Analysis and Classification of Urban Facilities Problems Based on Comments in Social Networks. Case for Saint Petersburg

Petr Begen[1] [0000-0002-0613-3133], Boris Nizomutdinov[1] [0000-0002-4090-9564], Aleksandr Tropnikov[1] [0000-0002-4179-536X]

[1] ITMO University, Kronverksky Pr. 49, 197101, Saint-Petersburg, Russia
petyabegen@mail.ru

Abstract. This paper describes an approach for conducting research on messages from users of urban communities in social networks using machine learning methods. This approach is proposed to be used to improve the efficiency of local government management, considering opinions of citizens who post their messages about urban facilities in social networks. The article presents the results of a pilot study, collected information (user reviews) from different communities, then, using machine learning methods, it was determined which object of urban economy is in issue, and then the sentiment type of such reviews was determined. The considered approach is a complex scenario for the classification of requests. The proposed classification approach has an accuracy of 78 %. The use of the proposed method is shown on the example of 2 administrative districts in Saint Petersburg – Petrogradsky and Kronshtadtsky.

Keywords: urban economy, artificial intelligence, machine learning, classification, clustering, social networks, sentiment analysis, neural networks, natural language processing.

1 Introduction

The main trend of modern municipal management is a participatory design. With this approach city residents are involved in planning the urban environment and designing improvement facilities. However, as a rule, citizens are not willing to take part in such events, especially given the current epidemiological situation. However, residents of the city are willing to post their comments on social networks, namely, in urban communities or groups of their home, area or district. This data is very interesting information, because it is an informal appeal of citizens to the authorities, it is possible to understand what are the needs of the population, what expectations and requirements for the surrounding urban space do they have.

Understanding what is written about the city in social networks will help to improve the efficiency of municipal management, will allow us to establish an information exchange between the population and the authorities. Criticism and negative reviews can help to create targeted improvement programs.

Currently, user engagement in social networks continues to grow, while the scope of application is expanding: citizens use communities to solve everyday problems, discuss urban management, improvement of their district or yard, and participate in participatory budgeting [1]. At the same time, the institution of public communications and channels of interaction with the authorities are being improved.

In online communities, a huge number of messages are generated, and this amount of information might seem weakly structured and not suitable for research and further use, however, there are approaches that allow all posts to be classified in an automated way and determine which object is in question.

The article presents the results of a study of messages and statements related to the urban environment and public spaces on the example of St. Petersburg. The task of the research team was to automatically determine which urban object is mentioned in the message or comment based on the available data set, which is user messages and collected on different sites or social networks, as well as to automatically determine the user's attitude to this type of object (positive, neutral or negative).

The article offers an implementation of this task using machine learning methods. It focuses on the use of tools for automated information collection. This complex is suitable for studying individual urban communities (groups), as well as, if the number of groups increases, for studying the situation in the city, district, or region.

The results obtained can be used at the municipal level of government in the implementation of various urban improvement projects, as well as to track the current state of urban and courtyard infrastructure, which will reduce the cost of additional information collection activities. For this purpose, it is necessary to solve the problem of collecting feedback in social network communities, then classify all messages into categories in order to select from the entire array of information exactly the one that is required for a specific improvement project. In addition, the classifier will allow you to assess the overall state of urban infrastructure in a particular area, which will ultimately reduce the cost of monitoring and reduce the decision-making time to eliminate the identified problems.

2 Software implementation of the classifier

To implement an automatic classifier of urban objects based on data from social networks, the following tasks were set:

- to automatically identify an object in the text of a user's review or comment, it is necessary to develop an algorithm for solving problems of text classification and clustering for subsequent comparison by experts of the results of a classifier using manual and / or automatic (machine) markup;
- to automatically determine the type of sentiment (positive, neutral, or negative) of the review or comment text, select a ready-made and public Russian-language text

array with the specified markup for positive and negative text groups, and use this array to prepare a model for automatically determining the sentiment type of the review text.

To solve the problem of automatic object type detection in text data, an algorithm was developed for solving 2 classic machine learning problems: text classification and clustering.

