The Method of Monitoring Incidents Based on Data from Social Networks: The Case of St. Petersburg

Boris Nizomutdinov[a], Petr Begen[a] and Daria Lipatova[a]

[a] *ITMO University, Kronverksky Pr. 49, Saint-Petersburg, 197101, Russia*

Abstract

The paper presents an approach to extracting the address and type of incident from a text data array formed based on posts in social networks. Data was uploaded from the Vkontakte community dedicated to incidents in St. Petersburg. A total of 48,943 records were collected and processed. A service has been developed for automatic recognition of the post topic and address extraction (if available) using natural language processing and machine learning methods using the free natasha library for Russian-language texts. Using the Geocoding API service from Google, the existing addresses were geocoded, and an array in GeoJSON format was obtained, which allows working with the dataset in various map services in real time.

Keywords

natural language processing, address extraction, parser, natasha, yargy, geocoding, GeoJSON.

1. Introduction

Currently, the interconnection of heterogeneous urban elements is supported primarily by information technologies that provide communication links between residents, the management sector and infrastructure. On the one hand, the modern information space allows residents to constantly observe the life of the city, to meet their needs and interests in a mobile way, and on the other hand, it creates high expectations in relation to the urban environment and increases the level of responsibility of city authorities.

Due to the lack of capacity to limit the impact of hazards, many cities still face a high level of threats. As threats to cities increase, improving the resilience of cities becomes a major challenge. To increase the sustainability of cities, there is a growing need for information that is relevant to all stages of urban development. Thus, a better understanding of the spatio-temporal patterns of public response is a key step towards reducing damage and improving the resilience of cities.

So, a team of researchers from New York University analyzed incident data from two different sources: from a traditional data provider that collects incident reports from multiple agencies, and user messages from Twitter during Hurricane Sandy, which flooded many areas of New York in 2012 [1]. The result showed that Twitter can provide detailed information about the location of a particular incident, as well as its intensity, duration, etc.

In recent years, interest in Twitter has been growing due to the fact that data in it is stored in real time. Microblogs are increasingly attracting attention as an important source of information in emergency management. Twitter is used as a way to predict accidents, natural disasters, and traffic. [2] For example, in China, local floods are studied using geo data from Twitter [3]. There are methods for monitoring the traffic situation in real time based on data in social networks using modern machine learning algorithms [4]. Metro passenger traffic forecasting is strategically important in the management of the metro transit system. Predicting the occurrence of events turns into a very difficult task, so today, forecasting in passenger transport is developing based on data from their social

IMS 2021 - International Conference "Internet and Modern Society", June 24-26, 2021, St. Petersburg, Russia
EMAIL: boris-wels@yandex.ru (A. 1); petyabegen@mail.ru (A. 2); sergaahsad@gmail.com (A. 3)
ORCID: 0000-0002-4090-9564 (A. 1); 0000-0002-0613-3133 (A. 2); 0000-0003-4500-3582 (A. 3)

networks [5]. In most cases, Twitter data is used [6-9], the use of the social network Vkontakte has not yet become widespread.

The purpose of this work is to study and analyze the use of methods for extracting addresses from Russian-language text messages about incidents in social networks Vkontakte to generate geospatial data in the GeoJSON format, which can be used later in GIS systems or in the hardware-software complex “Safe City”.

The result of this approach is the distribution of incidents in the city on the map since the text will consist of 2 entities: the address and the type of incident. At the first stage of the study, 5 main topics were selected: Car theft, Accident, Fire, Robbery and Assault. Using this approach, we can identify the most dangerous or problematic area in the city or find the area where the most theft occurs and so on [10]. By the way, this information is available in official sources, it is not always available to urbanists and researchers and does not always reflect the current state of the city. This approach can expand the data set for researchers and citizens.

In addition, the data obtained can be used in the hardware-software complex “Safe City”, since often the incident message does not arrive in the system immediately, and this approach allows you to generate information online. For example, information about an accident can be included in the statistics in a week or even a month, if the accident was registered according to the Euro protocol. But active citizens quickly highlight road accidents and publish them into this community, which allows to search for problem areas in the city online.

2. Development of an address extraction and incident recognition service

2.1. Data preparation

In the case of St. Petersburg, we considered the possibility of extracting data about incidents that citizens write about in the social network called Vkontakte. The community about road accidents and emergencies in St. Petersburg (https://vk.com/spb_today) was chosen as a site for the study.

