Классификатор объектов городского хозяйства для данных из социальных сетей

П.Н. Беген[1], Б.А. Низомутдинов[1], А.С. Тропников[1]

1. Университет ИТМО

Аннотация. В статье приводятся результаты анализа сообщений из социальных сетей, основная задача – по имеющемуся набору данных, представляющих собой сообщения пользователей и собранных на разных площадках, автоматически определить, какой объект городского хозяйства упоминается в тексте отзыва, а также определить отношение пользователя к данному объекту (позитивное, нейтральное или негативное). Авторы представляют комплексный сценарий для проведения исследований, включающий сбор информации, контент-анализ, методы классификации и кластеризации текста, анализ тональности сообщений, изучение обмена информацией между пользователями социальных сетей. Пилотное исследование проведено на примере 2 административных районов в Санкт-Петербурге – Петроградском и Кронштадтском. Средняя точность классификации модели алгоритма составила 78 %. Обученная модель алгоритма была применена в разработке онлайн сервиса для проведения классификации постов.

Ключевые слова: машинное обучение, классификация, кластеризация, обработка естественного языка, нейронные сети, социальные сети

Classifier of urban objects for data from social networks

P.N. Begen[1], B.A. Nizomutdinov[1], A.S. Tropnikov[1]

1. ITMO University

Abstract. The article presents results of messages analysis from social networks. The main task was to automatically determine which urban object is mentioned in the review text using available set of data, which are user messages collected on different sites, in social networks, as well as to determine user's attitude to this object (positive, neutral or negative). The authors present a complex scenario for conducting research, including information collection, content analysis, methods for classifying and clustering text, analysis of the messages tone, and studying the information exchange between users of social networks. The pilot study was considered with the example of 2 districts in St. Petersburg – Petrogradsky and Kronshtadtsky. The average classification

precision of the algorithm model was 78 %. The trained algorithm model was used in the development of an online service for classifying posts.

Keywords: Machine Learning, classification, clustering, Natural Language Processing, neural networks, social networks

1. Введение

Для эффективного управления территорией при реализации проектов по благоустройству городских территорий необходимо проводить обсуждения с жителями, собирать мнение граждан, организовывать встречи для соучастного проектирования. Это сложные мероприятия, требующие финансовых затрат. Однако, часто необходимо собирать отзывы и общественное мнение в короткие сроки или в режиме онлайн. Проблема возникает, когда жители игнорируют встречи по проведению общественных обсуждений с муниципальными властями. Все это понижает качество реализуемых проектов по благоустройству, т. к. нет достаточного количества идей или отзывов.

Помимо прочего, граждане активно высказываются и критикуют благоустройство территорий в социальных сетях, ежедневно оставляя комментарии в различных группах, например, в сообществах своего дома, района или города. Это большой пласт информации, который сможет помочь в принятии решений для органов власти при реализации проектов благоустройства, а также для обнаружения недочетов в существующей инфраструктуре.

В настоящее время продолжается рост вовлечения пользователей в социальные сети, при этом расширяется сфера применения: граждане используют сообщества для решения повседневных проблем, обсуждают городское хозяйство, благоустройство своего района или двора, участвуют в партисипаторном бюджетировании [1]. Одновременно с этим происходит преобразование института общественных коммуникаций и каналов взаимодействия с властью.

В онлайн сообществах генерируется огромное число сообщений, и такой объем информации может показаться слабоструктурированным и не пригодным для исследования и дальнейшего использования, однако, существуют подходы, благодаря которым все посты можно классифицировать автоматизированным способом и определить, о каком объекте идет речь.

В настоящей статье приводятся результаты исследования сообщений и высказываний, связанных с городской средой и общественными пространствами на примере Санкт-Петербурга. Перед группой исследователей стояла задача: по имеющемуся набору данных, представляющих собой сообщения пользователей и собранных на разных площадках или в социальных сетях, автоматически определить, какой

объект городского хозяйства упоминается в тексте отзыва, а также автоматически определить отношение пользователя к данному типу объекта (позитивное, нейтральное или негативное).