Classification is the division of a set of documents into pre-known groups based on some parameters or properties.

Clustering is based on splitting a set of similar documents into clusters or subsets whose parameters are unknown in advance. The number of clusters can be arbitrary (i.e. the algorithm determines the required number itself) or fixed (i.e. set by the user at the initial stage of the algorithm implementation).

Machine learning is a broad section of the field of artificial intelligence that studies methods for building self-learning algorithms. In his article [2], V.N. Vapnik was one of the first to consider "the theory of statistical learning as one of the possible options for the development and operation of machine learning", which allowed to give a clearer statement of the problem of machine learning as a whole and form a scheme of the mechanisms of various algorithms for years to come.

Yoav Goldberg in his works [3, 4], T.V. Batura in a review study of automatic text classification methods [5] and A.O. Zibert and V.I. Khrustalev in [6] present the main methods and approaches in the field of Natural Language Processing (NLP) in the framework of working with neural networks of different architectures and with standard statistical models for implementing deep learning methods, as well as the main results of testing, experiments and indicators obtained during the implementation of various methods.

The main methods and approaches of Natural Language Processing include the following:

- Tokenization or segmentation.
- Creating and using a stop word list.
- Stemming.
- Lemmatization.
- Named Entity Recognition.
- Bag-of-words model.
- Calculation of the TF-IDF function, dimension reduction.
- Word2Vec algorithms.

The presented methods and approaches for Natural Language Processing formed the basis of the developed solution for analyzing text data from social networks.

3 Collected data

For the pilot study two districts, Petrogradsky and Kronshtadtsky, were selected in St. Petersburg with different administrative characteristics. The main parameters of both

districts: the population of Kronshtadt is 43,687 people (2017), the island area is 1,584 ha, the total territory of the Kronshtadt district within the existing approved borders is 1,935 ha. Petrogradsky district: the geographical area of the district is 24 square km; the population is 131,356 people [7].

The following platforms were selected as sites for the study:

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

These sources cover different groups of residents, which makes the sample more representative. It is important to keep in mind that only the text of requests was collected, and all user data was depersonalized and not saved.

The largest area that contains the urban communities of "VKontakte" or "VK" (vk.com). It features many groups where residents of a city or a particular area discuss infrastructure or landscaping in an informal setting. These communities are gaining popularity as more and more residents want a comfortable urban environment. In the pilot study, information was collected by parsing HTML content, without using the API, and then parsing the content with a special script that selected only text from the markup. Only basic information of interest was saved for the specified criteria. To collect the main parameters of a post (user's message) on the community wall (VK groups), the parsing boundaries for each parameter were set using the example of a single message. The source code of the page was analyzed and HTML tags containing the necessary parameters were highlighted. This method has several limitations: 1) time to collect information; 2) the inability to track the dynamics in real time; 3) work only with a ready-made database. During parsing, more than 200,000 records were collected.

The next step after this work is planned to study and collect data using available methods via the VK API. This will increase performance and improve the accuracy of the collected data.

The VK API is an external interface that allows you to get information from the database vk.com using HTTP requests to a special server. The query syntax and the type of data returned are strictly defined on the service side. VK is a social network that has friendships, privacy settings, and blacklists. A lot depends on who is viewing the page: someone will see all the same information on it as the owner, and someone will see only publicly available data. This principle is also preserved in the API. The VK API features are described here: https://vk.com/dev/manuals.

Text messages from Twitter were collected using the standard Twitter API. Twitter provides access to some of its services using APIs so that developers can create services that use data from Twitter. the method provides a set of ready-made classes, functions, or structures for working with existing data. To work, you had to register the app, go through the verification procedure, and get the necessary keys. Given that any tweet is a message of no more than 280 characters, users put the most significant information

in the text, which is well suited for classification. The research team prepared a parser in the Golang language to extract and save information

Using the Twitter API, you can extract and analyze a wide variety of information [8]. The Twitter API was used to collect an array of tweets that users posted at a distance of 10 km from a given point, on specified dates, and in Russian.