The collection of information for the study was carried out using a set of tools that included the Vkontakte API, a content parser, and a public service. The method API Vkontakte “Wall.get” returns a list of posts from the user's wall or community, using this method you can collect all the comments in the community. When conducting research on social networks, there is a complex problem of personal data security during parsing. Personal data, according to art. 3 of the federal law "On Personal Data", is called " any information related directly or indirectly to a certain or identifiable individual (subject of personal data)". No personal information was collected or stored in this report. A total of 48,943 records were collected.

For the collected sample, preliminary automatic processing of posts was carried out, which consisted in deleting entries starting with the words “News of our metropolis:”. In this case, we did not consider the daily news reports when forming the final sample for analysis, because its only summed up the results of the day. The size of the final sample after deleting such posts was 48,408 entries.

2.2. Using natasha library and yargy parser for extracting address and incident topic from posts

One of the tasks of the research was also the development of tools for automatic processing and analysis of the data array of posts from social networks. The main functionality of the tool was the automatic detection of the topic of the post and recognition of the address or its component part: street, block, house number, district, etc. To get the primary results of the tool, it was decided to recognize five topics of posts: Car theft, Accident, Fire, Robbery and Assault.

To develop the toolkit, we used the open neural network library natasha (v.1.4.0), which was updated in 2020, for recognizing addresses in Russian-language text [11]. To recognize the one of 5 incident topics a yargy parser was configured. To extract user entities in yargy, special rules are created using context-free grammars and specialized dictionaries. As part of the research work, simple rules were added with ready-made parser predicates that recognize words for highlighting the topic:

“theft, stolen, stealing, car theft” – Car theft, or “accident, road accident, lapped, collided, crash” – Accident.

To recognize the topics of posts, we used self-written rules for the yargy parser. As a result of setting up the parser, a sample of these posts was obtained, which contained only five topics. The Figure 1 shows the ratio of posts with five selected topics to the total number of posts.

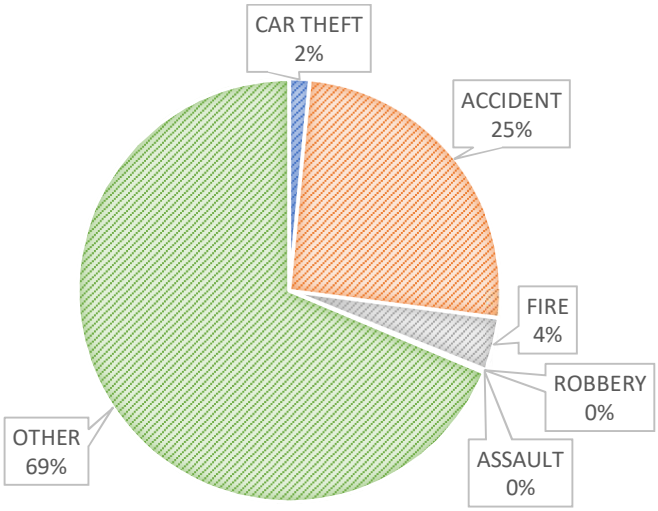


Figure 1: Distribution of posts by topic

Based on this sample in semi-automatic mode (SQL query + manual markup), the average accuracy of the yargy parser recognition of the five selected topics was calculated, the results are presented below in Table 1.

Table 1. Results of post topics recognition using yargy parser

Topic	Recognized	Not recognized	Total number	Accuracy
Car theft	694	61	755	91.92%
Accident	11 568	774	12 342	93.73%
Fire	1801	153	1954	92.17%
Robbery	25	6	31	80.65%
Assault	77	8	85	90.59%

The result of the average accuracy of determining the topics in the text using the configured yargy parser is good (more than 90% except for Robbery). Among the disadvantages of this approach, the long duration of the parser's operation time was highlighted, since which we can assume about the slowness of the algorithm itself, especially when increasing the data sample. Thus, the next stage will require optimization of the parsing rules and more efficient data processing.

To recognize addresses in the text, the built-in “AddrExtractor” function from the natasha library was used. Address recognition was performed on a sample of data, including posts with recognized topics. A total of 15,167 records were selected in the sample. To calculate the average recognition accuracy, the condition was set that if at least one part of the address is recognized in the text of the post (for example, street, name, house number, etc.), then the address is considered recognized. The results of address recognition are shown in Table 2.

