Spatiotemporal Filtering Pipeline for Efficient Social Networks Data Processing Algorithms

Ksenia Mukhina[✉], Alexander Visheratin, and Denis Nasonov

ITMO University, Saint Petersburg, Russia
mukhinaks@gmail.com, alexvish91@gmail.com, denis.nasonov@gmail.com

Abstract. One of the areas that gathers momentum is the investigation of location-based social networks (LBSNs) because the understanding of citizens' behavior on various scales can help to improve quality of living, enhance urban management, and advance the development of smart cities. But it is widely known that the performance of algorithms for data mining and analysis heavily relies on the quality of input data. The main aim of this paper is helping LBSN researchers to perform a preliminary step of data preprocessing and thus increase the efficiency of their algorithms. To do that we propose a spatiotemporal data processing pipeline that is general enough to fit most of the problems related to working with LBSNs. The proposed pipeline includes four main stages: an identification of suspicious profiles, a background extraction, a spatial context extraction, and a fake transitions detection. Efficiency of the pipeline is demonstrated on three practical applications using different LBSN: touristic itinerary generation using Facebook locations, sentiment analysis of an area with the help of Twitter and VK.com, and multiscale events detection from Instagram posts.

Keywords: Location-based social network · Data processing · Event detection · Sentiment analysis · Tourist path construction · Data filtering pipeline

1 Introduction

In today's world, the idea of studying cities and society through location-based social networks (LBSNs) became a standard for everyone who wants to get insights about people's behavior in a particular area in social, cultural, or political context [11]. Nevertheless, there are several issues concerning data from LBSNs in research. Firstly, social networks can use both explicit (i.e., coordinates) or implicit (i.e., place names or toponyms) geographic references [3]; it is a common practice to allow manual location selection and changing user's position. The Twitter application relies on GPS tracking, but user can correct the position using the list of nearby locations, which causes potential errors from both GPS and user sides [2]. Another popular source of geo-tagged data – Foursquare –

also relies on a combination of the GPS and manual locations selection and has the same problems as Twitter. Instagram provides a list of closely located points-of-interest [16], however, it is assumed that a person will type the title of the site manually and the system will advise the list of locations with a similar name. Although this functionality gives flexibility to users, there is a high chance that a person mistypes a title of the place or selects the wrong one. In Facebook, pages for places are created by the users [25], so all data including title of the place, address and coordinates may be inaccurate.

In addition to that, a user can put false data on purpose. The problem of detecting fake and compromised accounts became a big issue in the last five years [8,19]. Spammers misrepresent the real level of interest to a specific subject or degree of activity in some place to promote their services. Meanwhile, fake users spread unreliable or false information to influence people's opinion [8]. If we look into any popular LBSN, like Instagram or Twitter, location data contains a lot of errors [5]. Thus, all studies based on social networks as a data source face two significant issues: wrong location information stored in the service (wrong coordinates, incorrect titles, duplicates, etc.) and false information provided by users (to hide an actual position or to promote their content).

Thus, in this paper, we propose a set of methods for data processing designed to obtain a clean dataset representing the data from real users. We performed experimental evaluations to demonstrate how the filtering pipeline can improve the results generated by data processing algorithms.

2   Background

With more and more data available every minute and with a rise of methods and models based on extensive data processing [1,13], it was shown that the users' activity strongly correlates with human activities in the real world [21]. For solving problems related to LBSN analysis, it is becoming vital to reduce the noise in input data and preserve relevant features at the same time [10]. Thus, there is no doubt that such problem gathers more and more attention in the big data era. On the one side, data provided by social media is more abundant that standard georeferenced data since it contains several attributes (i.e., rating, comments, hashtags, popularity ranking, etc.) related to specific coordinates [3]. On the other side, the information provided by users of social networks can be false and even users may be fakes or bots. In 2013, Goodchild in [9] raised questions concerning the quality of geospatial data: despite that a hierarchical manual verification is the most reliable data verification method, it was stated that automatic methods could efficiently identify not only false but questionable data. In paper [15], the method for pre-processing was presented, and only 20% of initial dataset was kept after filtering and cleaning process.