Авторы предлагают подход, основанный на методах машинного обучения. Подчеркивается возможность применения автоматических или автоматизированных инструментов для исследования данных из социальных сетей. Предложенный подход предназначен для изучения отдельных онлайн сообществ в социальных сетях, а также может быть применен для изучения процессов, происходящих в отдельном городе, районе, области.

Полученные результаты можно использовать на муниципальном уровне власти при реализации различных городских проектов по благоустройству, а также для отслеживания текущего состояния городской и дворовой инфраструктуры, что позволит сократить расходы на проведение дополнительных мероприятий по сбору информации. Для этой цели необходимо решить задачу сбора отзывов в сообществах социальных сетей, затем классифицировать все сообщения по категориям, чтобы из всего массива информации отобрать именно ту, которая требуется для конкретного проекта благоустройства. Кроме того, классификатор позволит оценивать общее состояние городской инфраструктуры в том или ином районе, что в итоге приведет к сокращению расходов на проведение мониторинга и сокращению времени принятия решения для устранения выявленных проблем.

2. Программная реализация классификатора

Для реализации автоматического классификатора объектов городского хозяйства на основе данных из социальных сетей были поставлены следующие задачи:

- для автоматического определения объекта в тексте отзыва или комментария пользователя необходимо разработать алгоритм решения задач классификации и кластеризации текста для последующего сравнения экспертами результатов классификатора, использующего ручную и/или автоматическую (машинную) разметки;
- для автоматического определения типа тональности (позитивный, нейтральный или негативный оттенок) текста отзыва или комментария необходимо выбрать готовый и публичный русскоязычный массив текста с заданной разметкой на положительные и отрицательные группы текстов, и на основе данного массива подготовить модель по автоматическому определению тональности текста отзывов.

Для решения задачи автоматического определения типа объекта в текстовых данных был разработан алгоритм для решения 2-ух

классических задач машинного обучения: классификации и кластеризации текста.

Классификацией считается разбиение множества документов на заранее известные группы на основе каких-либо параметров или свойств.

Кластеризация — разбиение множества схожих документов на кластеры или подмножества, параметры которых заранее неизвестны. Количество кластеров может быть произвольным (т. е. алгоритм сам определяет нужное количество) или фиксированным (т. е. заданных пользователем на начальном этапе реализации алгоритма).

Машинное обучение представляет собой широкий раздел области искусственного интеллекта, который изучает методы построения самообучающихся алгоритмов. В своей статье [2] В. Н. Вапник одним из первых рассмотрел «Теорию статистического обучения в роли одного из возможным вариантов развития и работы машинного обучения», что позволило дать более четкую постановку задачи машинного обучения в целом и сформировать схему механизмов работы различных алгоритмов на годы вперед.

Йоав Голдберг (Yoav Goldberg) в своих работах [3, 4], Т. В. Батура в обзорном исследовании методов автоматической классификации текстов [5] и А. О. Зиберт с В. И. Хрусталевым в [6] приводят основные методы и подходы в области обработки естественного языка (англ. Natural Language Processing, сокр. NLP) в рамках работы с нейронными сетями разных архитектур и со стандартными статистическими моделями для реализации методов глубокого обучения (англ. deep learning), а также приводят основные результаты тестирования, экспериментов и показатели, полученные в ходе реализации различных методов.

Среди основных методов и подходов обработки естественного языка можно выделить следующие:

- токенизация или сегментация;
- составление и использование списка стоп-слов;
- стемминг;
- лемматизация;
- извлечение именованных сущностей (англ. Named Entity Recognition);
- модель «мешок слов» (англ. bag-of-words);
- вычисление функции TF-IDF, сокращение размерности;
- алгоритмы Word2Vec.

Представленные методы и подходы по обработке естественного языка легли в основу разработанного решения для анализа текстовых данных из социальных сетей.

3. Данные для обработки

В работе проведено исследование данных в социальных сетях на примере 2 административных районов Санкт-Петербурга — Петроградский и Кронштадтский. Приведем краткие характеристики по каждому: население Кронштадта составляет 43687 человек (2017 год), площадь острова — 1584 га, всего территория Кронштадтского района в существующих утвержденных границах составляет 1935 га. Петроградский район: географическая площадь района 24 км^2, численность населения 131 356 человек [7].

