. **Bissan Ghaddar, Joe Naoum-Sawaya.** High dimensional data classification and feature selection using support vector machines. *European Journal of Operational Research*, 2018, Vol. 265, Iss. 3, Pp. 993–1004.

23. **Dey A., Jenamani M., Thakkar J.J.** Senti-N-Gram: An N-gram lexicon for sentiment analysis. *Expert Systems with Applications*, 2018, Vol. 103, Pp. 92–105. DOI: 10.1016/j.eswa.2018.03.004

24. **Sizov A.A., Nikolenko S.I.** Naivnyy Bayesovskiy klassifikator. DOCPLAYER [Naive Bayes classifier. DOCPLAYER]. Available: *https://docplayer.ru/45424867-Naivnyy-bayesovskiy-klassifikator.html* (Accessed: 25.01.2019). (rus)

25. **Vorontsov K.V.** Veroyatnostnoye tematicheskoye modelirovaniye [Probabilistic thematic modeling]. Available: *http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf* (rus)

26. **Vorontsov K.V.** Lektsii po lineynym algoritmam klassifikatsii [Lectures on linear classification algorithms]. Available: *http://www.machinelearning.ru/wiki/images/6/68/voron-ML-Lin.pdf*

27. **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.

28. **Seliverstov S.A., Seliverstov Ya.A.** Review of megalopolis transportprovision indicators. *Vestnik grazhdanskikh inzhenerov* [*Bulletin of Civil Engineers*], 2015, No. 5 (52), Pp. 237–247. (rus)

29. **Seliverstov S.A., Seliverstov Ya.A.** O metode otsenki effektivnosti organizatsii protsessa dorozhnogo dvizheniya megapolisa [On a method for assessing the effectiveness of the organization of the process of traffic in a megalopolis]. *Vestnik transporta Povolzhya*, 2015, No. 2 (50), Pp. 91–96. (rus)

30. **Seliverstov S.A., Seliverstov Ya.A.** The determination method of the subjective preferred route. *Izvestiya SPbGETU LETI*, 2016, Vol. 4. Pp. 31–37. (rus)

31. **Seliverstov Ya.A., Seliverstov S.A.** Use of GATLOSAMI to prevent causes of traffic accidents and adverse social accidents in a 'SMART CITY'. *St. Petersburg State Polytechnical University Journal. Computer Science. Telecommunications and Control Systems*, 2016, No. 1 (236), Pp. 65–81. DOI: 10.5862/JCSTCS.236.7. (rus)

32. **Shatalova N.V.** Razvitiye transportnoy otrasli kak osnovopolagayushchey pri reshenii problem strategicheskogo i ekonomicheskogo kharaktera [Development of the transport industry as fundamental in solving problems of a strategic and economic nature]. *Modernizatsiya i nauchnyye issledovaniya v transportnom komplekse* [*Modernization and research in the transport sector*], 2017, Vol. 1, Pp. 230–233. (rus)

Received 30.06.2019.

СВЕДЕНИЯ ОБ АВТОРАХ / THE AUTHORS

СЕЛИВЕРСТОВ Ярослав Александрович
SELIVERSTOV Yaroslav A.
E-mail: maxwell_8-8@mail.ru

НИКИТИН Кирилл Вячеславович
NIKITIN Kirill V.
E-mail: execiter@mail.ru

ШАТАЛОВА Наталья Викторовна
SHATALOVA Natalya V.
E-mail: shatillen@mail.ru

КИСЕЛЕВ Арсений Алексеевич
KISELEV Arseny A.
E-mail: ars8ars@mail.ru