One of the reasons for the emergence of fake geotags is a location spoofing. In [28], authors used the spatiotemporal cone to detect location spoofing on Twitter. It was shown that in the New York City, the majority of fake geotags are located in the downtown Manhattan, i.e., users tend to use popular places or

locations in the city center as spoofing locations. The framework for the location spoofing detection was presented in [6]. Latent Dirichlet Allocation was used for the topic extraction. It was shown that message similarity for different users decreases with a distance increase. Next, the history of user check-ins is used for the probability of visit calculation using Bayes model.

The problem of fake users and bots identification become highly important in the last years since some bots are designed to distort the reality and even to manipulate society [8]. Thus, for scientific studies, it is essential to exclude such profiles from the datasets. In [27], authors observed tweets with specific hashtags to identify patterns of spammers' posts. It was shown that in terms of the age of an account, retweets, replies, or follower-to-friend ratio there is no significant difference between legitimate and spammer accounts. However, the combination of different features of the user profile and the content allowed to achieve a performance of 0.95 AUC [23]. It was also shown that the part of bots among active accounts varies between 9% and 15%. This work was later improved by including new features such as time zones and device metadata [26]. In contrast, other social networks do not actively share this information through a public API. In [12], available data from 12 social network sites were studied, and results showed that social networks usually provide information about likes, reposts, and contacts, and keep the data about deleted friends, dislikes, etc., private. Thus, advanced models with a high-level features are applicable only for Twitter and cannot be used for social networks in general.

More general methods for compromised accounts identification on Facebook and Twitter were presented in [22]. The friends ratio, URL ratio, message similarity, friend number, and other factors were used to identify spam accounts. Some of these features were successfully used in later works. For example, in [7], seven features were selected to identify a regular user from a suspicious Twitter account: mandatory – time, message source, language, and proximity – and optional – topics, links in the text, and user interactions. The model achieved a high value of precision with approximately 5% of false positives. In [20], Random Forest classifier was used for spammers identification on Twitter, which results in the accuracy of 92.1%. This study was focused on five types of spam accounts: sole spammers, pornographic spammers, promotional spammers, fake profiles, and compromised accounts. Nevertheless, these methods are user-centered, which means it is required to obtain full profile information for further analysis.

However, there is a common situation where a full user profile is not available for researches, for example, in spatial analysis tasks. For instance, in [17], authors studied the differences between public streaming API of Twitter and proprietary service Twitter Firehose. Even though public API was limited to 1% sample of data, it provided 90% of geotagged data, but only 5% of all sample contains spatial information. In contrast, Instagram users are on average 31 times more likely post data with geotag comparing to Twitter users [16]. Thus, LBSN data processing requires separate and more sophisticated methods that would be capable of identifying fake accounts considering incomplete data. In addition to that, modern methods do not consider cases when a regular user tags a false location for some reason, but it should be taken into account as well.

3   Pipeline Scheme

As it was discussed above, it is critical to use as clean data as possible for research. However, different tasks require different aspects of data to be taken into consideration. In this work, we focus on the main features of the LBSN data: space, time, and messages content. First of all, any LBSN contains data with geotags and timestamps, so the proposed data processing methods are applicable for any LBSN. Secondly, the logic and level of complexity of data cleaning depend on the study goals. For example, if some research is dedicated to studying daily activity patterns in a city, it is essential to exclude all data with wrong coordinates or timestamps. In contrast, if someone is interested in exploring the emotional representation of a specific place in social media, the exact timestamp might be irrelevant. In Fig. 1, elements of a pipeline are presented along with the output data from each stage. As stated in the scheme, we start from general methods for a large scale analysis, which require fewer computations and can be applied on the city scale or higher. Step by step, we eliminate accounts, places, and tags, which may mislead scientists and distort results.
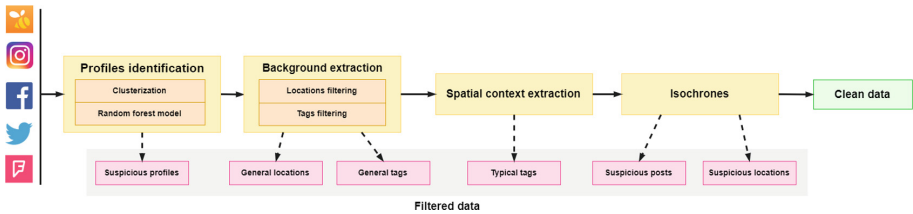