Text information from Yandex.Market, TripAdvisor, Peekaboo, and other forums on Invision Power Board platforms was collected using special parsing software, as in the case of the VK social network, since these services do not have an external API. All collected information was depersonalized, only the text of requests is collected, without the authors data.

3.1 Database creation

To store a large amount of collected information (text messages, posts, comments), it was necessary to determine the storage format and create a simple database. In the first step, a logical database model was prepared using the toad Data Modeler database design tool. The designed model has 7 entities and reflects the following data:

- Message text.
- Link/address of a resource.
- The object of the urban environment.
- Reference words for object evaluation analysis.
- Type of evaluation.
- Types of urban environment objects.
- Geolocation data.

Then, using the software for creating databases (ERwin Data Modeler and pgModeler), a physical database model was created that has the structure of the PostgreSQL DBMS.

4 Automatic text classification service developing toolset

The intermediate result of this research was the development of a web service prototype that provides API methods based on REST principles for loading, unloading, analyzing, visual representation of data, automatic text classification by object types, and automatic detection of text sentiment types. Methods include basis of Machine Learning, Natural Language Processing, data analysis, statistics.

The service allows to upload an array of data with posts from social networks and perform their processing and analysis. The service result is a modified uploaded file with 4 additional columns: “Object type”, “Probability of object type”, “Sentiment type”, “Probability of sentiment type”.

The service also provides uploading (downloading) ready-made processed data in “.csv” format, which is comfortable for future analyses and works. The file contents can be viewed using a preview in the web interface, presented as a table with data.

The data analysis service was developed with Python, version 3.6.5, also using web development framework Django, version 3.0.4, for implementing the web interface, and additional plug-in libraries. PostgreSQL was used as the main database.

At the beginning stage of implementing the algorithm training, the collected data was pre-processed, i.e. punctuation, numbers and some "noise" were removed, and presented in a vector or numeric form. For this purpose, Natural Language Processing methods and approaches were used.

The "pandas" library is used to extract and import the collected array of user reviews into the program code. As training data, the text of user reviews was previously marked up by a group of researchers with 6 objects, which are following:

1. Building.
2. Yard.
3. Road.
4. Landscaping area.
5. Architectural landscape element.
6. Water object.

The received data was pre-processed as follows: using regular expressions from "re" library punctuation marks, invisible characters, Latin letters, single letters were removed, and extra spaces and tabs were also eliminated.

Using "pymorphy2" library all words were reduced to their initial form in accordance with the rules of Russian language. This approach allowed to reduce the size of the data array without losing significant features in the text.

Text was converted to a vector (numeric) form using Tokenizer class from the Keras framework. This class converts text to a vector form by creating a sparse matrix of weights for each word based on TF-IDF function calculation approach.

The processed data was then passed to the algorithms of LSTM and KMeans classes to solve the classification and clustering problems.

To solve the problem of automatically determining the type of text sentiment (i.e. positive, neutral, or negative), it was proposed to use a ready-made marked-up Russian-language text array placed in the public domain for research. The RuSentiment array was used as such an array [11], containing 30521 posts from the popular in Russia social network "VK". This Russian-language array was divided into the following types of keys:

- Positive.
- Negative.
- Neutral.
- Skip (omitted values, i.e. values without a clearly defined type of sentiment or texts that contain features of an artistic style: poems, prose, anecdotes, aphorisms etc.
- Speech (the text contains many speeches cliches, such as greetings etc.

The third-party open library "dostoevsky", which is distributed under the MIT license, is used to automatically determine the type of text sentiment. This library contains a

model trained on the Russian-language array RuSentiment, which showed a determination accuracy of the text sentiment types of about 71 %, which is quite a good indicator for sentiment analysis.

To solve the problem of object classification we used a method based on Recurrent Neural Network with long-short-term memory (RNN+LSTM), since the text we had to work with was quite short and did not contain a large number of distinctive properties or features. Recurrent Neural Networks cope well with this type of task because they can correct their own results based on the previous ones [9, 10]. TensorFlow library is used as a backend and computing core, and the Keras framework is used as a top-level add-in above TensorFlow for the whole service algorithm implementation.