В качестве источников информации для формирования набора данных были выбраны следующие ресурсы:

- https://vk.com/
- https://twitter.com/
- https://pikabu.ru/
- https://www.tripadvisor.ru/
- https://gorod.gov.spb.ru/
- https://local.yandex.ru/

В социальной сети «ВКонтакте» (vk.com) были отобраны сообщества, которые имеют отношение к указанным районам по географическому признаку в названии. Парсер контента из сообществ осуществлялся без использования специализированного API, т. е. методом парсинга. Работа парсера устроена следующим образом: на примере одной страницы задается граница парсинга, откуда нужно собирать информацию, далее настраивается шаблон вывода и формат сохранения информации. В парсер загружается список URL для сбора, она выгружает весь HTML код заданной страницы, макрос «прокрутки страницы» позволяет загрузить все записи.

Следующий шаг парсинга — извлечение из полученного кода нужной информации. Получив исходный код HTML-страницы, необходимо выполнить над ним обработку, т. е. отделить искомый текст от гипертекстовой разметки, выстроить иерархическое дерево элементов документа (DOM) и извлечь из страницы искомую информацию. По заданным критериям выделить только основную информацию, которая представляет интерес. Для сбора основных параметров публикации (сообщения пользователя) на стене сообщества (групп «ВКонтакте») были заданы границы парсинга каждого параметра на примере одного сообщения. Анализировался исходных код страницы и выделялись теги HTML, содержащие необходимые параметры. Данный способ имеет ряд ограничений: 1) время сбора информации; 2) отсутствие возможности отслеживать динамику в режиме реального времени; 3) работа только с готовой базой. В ходе парсинга был осуществлен сбор данных объемом более 200000 записей.

В настоящее время коллектив исследователей ведет разработку парсера «ВКонтакте» по API для увеличения скорости сбора данных. Парсинг контента через API ВКонтакте. API ВКонтакте — это внешний интерфейс, который позволяет получать информацию из базы данных vk.com с помощью HTTP-запросов к специальному серверу. Синтаксис запросов и тип возвращаемых данных строго определены на стороне самого сервиса. «ВКонтакте» — это социальная сеть, где есть дружеские связи, настройки приватности и черные списки. Многое зависит от того, кто просматривает страницу: кто-то увидит на ней всю ту же информацию, что и владелец, а кто-то — лишь общедоступные данные. В API этот принцип также сохраняется. Описание возможностей API ВКонтакте приведено здесь: https://vk.com/dev/manuals.

Среди социальных сетей Твиттер лучше других подходит для добычи текстовых данных в силу жесткого ограничения на длину сообщения (280 символов), в которое пользователи вынуждены поместить все самое существенное. Авторы работы настроили парсинг данных через API Twitter. API упрощает создание самого кода, поскольку предоставляет набор готовых классов, функций или структур для работы с имеющимися данными. Научным коллективом был разработан парсер Twitter, который собирает данные с платформы по API. Язык разработки — Golang. Используя Twitter API, можно извлекать и анализировать самую разнообразную информацию [8]. С помощью API Twitter был собран массив твиттов, которые пользователи публиковали на расстоянии в 10 км от заданной точки, в заданные даты и на русском языке.

Для площадок Яндекс-районы, TripAdvisor, Пикабу и других форумов на платформах Invision Power Board или phpBB3 применялся метод парсинга контента страниц, т. к. данные сервисы не имеют публичного API.

3.1. Формирование базы данных

Для определения входящих и исходящих данных разрабатываемой системы были подготовлены и проанализированы тестовые выборки данных, а также выбран предварительный формат их хранения. Используя инструмент проектирования баз данных Toad Data Modeler, была разработана логическая модель базы данных (сокр. БД).

Модель БД содержит 8 сущностей и отражает следующие данные:

- текст сообщения;
- ник автора;
- ссылка/адрес ресурса;
- объект городской среды;
- опорные слова для анализа оценки объекта;
- типы оценки;

- типы объектов городской среды;
- геолокационные данные.