Fig. 1. Pipeline scheme

Suspicious Profiles Identification. First, we identify suspicious accounts. The possibility of direct contact with potential customers attracts not only global brands or local business but spammers, which try to behave like real persons and advertise their products at the same time. Since their goal differs from real people, their geotags often differ from the actual location, and they use tags or specific words for advertising of some service or product. Thus, it is important to exclude such accounts from further analysis. The main idea behind this method is to group users with the same spatial activity patterns. For the business profiles such as a store, gym, etc. one location will be prevalent among the others. Meanwhile, for real people, there will be some distribution in space. However, it is a common situation when people tag the city only but not a particular place, and depending on the city, coordinates of the post might be placed far from user's real location, and data will be lost among the others. Thus, on the first stage, we exclude profiles, who do not use geotags correctly, from the dataset.

We select users with more than ten posts with location to ensure that a person actively uses geotag functionality and commutes across the city. Users with less than ten posts do not provide enough data to correctly group profiles. In addition, they do not contribute sufficiently to the data [7]. Then, we calculate all distances between two consecutive locations for each user and group them by 1000 m, i.e., we count all distances that are less than 1 km, all distances between 1 and 2 km and so on. Distances larger than 50 km are united into one group. After that, we cluster users according to their spatial distribution. The cluster with a deficient level of spatial variations and with the vast majority of posts being in a single location represents business profiles and posts from these profiles can be excluded from the dataset.

At the next step, we use a Random Forest (RF) classifier to identify bots, business profiles, and compromised accounts – profiles, which do not represent real people and behave differently from them. It has been proven by many studies that a RF approach is efficient for bots and spam detection [23, 26]. Since we want to keep our methods as general as possible and to keep our pipeline applicable to any social media, we consider only text message, timestamp, and location as feature sources for our model. We use all data that a particular user has posted in the studied area and extract the following spatial and temporal features: number of unique locations marked by a user, number of unique dates when a user has posted something, time difference in seconds between consecutive posts. For time difference and number of posts per date, we calculated the maximum, minimum, mean, and standard deviation. From text caption we have decided to include maximum, minimum, average, mean, standard deviation of following metrics: number of emojis per post, number of hashtags per post, number of words per post, number of digits used in post, number of URLs per post, number of mail addresses per post, number of user mentions per post. In addition to that, we extracted money references, addresses, and phone numbers and included their maximum, minimum, average, mean, and standard deviation into the model. In addition, we added fraction of favourite tag in all user posts. Thus, we got 64 features in our model. As a result of this step, we obtain a list of accounts, which do not represent normal users.

City Background Extraction. The next stage is dedicated to the extraction of basic city information such as a list of typical tags for the whole city area and a set of general locations. General locations are places that represent large geographic areas and not specific places. For example, in the web version of Twitter user can only share the name of the city instead of particular coordinates. Some social media like Instagram or Foursquare are based on a list of locations instead of exact coordinates, and some titles in this list represent generic places such as streets or cities. Data from these places is useful in case of studying the whole area, but if someone is interested in studying actual temporal dynamics or spatial features, such data will distort the result. Also, it should be noted that even though throughout this paper we use the word 'city' to reference the particular geographic area, all stages are applicable on the different scales starting from city districts and metropolitan regions to states, countries, or continents.

Firstly, we extract names of administrative areas from Open Street Maps (OSM). After that, we calculate the difference between titles in social media data and data from OSM with the help of Damerau-Levenshtein distance. We consider a place to be general if the distance between its title and some item from the list of administrative objects is less than 2. These locations are excluded from the further analysis. For smaller scales such as streets or parks, there are no general locations.

Then, we analyze the distribution of tags mentions in the whole area. The term 'tag' denotes the important word in the text, which characterizes the whole message. Usually, in LBSN, tags are represented as hashtags. However, they can also be named entities, topics, or terms. In this work, we use hashtags as an example of tags, but this concept can be further extrapolated on tags of different types. The most popular hashtags are usually related to general location (e.g., #nyc, #moscow) or a popular type of content (#photo, #picsoftheday, #selfie) or action (#travel, #shopping, etc.). However, these tags cannot be used to study separate places and they are not relevant either to places or to events since they are actively used in the whole area. Nevertheless, scientists interested in studying human behavior in general can use this set of popular tags because it represents the most common patterns in the content. In this work, we consider tag as general if it was used in more than 1% of locations.