The model based on RNN+LSTM consists of an input layer, a data convolution layer to the desired dimension, an LSTM block of a Recurrent Neural Network, and a fully connected layer with 6 outputs corresponding to the number of object types, with the "softmax" activation function (i.e. for correct operation of multidimensional classification).

Kmeans method was used for clustering. The goal of the algorithm is to minimize the total square deviation of cluster points from the centers of these clusters themselves. Kmeans class from "sklearn.cluster" library is used for software implementation of the algorithm. Clustering is used for automatic search and removal of spam and ads in text data.

This model was trained on a training sample and tested on each training iteration on a test sample with ratio 80/20. The size of the entire sample was 1,864 records. This number will be increased for future works, since such a relatively small number does not allow to say confidently about the representativeness of the collected set. The training sample was marked up as follows: based on the analysis of the user's text, one of 6 types was recorded in the additional column "Object type" in data file, if the only exact object type can be determined for the entire text statement. If the user's text cannot accurately determine the object type or the text does not relate to the subject of urban objects, then this text was assigned the label "undefined", and such records did not participate in the main sample formation for training and testing the model.

5 Testing the text classification service

The developed algorithm and prepared models of service were used to process an array of text data obtained from social networks and forums in a case for two St. Petersburg administrative districts: Kronshtadtsky and Petrogradsky districts.

5.1 Kronshtadtsky district

The developed tools were used to analyze the collected text data of user reviews and comments about urban objects located in the Kronshtadtsky district.

A total of 4,935 records were processed and analyzed. The distribution of records by six object types is presented on Fig.1.

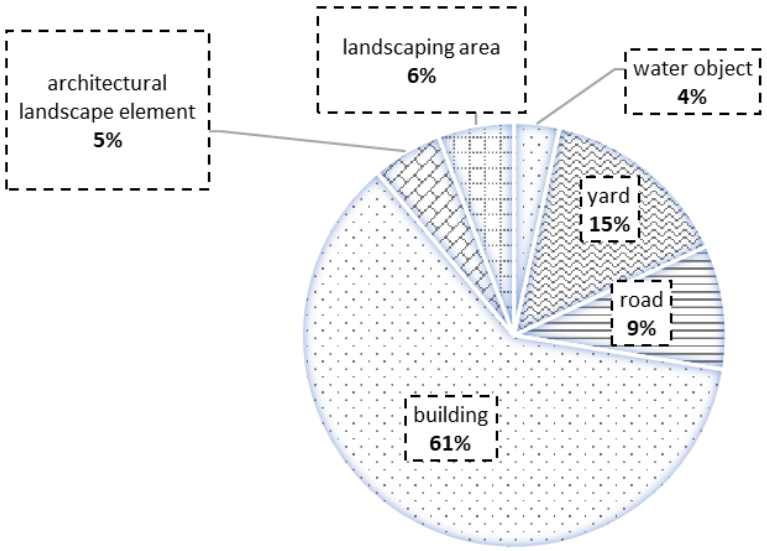

Fig. 1. Percentage by object types in the Kronshtadtsky district

The most frequent item type that users wrote about or mentioned was “Building” (3,025 – 61 %). The popularity of this type is explained by the fact that most people attractions (e.g. temples, cathedrals, mansions etc.) or other urban objects that are in demand among residents have literally the “Building” type. However, this type was often chosen because it is most detailed in the training sample and has a significant advantage in the amount of data, which requires balancing in future works. The same situation is observed for the least popular types (“Water object”, 171 – 4%). Therefore, it is necessary to bring the data set to a balanced form in the future and add a new amount of representative data to confirm the results.

The Figure 2 shows the distribution of values by 5 types of text sentiment.