После этого в средствах разработки баз данных — ERwin Data Modeler и pgModeler — была разработана физическая модель базы данных, отражающая её структуру в СУБД PostgreSQL.

4. Методология и методы исследования

Разработанный классификатор объектов городского хозяйства представляет собой веб-сервис с REST API-методами загрузки, выгрузки, машинного обучения, обработки естественного языка и статистического анализа данных.

4.1. Программные средства реализации алгоритма

Разработка алгоритма и моделей для автоматического классификатора велась на языке программирования *Python (3.6.5)* с использованием открытых библиотек для проведения различных исследований в области машинного обучения и анализа данных.

Для реализации решения задачи классификации объектов был использован метод, основанный на рекуррентных нейронных сетях с долгой-краткосрочной памятью (RNN + LSTM), т. к. текст, с которым нам предстояло работать, достаточно короткий и не содержит большое число отличительных свойств или признаков. Рекуррентные нейронные сети хорошо справляются с подобным типом задач, т. к. способны корректировать собственный результат на основе предыдущих [9, 10]. Для реализации алгоритма использована библиотека *TensorFlow* в качестве бэкенда и ядра вычислений, в качестве верхнеуровневой надстройки использован фреймворк *Keras*.

Для реализации решения задачи кластеризации был использован метод KMeans (метод k-средних). Действие алгоритма заключается в задаче минимизации суммарного квадратичного отклонения точек кластеров от центров самих этих кластеров. Для программной реализации алгоритма использован класс *Kmeans* из библиотеки *sklearn.cluster.* Кластеризация применена для автоматического поиска и удаления в текстовых данных спама, рекламы.

На подготовительном этапе запуска обучения алгоритма собранные данные были предварительно обработаны (удалены знаки пунктуации, убраны «шумы» и т. д.) и представлены в векторном или числовом виде. Для этого были использованы основные методы и подходы по обработке естественного языка.

Для извлечения и импорта собранного массива отзывов пользователей в программу использована библиотека *pandas.* В качестве обучающих текстовых данных был выбран предварительно размеченный

группой исследователей текст отзывов пользователей, в котором выделено 6 типов объектов:

- здание;
- двор;
- дорога;
- территория озеленения;
- малая архитектурная форма;
- водный объект.

Полученные данные были обработаны следующим образом: с помощью регулярных выражений (библиотека *re*) удалены знаки пунктуации, невидимые символы, латинские буквы, одиночные буквы, убраны лишние пробелы и табуляция.

С помощью библиотеки *pymorphy2* все слова приведены к начальной форме в соответствии с правилами русского языка (например, прилагательное слово «электронную» приводится к форме «электронный», т. е. прилагательное в единственном числе в мужском роде). Данный подход позволил снизить размерность массива данных без потери существенных признаков в тексте.

Преобразование текста в векторный (числовой) вид проведено с помощью класса *Tokenizer* из фреймворка *Keras*. Данный класс преобразует текст в векторный вид с помощью составления матрицы весов каждого слова на основе подхода вычислений функции TF-IDF.

Затем обработанные данные были переданы в алгоритмы классов *LSTM* и *KMeans* для решения задачи классификации и кластеризации соответственно.

Для решения задачи автоматического определения типа тональности текста (позитивный, нейтральный или негативный оттенок) было предложено использовать готовый размеченный русскоязычный массив текста, размещенный в открытом доступе для исследований. В качестве такого массива был взят массив RuSentiment (https://github.com/text-machine-lab/rusentiment) [11], содержащий 30521 постов из социальной сети «ВКонтакте». Русскоязычный массив размечен на следующие типы тональностей:

- positive (позитивный оттенок);
- negative (негативный оттенок);
- neutral (нейтральный оттенок);
- skip (пропущенные значения, т. е. значения без четко определенного типа тональности или тексты, которые содержат признаки художественного стиля: стихи, проза, анекдоты, афоризмы и т. д.);
- speech (текст содержит большое количество речевых клише, например приветствия, поздравления и т. д.).

Для автоматического определения тональности собранных нами текстов использована сторонняя открытая библиотека *dostoevsky*, распространяемая под лицензией MIT. В данной библиотеке находится обученная на русскоязычном массиве RuSentiment модель, показавшая точность определения тональности текста около 71 %, что является неплохим показателем в области решения задач анализа тональности.

