# ИССЛЕДОВАНИЕ МЕХАНИЗМОВ ЭЛЕКТРОННОГО ВЗАИМОДЕЙСТВИЯ ГОРОДСКИХ ВЛАСТЕЙ И ЖИТЕЛЕЙ: РЕЗУЛЬТАТЫ РАЗРАБОТКИ КЛАССИФИКАТОРА ОБРАЩЕНИЙ ГОРОЖАН В ПЕТЕРБУРГЕ

***Л. А. Видясова, А. С. Антонов***
*Университет ИТМО*
Санкт-Петербург

В докладе представлены результаты разработки метода выявления общественных пространств и реакции на них горожан. В основе работы лежит применение методов обработки естественного языка (НЛП) к текстовым сообщениям граждан, полученным на площадках органов власти в социальных сетях. При разработке классификатора были выделены 12 основным категорий. Кроме того, в работе использовался метод распознавания городских объектов по детекции геолокационных параметров. Исследовательская группа разработала модель с использованием Python и специальных библиотек.

**Ключевые слова:** обращения граждан, обнаружение сущностей, социальные сети, анализ текстов

# STUDY OF THE MECHANISMS OF ELECTRONIC INTERACTION BETWEEN CITY AUTHORITIES AND RESIDENTS: RESULTS OF THE DEVELOPMENT OF A CLASSIFIER OF CITIZENS' APPEALS IN ST. PETERSBURG

***L. A. Vidiasova, A. S. Antonov***
*ITMO University*
Saint-Petersburg

The report presents the results of developing a method for identifying public spaces and citizens' reactions to them. The work is based on the application of natural language processing (NLP) methods to text messages of citizens received on the platforms of government bodies in social networks. When developing the classifier, 12 main categories were identified. In addition, the work used the method of recognizing urban objects by detecting geolocation parameters. The research team developed the model using Python and special libraries.

**Keywords:** citizens’ applications, entities detection, social media, text analysis

В последние годы в крупных городах наметилась тенденция цифровизации и, как следствие, формирование цифрового пространства взаимодействия населения и городских властей. С 2021 года в Санкт-Петербурге активно реализуется концепция экосистемы городских цифровых сервисов [1] как пространства взаимодействия государственных и коммерческих сервисов с жителями города при развитии цифровой среды города. Внедрение такого подхода означает признание того факта, что город может полностью удовлетворить потребности горожан только в сотрудничестве с партнерами и заинтересованными сторонами. Получение обратной связи от граждан необходимо для эффективного и полезного функционирования общественных пространств.

В то же время обращения граждан представляют собой большой поток плохо структурированной информации, которая может служить важными индикатором как очагов городских проблем, социальной напряженности, так и с целом отражать социальное самочувствие в той или иной городской среде. Исследовательская повестка для анализа форм гражданского электронного участия достаточно широка. В виду все возрастающей критики традиционных методик оценки, на арену выходят методы автоматизированного анализа данных [3]. Неструктурированные данные, ежесекундно создаваемые пользователями цифровых платформ по всему миру, нуждаются в инструментах и методах, которые позволят автоматически извлекать полезную информацию из текстов [4].

Интеллектуальный анализ текста предполагает обнаружение и извлечение интересных, нетривиальных знаний из неструктурированного или слабо структурированного текста. Применение этого метода включает в себя ряд процессов от непосредственного поиска информации до классификации текста, последующей кластеризации, извлечения сущностей, отношений и событий [5]. Интеллектуальный анализ текста и обработка естественного языка становятся важными там, где размер изучаемых текстовых материалов не позволяет провести ручной анализ.

В докладе представлены результаты разработки классификатора сообщений горожан, поступающих через официальные аккаунты органов власти в социальных сетях. Исследование было проведено на материале Санкт-Петербурга. Целью работы являлось применение методов машинной обработки естественного языка для разработки экстрактора информации из обращений горожан [2].

**Методы и инструменты**

Использованный авторами метод основан на применении методов обработки естественного языка (НЛП) к текстовым сообщениям граждан, полученным в социальных сетях. Первый метод НЛП представляет собой каскад классификаторов на основе предварительно обученной языковой модели. Здесь мы делим сообщения на основные категории и функциональные подкатегории. Затем был использован второй метод для определения примерного места событий в сообщениях.

Для разработки модели были использованы 2 набора данных:

– массивы обращений и сообщений горожан Петербурга из социальных сетей, содержащие 71 тыс. записей за период с 30.12.2021 по 23.02.2022;

– массив сообщений граждан из социальных сетей одного района Санкт-Петербурга, содержащий 18168 записей за период с 05.06.2017 по 25.01.2023.

Рубрикатор в массиве включал такие блоки, как коммунальные услуги, дороги, жилье, здравоохранение, образование, социальная защита, строительство, вывоз мусора, транспорт, охрана окружающей среды, энергетика, безопасность.