The prevalence of the “neutral” type (4,316 – 87 %) is since the training sample mostly includes comments containing an ad or a single-word phrase. It is also worth noting that the key type was determined automatically using a ready-made model from the “dostoevsky” library, so in future works it is necessary to consider the case without using this model, or building a model using our own manual markup of the text set. Based on this, we need to better configure the clustering algorithm for spam and advertising search, as well as implement your own sentiment detection model. We also note that the total number of statements with a negative connotation (6 %) prevails over statements with a positive connotation (4 %).

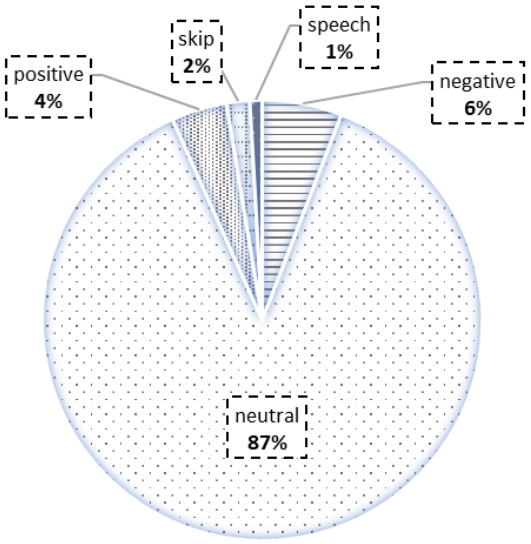

Fig. 2. Percentage by sentiment types in the Kronstadtsky district

A detailed examination of the section of object types and sentiment types also confirmed the prevalence of neutral shades in statements for each type, as well as the preponderance of negative shades over positive ones for almost all types (i.e. average 35% more than negative ones).

5.2 Petrogradsky district

An analysis of the collected statements data in Petrogradsky district was also carried out.

A total of 17,228 records were processed and analyzed. Figure 3 shows the distribution of records by six object types in this location area.

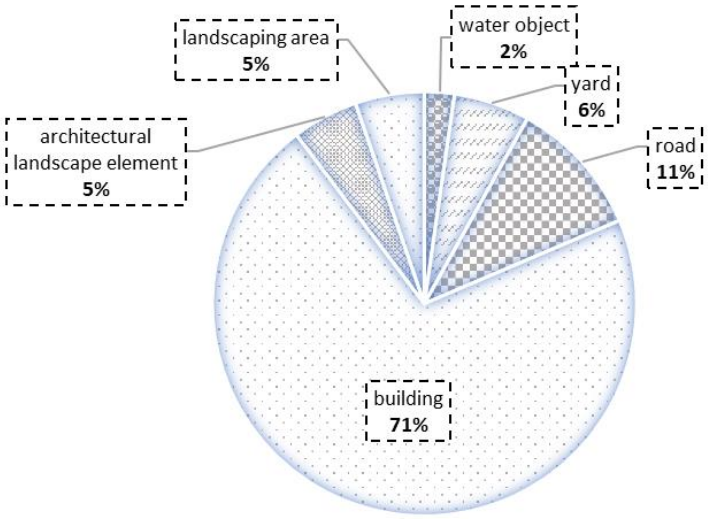

Fig. 3. Percentage by object types in the Petrogradsky district

Based on the analysis, as well as in the Kronshtadtsky district, the “Building” type has the highest demand (12,250 – 71 %). This is since abundance of attractions in the Petrogradsky district and the density of buildings, as well as the highest representation of this category in the training data set. The least popular among the users turned out to be a type of “Water object” again. The service identified 393 entries (2 %).

Figure 4 shows the distribution of values by types of text sentiment.