4.2. Автоматическая классификация типов объектов

Для решения задачи классификации типов объектов была сформирована нейронная модель, представляющая класс *LSTM*, рекуррентную нейронную сеть с LSTM-блоком (RNN + LSTM), т. к. данный тип архитектуры достаточно успешно справляется с задачей классификации коротких текстов.

Модель состоит из входного слоя, слоя свертки данных в нужную размерность, LSTM-блока рекуррентной нейронной сети, полно связного слоя с 6 выходами (соответствует количеству типов объекта) с функцией активации «softmax» (для корректной работы многомерной классификации).

Данная модель была обучена на тренировочной выборке (80 % от объема основной выборки), и проверена на каждой итерации обучения на тестовой (20 % от объема основной выборки). Размер всей выборки составил 1864 записи, и в дальнейшем это количество будет увеличено, т. к. такое относительно малое количество не позволяет с уверенностью сказать о репрезентативности собранного набора. Обучающая выборка была размечена следующим образом: на основе анализа текста пользователя в дополнительный столбец «Тип объекта» был записан один из 6-ти типов, в случае если для всего текстового высказывания можно определить единственный точный тип объект. Если по тексту пользователя невозможно точно определить тип объекта или текст не относится к тематике городских объектов, то данному тексту присваивалась метка «не определено», и такие записи не участвовали в формировании основной выборки для обучения и тестирования модели.

5. Апробация алгоритма

С помощью разработанного алгоритма и подготовленных моделей был обработан массив текстовых данных, полученный из социальных сетей и форумных площадок для двух административных районов Санкт-Петербурга — Кронштадтского и Петроградского.

5.1. Кронштадтский район

С помощью разработанного инструментария был осуществлен анализ собранных текстовых данных отзывов и комментариев пользователей об объектах городской среды в Кронштадтском районе.

Всего было обработано и проанализировано 4935 записей. На рис. 1 представлено распределение записей по 6 типам объектов.

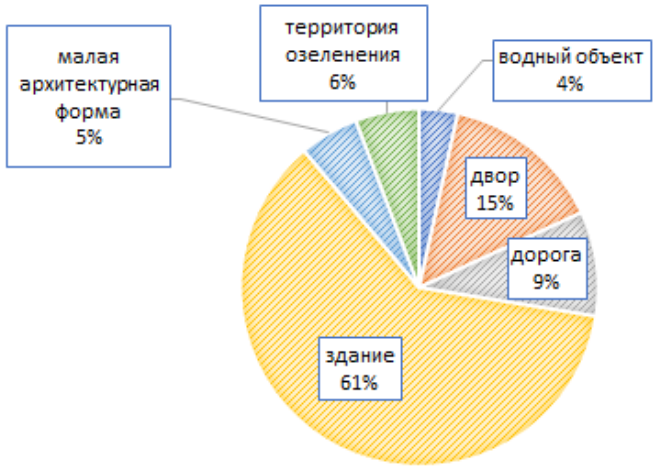

Рис.1 – Распределение записей по типам объектов (Кронштадтский район)

Наиболее частым типом объекта, по которому писали пользователи, оказалось «Здание» (3025, 61 %). Популярность данного типа объясняется тем, что большинство достопримечательностей (храмы, соборы, усадьбы, особняки и т. д.) или другие востребованные у жителей объекты городской среды действительно имеют тип «Здание». Однако было замечено, что данный тип часто выбирался и по тому, что эта категория наиболее подробно представлена в обучающей выборке и имеет существенный перевес в количестве данных, что в дальнейшем требует балансировки. Такая же ситуация наблюдается и для наименее востребованных типов («водный объект», 171, 4 %). Поэтому необходимо в дальнейшем привести набор данных к сбалансированному виду и дополнить новым количеством репрезентативных данных для подтверждения полученных результатов.

На рис. 2 представлено распределение записей по 5-ти типам тональности высказывания.