However, it is possible to exclude tags related to public holidays. We want to avoid such situations and keep tags, which have a large spatial distribution but narrow peak in terms of temporal distribution. Thus, we group all posts that mentioned a specific tag for the calendar year and compute their daily statistics. We then use the Gini index G to identify tags, which do not demonstrate constant behavior throughout the year. If $G \geq 0.8$ we consider tag as an event marker because it means that posts distribution have some peaks throughout the year. This pattern is common for national holidays or seasonal events such as sports games, etc. Thus, after the second stage, we obtain the dataset for further processing along with a list of common tags and general locations for the studying area.

Spatial Context Extraction. Using hashtags for events identification is a powerful strategy, however, there are situations where it might fail. The main problem is that people often use hashtags to indicate their location, type of activity, objects on photos and etc. Thus, it is important to exclude hashtags which are not related to the possible event. To do that, we grouped all hashtags by locations, thus we learn which tags are widely used throughout the city and which are place related. If some tag is highly popular in one place, it is highly likely that the tag describes this place. Excluding common place-related tags like #sea or #mall for each location, we keep only relevant tags for the following analysis. In other words, we get the list of tags which describe a normal state of particular places and their specific features. However, such tags cannot be indicators of events.

Fake Transitions Detection. The last stage of the pipeline is dedicated to suspicious posts identification. Sometimes, people cannot share their thoughts or

photos immediately. It leads to situations where even normal users have a bunch of posts, which are not accurate in terms of location and timestamp. At this stage, we exclude posts that cannot represent the right combination of their coordinates and timestamps. This process is similar to the ideas for location spoofing detection – we search for transitions, which someone could not make in time. The standard approach for detection of fake transitions is to use space-time cones [28], but in this work, we suggest the improvement of this method – we use isochrones for fake transitions identification. In urban studies, isochrone is an area that can be reached from a specified point in equal time. Isochrone calculation is based on usage of real data about roads, that is why this method is more accurate than space-time cones. For isochrone calculation, we split the area into several zones depending on their distance from the observed point: pedestrian walking area (all locations in 5 km radius), car/public transport area (up to 300 km), train area (300–800 km) and flight area (further than 800 km). This distinction was to define a maximum speed for every traveling distance. The time required for a specific transition is calculated by the following formula:

$$t = \frac{1}{v} \sum_{i=1}^{N} s_i, \text{ where } v = \begin{cases} 5, & \text{if } S \leq 5 \\ 120, & 5 < S \leq 300 \\ 300 & 300 < S \leq 800 \\ 900 & S \geq 800 \end{cases} \tag{1}$$

where s_i is the length of the road segment and v is the maximum possible velocity depending on the inferred type of transport. The road data was extracted from OSM. It is important to note that on each stage of the pipeline, we get output data, which will be excluded, such as suspicious profiles, baseline tags, etc. However, this data can also be used, for example, for training novel models for fake accounts detection.

4   Experiments

4.1   Touristic Path Construction

The first experiment was designed to highlight the importance of general location extraction. To do that, we used the points-of-interest dataset for Moscow, Russia. The raw data was extracted from Facebook using the Places API and contained 40,473 places. The final dataset for Moscow contained 40,215 places, and 258 general sites were identified. However, it should be noted that among general locations, there were detected 'Russia' (8,984,048 visitors), 'Moscow', 'Russia' (7,193,235 visitors), 'Moscow Oblast' (280,128 visitors). For instance, the most popular non-general locations in Moscow are Sheremetyevo Airport and Red Square, with only 688,946 and 387,323 check-ins, respectively. The itinerary construction is based on solving the orienteering problem with functional profits (OPFP) with the help of the open-source framework FOPS [18]. In this approach, locations are scored by their popularity and by farness distance.

We used the following parameters for the Ant Colony Optimization algorithm: 1 ant per location and 100 iterations of the algorithm, as it was stated in the original article. The time budget was set to 5 h, the Red Square was selected as a starting point, and Vorobyovy Gory was used as a finish point since they two highly popular touristic places in the city center.