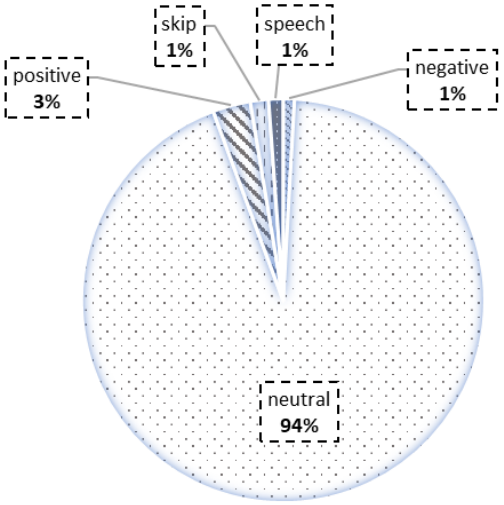

Fig. 4. Percentage by sentiment types in the Petrogradsky district

We found that for statements concerning the Petrogradsky district, the majority of values have a neutral shade (16,100 – 94 %), which also creates prerequisites for the hypothesis that the finished model from the "dostoevsky" library is irrelevant. Statements containing common speech phrases and greetings (1 %), as well as having a negative connotation (1 %), are quite rare. It is noteworthy that statements with a positive connotation for the Petrogradsky district (3 %) prevail over statements with a negative connotation.

A detailed analysis of objects types and types of text sentiment confirmed that positive statements mostly prevail (especially in two categories: "Building" and "Road"). However, in the categories "Water object" and "Architectural landscape element", there were about 50% more negative statements than positive ones.

6 Discussion and Conclusion

One of research outcomes is an algorithm based on Machine Learning methods and Natural Language Processing that was developed for automatically determining one of the six types of urban objects and one of five types of text sentiment using a ready-made model from the open third-party "dostoevsky" library, presented in text form and obtained from data from various social platforms and nets. This algorithm is embedded in the development of a web service that provides methods for loading, uploading, analyzing, visual representation of data, automatic classification of texts by object types, and automatic detection of the text sentiment types. The service allows to conduct real-time research of districts based on data left by users in social networks or forums, which has a positive effect on the speed of receiving and processing results for further analysis for city stakeholders: citizens, business, government, city management etc.

As further stages of developing service and improving city analysis outcomes, it is planned to improve the accuracy of models for tasks of classification, clustering, sentiment analysis, and expand the categories of objects. It is also planned to finalize the service with the possibility of additional training based on loaded arrays (i.e. the principle of training with partial involvement of a teacher, semi-supervised learning). The text arrays database will also be expanded by using API of new sites and platforms. These improvements would be a basis for developing the whole big system for the city authorities and citizens under the concept of "Smart City".

References

1. Doklad o luchshej praktike razvitiya iniciativnogo byudzhetirovaniya v sub'ektah Rossijskoj Federacii i municipal'nyh obrazovaniyah. https://www.minfin.ru/common/upload/library/2019/10/main/1070_Doklad.pdf, last accessed 2020/11/23.
2. Vapnik, V.N.: An Overview of Statistical Learning Theory. Neural Networks, IEEE Transactions on 10 (5), 988–999 (1999).

3. Goldberg, Y.: Neural Network Methods in Natural Language Processing. Morgan & Claypool Publishers (2017).
4. Goldberg, Y.: A Primer on Neural Network Models for Natural Language Processing. Journal of Artificial Intelligence Research 57, 345–420 (2016).
5. Batura, T.V.: Automatic text classification methods. Software & Systems 30 (1), 85–99 (2017). DOI: 10.15827/0236-235X.117.085-099
6. Zibert, A.O., Hrustalev, V.I.: Development of a system for determining the existence of adoption in the works of the students. Methods of preparation of automatic text processing. Universum: Tekhnicheskie nauki: elektron. nauchn. zhurn. 4 (5), (2014).
7. Common information about Petrogradsky district, https://www.gov.spb.ru/gov/terr/reg_petrograd/information/, last accessed 2020/11/23.
8. Docs – witter Developer, https://developer.twitter.com/en/docs, last accessed 2020/11/23.
9. Colas, F., Brazdil, P.: Comparison of SVM and Some Older Classification Algorithms in Text Classification Tasks. In: Bramer, M. (eds.) IFIP AI 2006: Artificial Intelligence in Theory and Practice, vol. 217, pp. 169–178 (2006). 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 7 (1.1), 603–606 (2018). 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, pp. 755–763 (2018).