Для классификации сообщений по категориям мы использовали предварительно размеченный набор. Сначала он был предварительно обработан с помощью пакета re python. Знаки препинания, хэштеги, повторяющиеся пробелы были разделены. Далее из библиотеки nltk был загружен список стоп-слов на русском языке, а также отфильтрованы слова из текстов. В данной работе в качестве исходной языковой модели для обучения, а также для лемматизации и морфологического анализа использовалась модель spacy-ru (https://github.com/buriy/spacy-ru). При таком подходе выделялись и лемматизировались слова, связанные с существительными, прилагательными, глаголами и наречиями. Эти наборы легли в основу дальнейшего обучения модели.

**Результаты**

Для обучения атрибут категорий был преобразован в отдельные атрибуты с помощью процедуры быстрого кодирования. Полученный набор текста и соответствующий вектор были рандомизированы и разделены на обучающий набор, который составлен 0,9 от исходного, и тестовый набор, который составил 0,1 от исходного. Подобное соотношение широко используется при обучении с ограниченным объемом данных. Обучение проводилось с размером пакетов, равным 32, и коэффициентом отсева = 0,2.

Для эксперимента использовался набор комментариев из социальной сети ВК от официальной группы Адмиралтейского района (Санкт-Петербург, Россия). Такие группы часто используются жителями как площадка для обращения к властям по различным вопросам. Первоначальный объем сообщений составил 18250 записей с 05.06.2017 по 25.01.2023.

Далее тексты адресов были предварительно обработаны: из списка русских стоп-слов nltk были удалены стоп-слова и знаки препинания. Остальные слова были лемматизированы и представлены в нижнем регистре. Полученный набор текстов был классифицирован с использованием предварительно обученной пространственной модели. Те тексты, которые не удалось отнести ни к одной категории с вероятностью больше 0,6, были удалены из набора, сократив его до 18168 попаданий. Распределение обращений по категориям на данный момент показано на рис. 1.

Далее к полученному набору был применен алгоритм геолокации: сначала были выделены топонимы из текстов с помощью предварительно обученной модели NER. Сообщения без определенных топонимов, а также с неверно определенными (длина результата более 5 слов) отфильтровывались. В результате набор сократился до 3845 записей. Затем методом нечеткого сопоставления имен для каждого топонима, находящегося в обращении, были добавлены геоданные, взятые из OpenStreetMaps. Таким образом, также была отсеяна часть сообщений с топонимами, не обнаруженными в адресной системе города (2727 сообщений). При этом, если в одном обращении упоминалось несколько мест, оно дублировалось для каждого из них. Таким образом, общее количество данных составило 16293 записи.

Также были выявлены тенденции распределения обращений в динамике. В течение нескольких лет основные пики обращений граждан выявлялись в зимние месяцы, связанные со сферой безопасности. Это можно объяснить большим потоком жалоб на низкую температуру в помещении. Такое распределение также согласуется с общей динамикой сообщений в Адмиралтейском районе, которая представлена на рисунке 2.