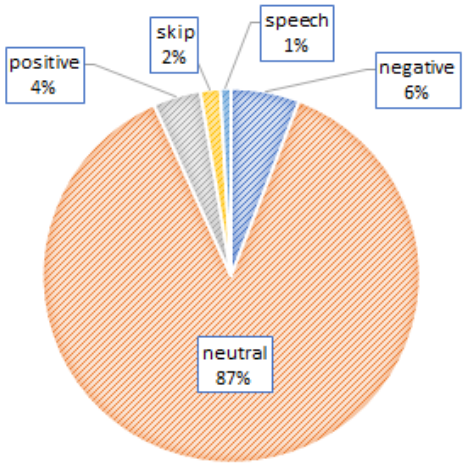

Рис. 2 – Распределение записей по типам тональности
(Кронштадтский район)

Превалирование типа «neutral» (4316, 87 %) обуславливается тем, что в обучающую выборку в основном попали комментарии, содержащую рекламу или фразу из одного слова. Также стоит отметить, что определение типа тональности происходило автоматически с помощью готовой модели из библиотеки *dostoevsky*, поэтому в перспективе необходимо рассмотреть случай без использования данной модели, или построение модели с использованием собственной ручной разметки текстового набора. Исходя из этого требуется более качественная настройка алгоритма кластеризации по поиску спама и рекламы, а также реализация собственной модели определения тональности. Отметим также, что общее количество высказываний с негативным оттенком (6 %) превалируют над высказываниями с положительным оттенком (4 %).

При детальном рассмотрении разреза типов объектов и типов тональности также подтвердилось превалирование нейтральных оттенков в высказываниях по каждому типу, а также перевес в сторону негативных оттенков над позитивными практически по всем типам (в среднем на 35 % больше негативных).

5.2. Петроградский район

Также был проведен анализ собранных данных высказываний по Петроградскому району.

Всего было обработано и проанализировано 17228 записей. На рис. 3 представлено распределение записей по 6 типам объектов.

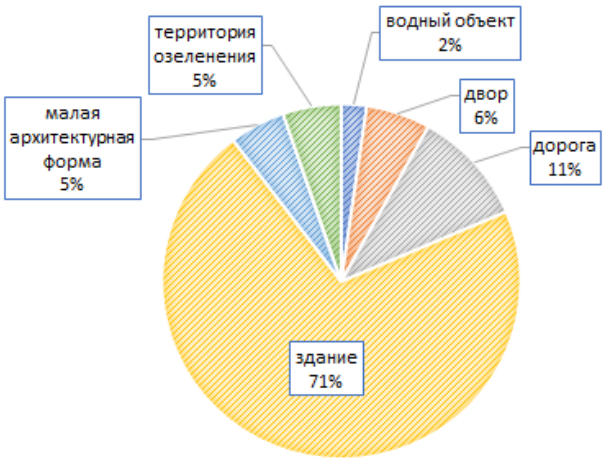

Рис. 3 – Распределение записей по типам объектов (Петроградский район)

На основе анализа также, как и в Кронштадтском районе, наибольшую востребованность имеет тип «Здание» (12250, 71 %). Это связано с обилием достопримечательностей в Петроградском районе и плотностью застройки, а также наибольшей представленностью данной категории в наборе обучающих данных. Наименее востребованным у пользователей оказался тип «водный объект». По нему было определено 393 записи (2 %).

На рис. 4 представлено распределение записей по типам тональности высказываний.

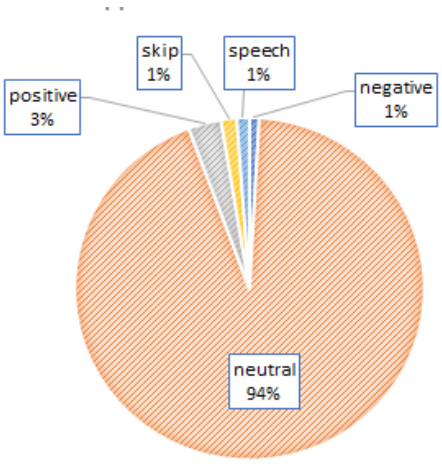

Рис. 4 – Распределение записей по типам тональности (Петроградский район)

Было выявлено, что и для высказываний, касающихся Петроградского района, большинство записей имеет нейтральный оттенок (16100, 94 %), что также создает предпосылки к гипотезе о нерелевантной работе готовой модели из библиотеки *dostoevsky*. Высказывания, содержащие обычные речевые фразы и приветствия (1 %), а также

имеющие негативный оттенок (1 %), встречаются довольно редко. Примечательно, что высказывания с положительным оттенком для Петроградского района (3 %) превалируют над высказываниями с негативным оттенком.