The resulting routes are presented in Fig. 2. Both routes contain extra places, including major parks in the city: Gorky park and Zaryadye park. However, there are several distinctions in these routes. The route based on the raw data contains four general places (Fig. 2, left) – 'Moscow', 'Moscow, 'Russia', 'Russia', and 'Khamovniki district', which do not correspond to actual places. Thus, 40% of locations in the route cannot be visited in real life. In contrast, in case of the clean data (Fig. 2, right), instead of general places algorithm was able to add real locations, such as Bolshoi Theatre and Central Children's Store on Lubyanka with the largest clock mechanism in the world and an observation deck with the view on Kremlin. Thus, the framework was able to construct a much better itinerary without any additional improvements in algorithms or methods.
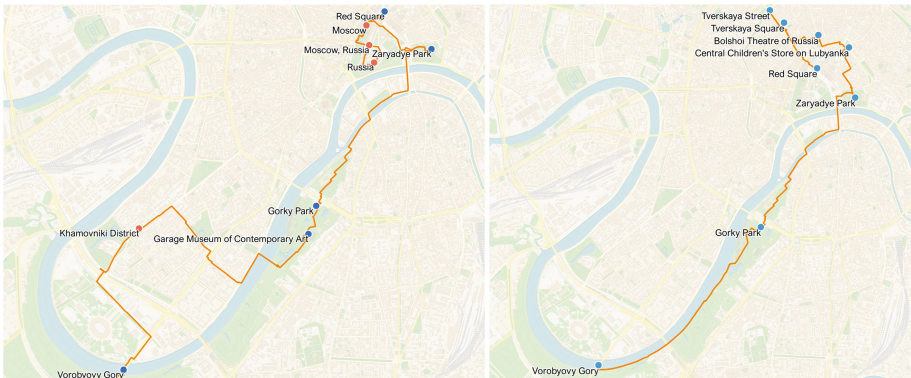

Fig. 2. Comparison of walking itineraries for original data (left) and cleaned dataset (right). Red dots related general locations, blue color indicates actual places (Color figure online)

4.2   Sentiment Analysis

To demonstrate the value of background analysis and typical hashtags extraction stages, we investigated the scenario of analysis of users' opinions in a geographical area via sentiment analysis. We used a combined dataset of Twitter and VK.com posts taken in Sochi, Russia, during 2016. Sochi is one of the largest and most popular Russian resorts. It was also the host of the Winter Olympics in 2014. Since Twitter and VK.com provide geospatial data with exact coordinates, we created a squared grid with a cell size equal to 350 m. We then kept only cells containing data (Fig. 3, right) – 986 cells in total. Each cell was considered as a separate location for the context extraction. The most popular tags in the area

are presented in Fig. 3 (left). Tag '#sochi' was mentioned in 1/3 of cells (373 and 321 cells for Russian and English versions of the tag, respectively). The follow-up tags '#sochifornia' (used in 146 cells) and '#sea' (mentioned in 130 cells) were twice less popular. After that, we extracted typical tags for each cell. We considered a post to be relevant to the place if it contained at least one typical tag. Thus, we can be sure that posts represent the sentiment in that area.

Fig. 3. The most popular tags in the area of Sochi (left) and results of sentiment analysis for raw and clean data (right)

The sentiment analysis was executed in two stages. First, we prepare the text for polarity detection. To do that, we delete punctuation, split text in words, and normalized text with the help of [14]. In the second step, we used the Russian sentiment lexicon [4] to get the polarity of each word (1 indicates positive word and -1 negative word). The sentiment of the text is defined as 1 if a sum of polarities of all words more than zero and -1 if the sum is less than zero. The sentiment of the cell is defined as an average sentiment of all posts. On the Fig. 3, results of sentiment analysis are presented, cells with average sentiment less than 0.2 were marked as neutral. It can be noted from maps that after the filtering process, more cells have a higher level of sentiment. For Sochi city center, the number of posts with the sentiment $|s| \geq 0.75$ increased by 26.6%. It is also important that number of uncertain cell with sentiment rate $0.2 \leq |s| \leq 0.75$ decreased by 36.6% from 375 to 239 cells. Thus, we highlighted the strong positive and negative areas and decreased the number of uncertain areas by applying the context extraction stage of the proposed pipeline.

4.3   Event Detection

In this experiment, we applied the full pipeline on the Instagram data. New York City was used as a target city in the event detection approach [24] similar to this study the area was bounded by this latitude interval – [40.4745, 40.9445] and longitude varied in [−74.3060, −73.6790]. We collected the data from over 108,846 locations for a period of up to 7 years. The total number of posts extracted from the New York City area is 67,486,683.