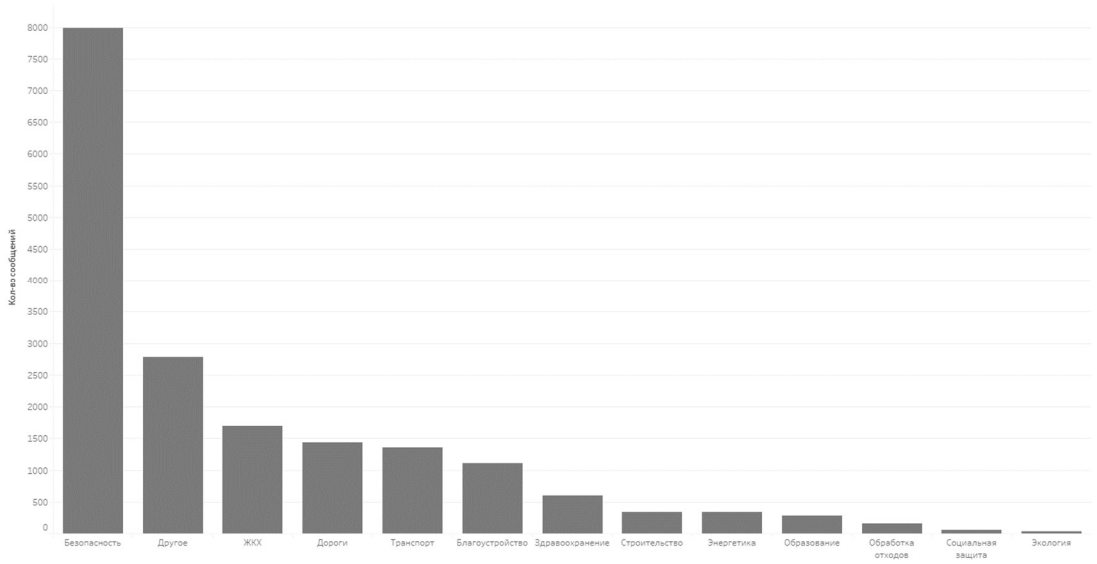


**Рис. 1.** Распределение обращений горожан Петербурга по категориям

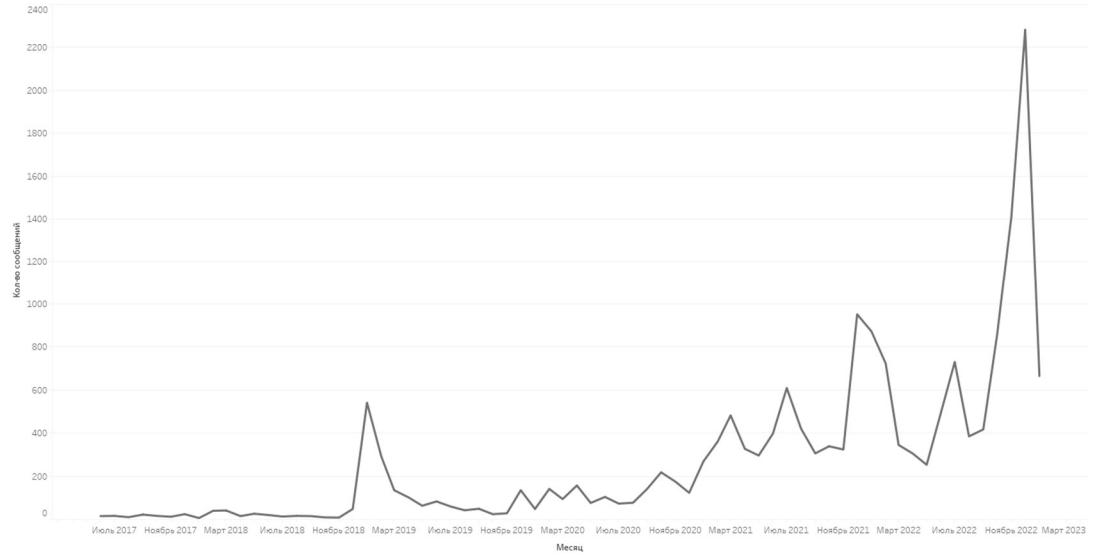


**Рис. 2.** Динамика публикации обращений гражданами в Адмиралтейском районе г. Санкт-Петербурга

**Выводы**

В работе представлены результаты разработанной авторской технологии – классификатора обращений граждан, апробированной на ретроспективных данных. Представленный метод помогает быстро получать структурированную информацию о возникающих проблемах, требующих решения. В качестве перспективных сфер для дальнейшего применение классификатора следует выделить: исследование субъективных оценок качество городской среды, а также выделение ситуаций и объектов повышенного социального риска в городе.

Однако в текущей версии алгоритм имеет некоторые ограничения. На данном этапе исследования сложно точно определить адрес на основании одного текста обращения в случае схожести географических названий. Кроме того, возникают сложности с определением более узких тематик обращений. Эти вопросы лягут в основу дальнейших исследований.

Исследование проведено в рамках НИР Университета ИТМО № 622264 «Разработка сервиса выявления объектов городской среды общественной активности и ситуаций повышенного риска на основе текстовых сообщений горожан».

## ЛИТЕРАТУРА

1. Frias-Martinez V., Frias-Martinez E. Spectral clustering for sensing urban land use using Twitter activity // Eng. Appl. Artif. Intell. 2014. № 35. P. 237–245.

2. Begen P., Chugunov A. Intellectual classifier development of citizens' messages on the “Our St. Petersburg” portal: Experience in using machine learning methods // CEUR Workshop Proceedings: SSI 2019 - Proceedings of the 21st Conference on Scientific Services and Internet, Novorossiysk-Abrau, 2020. Vol. 254. P. 82-92.
3. Kabanov Y. Refining the UN E-participation Index: Introducing the deliberative assessment using the Varieties of Democracy data // Government Information Quarterly. 2022. Vol. 39. Iss. 1. Art. 101646. DOI: 10.1016/j.giq.2021.101656.
4. Mead M., Popoola O., Popoola G., Landshoff P., Calleja M., Hayes M., Baldovi J. The use of Eletrochemical Sensors for monitoring urban air quality in lowcost, high-density netwroks // Atmospheric Environment. 2013. № 70. P. 186-203.
5. Poteet S.R. Natural Language Processing and Text Mining. Springer Publishing Company, 2006.