При детальном рассмотрении в разрезе типов объектов и типов тональности было подтверждено, что по большей части превалируют позитивные высказывания (особенно в категориях «Здание» и «Дорога»). Однако в категориях «Водный объект» и «Малая архитектурная форма» примерно на 50 % больше встретилось негативных высказываний, чем позитивных.

5.3. Сервис автоматической классификации текстов

Промежуточным итогом данного исследования стала разработка прототипа веб-сервиса, предоставляющего функции загрузки, выгрузки, анализа, визуального представления данных, автоматической классификации текстов по типам объектов, автоматического определения тональности текста на основе принципов REST.

Сервис позволяет произвести загрузку массива данных с постами из социальных сетей и произвести их обработку и анализ. Результатом работы методов является модифицированный загруженный файл с 4-мя дополнительными столбцами: «Тип объекта», «Вероятность типа объекта», «Тональность», «Вероятность тональности».

На рис. 5 представлена диаграмма действий в нотации UML, раскрывающая функциональные особенности сервиса.

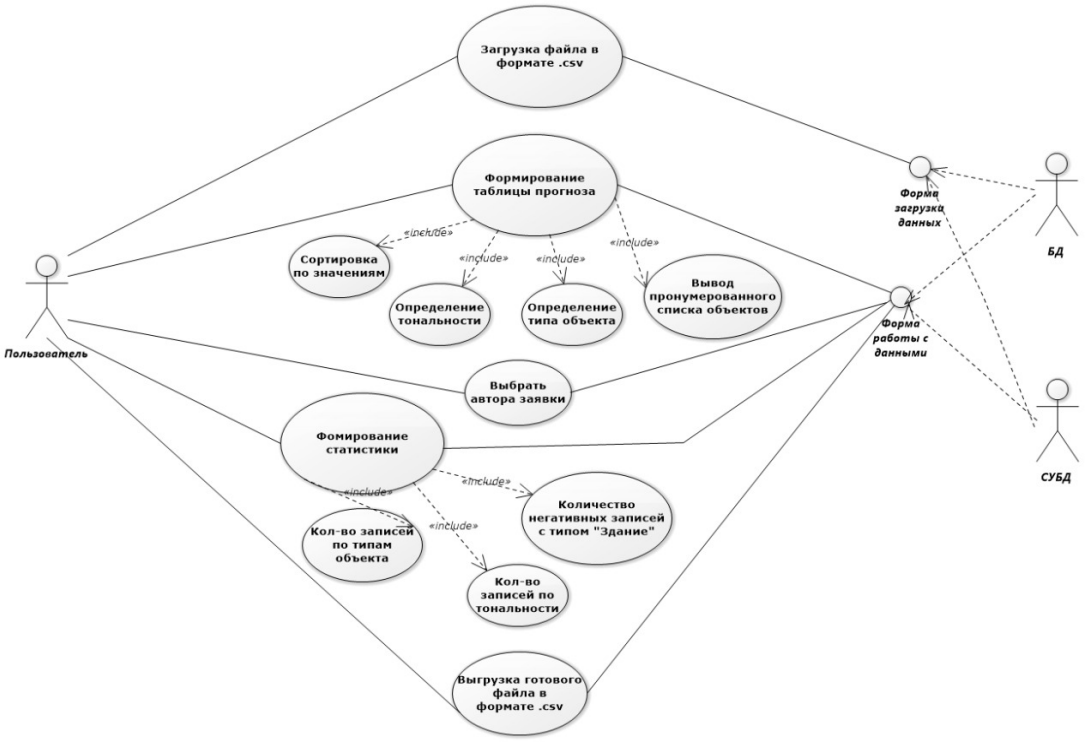

Рис.5 – Диаграмма действий

Система позволяет выгрузить (скачать) готовые обработанные данные в формате .csv. Содержимое файла возможно увидеть с помощью предварительного просмотра в веб-интерфейсе, представленного в виде таблицы с данными.