In the first step, we try to exclude from the dataset all users who provide incorrect data, i.e. use several locations instead of the whole variety. We group users with the help of K-means clustering method. The appropriate number of clusters was obtained by calculating the distortion parameter. Deviant cluster contained 504,558 users out of 1,978,403. The shape of deviant clusters can be seen in Fig. 4. Suspicious profiles mostly post in the same location. Meanwhile, regular users have variety in terms of places.
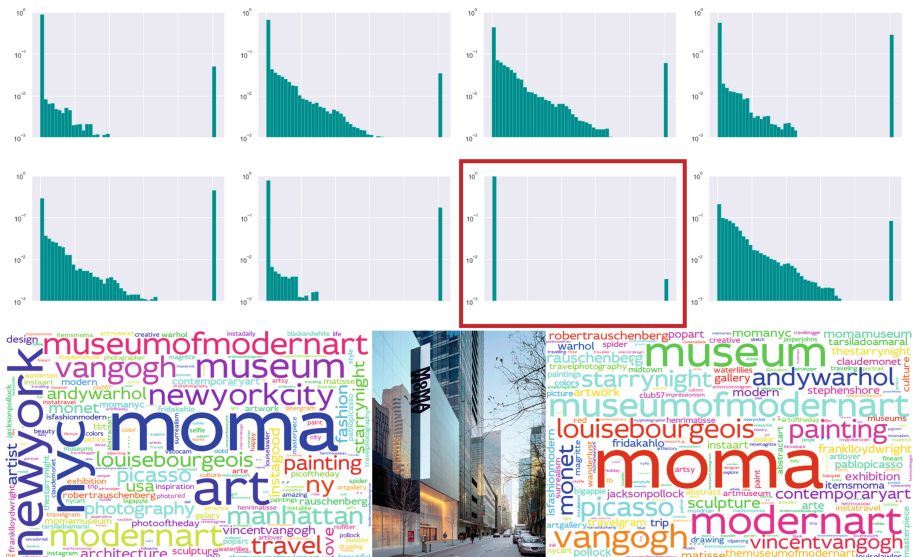

Fig. 4. Obtained clusters' shape (top, red square indicates deviant clusters) and popular tags for Museum of Modern Art before and after filtering (bottom) (Color figure online)

After that, we trained our RF model using manually labelled data from both datasets. The training dataset contains 223 profiles with 136 ordinary users and 87 fake users; test data consists of 218 profiles including 146 normal profiles and 72 suspicious accounts. The model distinguishes a regular user from suspicious successfully. 122 normal user were detected correctly and 24 users were marked as suspicious. 63 suspicious users out of 72 were correctly identified. Thus, there

were obtained 72% of precision and 88% of recall. Since the goal of this work is to get clean data as a result, we are interested in a high value of recall and precision is less critical. As a result, we obtained a list of 1,132,872 profiles which related to real people.

At the next step, we used only data from these users to extract background information about cities. 304 titles of general locations were derived for New York. These places were excluded from further analysis. After that, we extracted general hashtags; the example of popular tags in location before and after background tags extraction is presented on the Fig. 4. General tags contain mostly different term related to toponyms and universal themes such as beauty or life. Then, we performed the context extraction for locations. For each location typical hashtags were identified as 5% most frequent tags among users. We consider all posts from one user in the same location as one to avoid situations where someone tries to force their hashtag. We will use extracted lists to exclude typical tags from posts.

After that, we calculated isochrones for each normal users to exclude suspicious posts from data. In addition to that, locations with a high rate of suspicious posts (75% or higher part of posts in location was detected as suspicious) were excluded as well. There was 16 locations in New York City. The final dataset for New York consists of 103,977 locations. For event detection we performed the same experiment which was described in [24]. In the original approach the spike in activity in particular cell of the grid was consider as an event. To find these spikes in data, historical grids is created using retrospective data for a calendar year. Since we decrease amount of data significantly, we set threshold value to 12. We used data for 2017 to create grids, then we took two weeks from 2018 for the result evaluation: a week with a lot of events during 12–18 of March and an ordinary week with less massive events 19–25 February.