Table 2. Results of address recognition using the natasha library

Entity	Recognized	Not recognized	Total number	Accuracy
Address	11 389	3 778	15 167	75,09%

The table shows that the result of the average accuracy of address recognition in the Russian text is satisfactory for solving the problem (more than 75%), but in the future, the algorithm and rules for address recognition also require improvements to improve the accuracy and quality of determining the recognized addresses. When recognizing the address, it was also noted that the highest percentage of accuracy is achieved when determining the name “Moscow” for the type “street”, for example, in the format “Moscow Street”. However, if you remove the marker word “street”, the recognition accuracy will significantly decrease, even if there are other markers nearby, such as “house number”, “building”, and other parts of the address.

As a result of the analysis, we can conclude that it is sufficiently possible to use open and free ready-made solutions that provide the functionality of flexible rule settings for performing tasks of this kind of analysis.

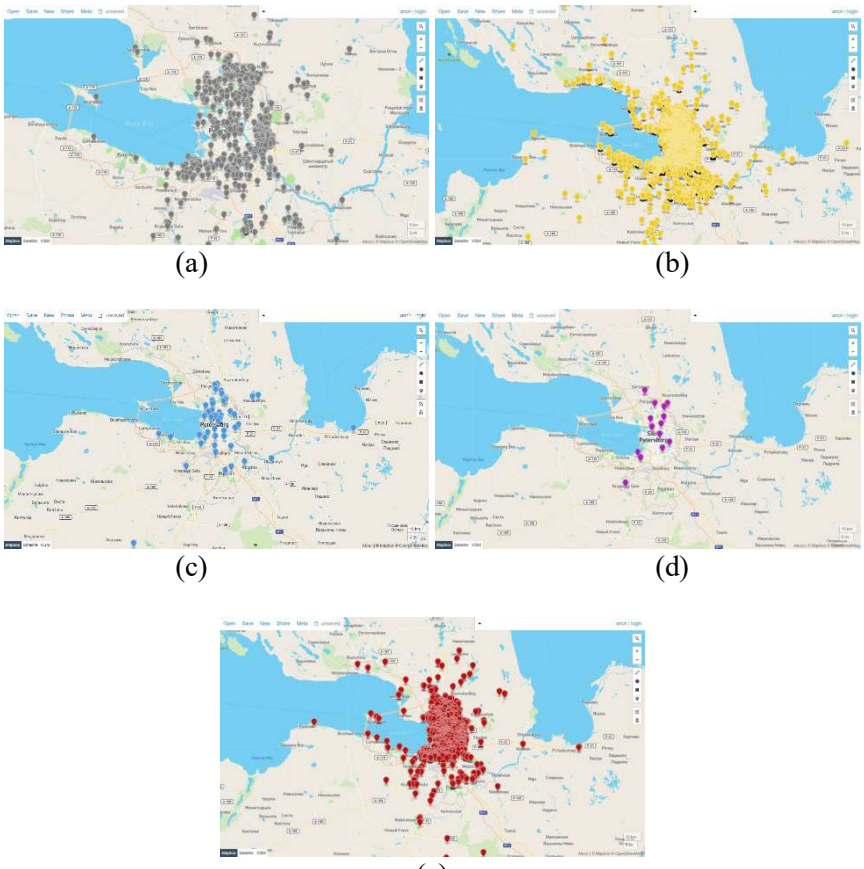
2.3. Representing incidents from the posts on a map

Geocoding recognized addresses and formatting dataset to GeoJSON was the next step of work. For this purpose, we used Geocoding API service from Google [12], that can convert address parts and return its coordinates. One of disadvantage of Geocoding API is that it returns standard JSON, so we also needed to convert it to GeoJSON format afterwards.

As we need to store and show topic on a map, we used the Feature and FeatureCollection objects according to GeoJSON specification. Here is the example of format we used:

{"features": [{"geometry": {"coordinates": [30.36091, 59.931058], "type": "Point"}, "properties": {"topic": "nan"}, "type": "Feature"}, {"geometry": {"coordinates": [30.516726, 59.73777], "type": "Point"}, "properties": {"topic": "Угон"}, "type": "Feature"}], "type": "FeatureCollection"}

The dataset of 15,1567 records with all recognized addresses was geocoded and converted to GeoJSON. Below in Figure 2 you can see the visualizations of points according to five incident topics.