Разработка системы анализа данных велась на языке программирования Python (версия 3.6.5), с использованием фреймворка для веб-разработки Django (версия 3.0.4), для реализации веб-интерфейса, и дополнительных подключаемых библиотек. В качестве базы данных и СУБД использовалась PostgreSQL.

6. Выводы и направления дальнейших исследований

В ходе исследования был разработан основанный на методах машинного обучения алгоритм автоматического определения одного из шести типов объектов городского хозяйства и одного из пяти типов тональности для высказываний (с использованием готовой модели из сторонней библиотеки *dostoevsky*), представленных в текстовом виде и полученных из данных различных социальных площадок. Данный алгоритм заложен в разработку веб-сервиса, предоставляющего функции загрузки, выгрузки, анализа, визуального представления данных, автоматической классификации текстов по типам объектов, автоматического определения типа тональности текста. Сервис позволяет проводить исследования районов в режиме реального времени, основываясь на оставляемых пользователями данных в социальных сетях или форумах, что положительным образом сказывается на скорости получения и обработки результатов.

В качестве дальнейших этапов планируется повышение точности моделей (классификации, кластеризации, определения типа тональности), расширение категорий объектов. Запланирована доработка сервиса с возможностью дополнительного обучения на основе загруженных массивов (т. е. реализуется принцип обучения с частичным привлечением учителя, англ. semi-supervised learning). Также будет расширена база данных сообщений за счет использования API новых площадок.

Литература

1. Доклад о лучшей практике развития инициативного бюджетирования в субъектах Российской Федерации и муниципальных образованиях – URL: https://www.minfin.ru/common/upload/library/2019/10/main/1070_Doklad.pdf (дата обращения: 01.04.2020).
2. Vapnik V. N. An Overview of Statistical Learning Theory // Neural Networks, IEEE Transactions on. – 1999. – Vol. 10. – № 5. – P. 988–999.
3. Goldberg Y. Neural Network Methods in Natural Language Processing. – Morgan & Claypool Publishers, 2017. – 310 p.

4. Goldberg Y. A Primer on Neural Network Models for Natural Language Processing // Journal of Artificial Intelligence Research – 2016. – № 57. – P. 345–420.
5. Батура Т. В. Методы автоматической классификации текстов // Программные продукты и системы / Software & Systems. – 2017. – Т. 30. – № 1. – С. 85–99. – DOI: 10.15827/0236-235X.117.085-099
6. Зиберт А. О., Хрусталев В. И. Разработка системы определения наличия заимствований в работах студентов высших учебных заведений. Методы предварительной обработки текста // Universum: Технические науки: электрон. научн. журн. – 2014. – №4 (5). – URL: http://7universum.com/ru/tech/archive/item/1258 (дата обращения: 10.04.2020).
7. Общая информация о Петроградском районе. – URL: https://www.gov.spb.ru/gov/terr/reg_petrograd/information/ (дата обращения: 25.04.2020).
8. Docs - Twitter Developer. – URL: https://developer.twitter.com/en/docs (дата обращения: 10.01.2020).
9. Colas F., Brazdil P. Comparison of SVM and Some Older Classification Algorithms in Text Classification Tasks // IFIP AI 2006: Artificial Intelligence in Theory and Practice / M. Bramer (eds.). – 2006. – Vol. 217. – P. 169–178. – DOI: 10.1007/978-0-387-34747-9_18
10. Prasanna P. L., Rao D. R. Text classification using artificial neural networks // International Journal of Engineering & Technology. – 2018. – Vol. 7. – No. 1.1. – P. 603–606. – DOI: 10.14419/ijet.v7i1.1.10785
11. Rogers A., Romanov A., Rumshisky A., Volkova S., Gronas M., Gribov A. RuSentiment: An Enriched Sentiment Analysis Dataset for Social Media in Russian // In Proceedings of COLING 2018. – 2018. – P. 755–763.