(a) (b) (c) (d) (e)

Figure 2: Car theft (a); Accident (b); Assault (c); Robbery (d); Fire (e)

Meanwhile we discovered that not all the address parts were recognized properly, for instance there can be only house number without street location or only street, which can be long. Thus, for future steps of research work we need to increase recognition accuracy and quality of address parts extraction.

3. Conclusions and Discussion

The selected source of information in the social network Vkontakte showed that users generate a large amount of information about incidents. In Russia, Vkontakte is more popular than Twitter [13], which is why it is important for researchers to have tools to work with this social network. The method has shown its promise, and the data obtained can be used by both researchers and representatives of government departments.

During the research work, a toolkit was developed for automatic recognition of the post topic and addresses. The considered experiment confirmed the good effectiveness of the selected open library natasha (v.1.4.0) with a yargy parser, which managed to extract the topic and address from the text of the posts. With the help of the Geocoding API from Google, we managed to get the coordinates of addresses, translate the result of geocoding into the standard GeoJSON format, which allows us to use this data in different map services, GIS, as well as in the hardware-software complex “Safe City”. In the future, it is planned to increase and improve the data sample by using methods of automatic collection and extraction of entities, improving the accuracy of extraction and recognition of the posts topics and address parts.

4. Acknowledgement

The work was done under the research topic of ITMO University No. 620179 “Development of a map service for monitoring the residents needs in the urban infrastructure development using automated data processing systems from social networks”.

5. References

[1] A. Kurkcu, F. Zuo, J. Gao, E. Morgul, K. Ozbay, Crowdsourcing Incident Information for Disaster Response using Twitter, Transportation Research Board, 2017.

[2] A. Ammari, I. Petalas, Traffic Event Detection Framework Using Social Media, in: International Conference on Smart Grid and Smart Cities, 2017. doi:10.1109/ICSGSC.2017.8038595.

[3] B. Wang, B.P.Y. Looc, F. Zhene, G. Xie, Urban resilience from the lens of social media data: Responses to urban flooding in Nanjing, China, Cities, volume 106, 2020, pp. 1–13. URL: https://doi.org/10.1016/j.cities.2020.102884.

[4] A. Pathak, B. Patra, A. Chakraborty, A. Agarwal, A City Traffic Dashboard using Social Network Data, in: the 2nd IKDD Conference, 2015. doi:10.1145/2778865.2778873.

[5] M. Ni, Q. He, J. Gao, Forecasting the Subway Passenger Flow Under Event Occurrences With Social Media, IEEE Transactions on Intelligent Transportation Systems, 2016, PP(99):1–10. doi:10.1109/TITS.2016.2611644.

[6] Y. Gua, Z. Qiana, F. Chenb, From Twitter to detector: Real-time traffic incident detection using social media data, Transportation Research Part C: Emerging Technologies, volume 67, 2016, pp. 321–342. URL: https://doi.org/10.1016/j.trc.2016.02.011.

[7] B. Hawelka, I. Sitko, E. Beinat, S. Sobolevsky, Geo-Located Twitter as Proxy for Global Mobility Patterns, Cartography and Geographic Information Science 41(3), 2013. doi:10.1080/15230406.2014.890072.

[8] S. Dabiriab, K. Heaslipa, Developing a Twitter-based traffic event detection model using deep learning architectures, Expert Systems with Applications, volume 118, 2019, pp. 425–439.

[9] E. Alomari, R. Mehmood, I. Katib, Road Traffic Event Detection Using Twitter Data, Machine Learning, and Apache Spark, in: The 3rd IEEE International Conference on Smart City

Innovations (SCI 2019), 2019. doi:10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00332.
[10] C. Gutierrez-Osorio, C. Pedraza, Modern data sources and techniques for analysis and forecast of road accidents: A review, Journal of Traffic and Transportation Engineering (English Edition) 7(4), 2020, pp. 432–446. URL: https://doi.org/10.1016/j.jtte.2020.05.002.
[11] natasha/natasha: Solves basic Russian NLP tasks, API for lower level Natasha projects, 2021. URL: https://github.com/natasha/natasha.
[12] Overview | Geocoding API | Google Developers, 2021. URL: https://developers.google.com/maps/documentation/geocoding/overview?hl=ru.
[13] Social networks in Russia: figures and trends, autumn 2020, 2021. URL: https://br-analytics.ru/blog/social-media-russia-2020.