The results of the recall evaluation are presented in Table 1. As can be seen from the table on an active week, the recall increment was 14.9% and for non-active week recall value increase on 32.6%. It is also important to note that some events, which do not have specific coordinates, such as snowfall in March or Saint Patrick's day celebration, were detected in the less number of places. This leads to lesser number of events in total and more significant contribution to the false positive rate. Nevertheless, the largest and the most important events, such as nationwide protest '#Enough! National School Walkout' and North American International Toy Fair are still detected from the very beginning. In addition to that due to the altered structure of historical grids, we were able to discover new events such as a concert of Canadian R&B duo 'dvsn', 2018 Global Engagement Summit at UN Headquarters, etc. These events were covered with a low number of posts and stayed unnoticed during the original experiment. However, the usage of clean data helped to highlight small events which are essential for understanding the current situation in the city.

Table 1. Comparison of event detection results

Recall	Active week	Non-active week	Total
Before	81.8%	63.9%	73%
After	96.7%	96.5%	96.7%

5   Conclusion and Future Works

In this work, we presented a spatiotemporal filtering pipeline for data preprocessing. The main goal of this process is to exclude unreliable data in terms of space and time. The pipeline consists of four stages: during the first stage, suspicious user profiles are extracted from data with the help of K-means clustering and Random Forest classifier. On the next stage, we exclude the buzz words from the data and filter locations related to large areas such as islands or city districts. Then, we identify the context of a particular place expressed by unique tags. In the last step, we find suspicious posts using the isochrone method. Stages of the pipeline can be used separately and for different tasks. For instance, in the case of touristic walking itinerary construction, we used only general location extraction, and the walking itinerary was improved by replacing 40% of places. In the experiment dedicated to sentiment analysis, we used a context extraction method to keep posts that are related to the area where they were taken, and as a result, 36.2% of uncertain areas were identified either as neutral or as strongly positive or negative. In addition to that, for event detection, we performed all stages of the pipeline, and recall for event detection method increased by 23.7%.

Nevertheless, there are ways for further improvement of this pipeline. In Instagram, some famous places such as Times Square has several corresponding locations including versions in other languages. This issue can be addressed by using the same method from the general location identification stage. We can use distance to find places with a similar name. Currently, we do not address the repeating places in the data since it can be a retail chain, and some retail chains include over a hundred places all over the city. In some cases, it can be useful to interpret a chain store system as one place. However, if we want to preserve distinct places, more complex methods are required. Despite this, the applicability of the spatiotemporal pipeline was shown using the data from Facebook, Twitter, Instagram, and VK.com. Thus, the pipeline can be successfully used in various tasks relying on location-based social network data.

References

1. Biessmann, F., Salinas, D., Schelter, S., Schmidt, P., Lange, D.: "Deep" learning for missing value imputation in tables with non-numerical data. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, pp. 2017–2025. ACM, New York (2018). https://doi.org/10.1145/3269206.3272005

2. Burton, S.H., Tanner, K.W., Giraud-Carrier, C.G., West, J.H., Barnes, M.D.: "Right time, right place" health communication on Twitter: value and accuracy of location information. J. Med. Inet. Res. **14**(6), e156 (2012). https://doi.org/10.2196/jmir.2121

3. Campagna, M.: Social media geographic information: why social is special when it goes spatial. In: European Handbook Crowdsourced Geographic Information, pp. 45–54 (2016). https://doi.org/10.5334/bax.d

4. Chen, Y., Skiena, S.: Building sentiment lexicons for all major languages. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pp. 383–389 (2014)

5. Cvetojevic, S., Juhasz, L., Hochmair, H.: Positional accuracy of Twitter and Instagram Images in urban environments. GI_Forum **1**, 191–203 (2016). https://doi.org/10.1553/giscience2016_01_s191

6. Ding, C., et al.: A location spoofing detection method for social networks (short paper). In: Gao, H., Wang, X., Yin, Y., Iqbal, M. (eds.) CollaborateCom 2018. LNICST, vol. 268, pp. 138–150. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-12981-1_9

7. Egele, M., Stringhini, G., Kruegel, C., Vigna, G.: COMPA: detecting compromised accounts on social networks (2013). https://doi.org/10.1.1.363.6606

8. Ferrara, E., Varol, O., Davis, C., Menczer, F., Flammini, A.: The rise of social bots. Commun. ACM **59**(7), 96–104 (2016). https://doi.org/10.1145/2818717

9. Goodchild, M.F.: The quality of big (geo)data. Dialogues Hum. Geogr. **3**(3), 280–284 (2013). https://doi.org/10.1177/2043820613513392

10. Silva, H., et al.: Urban computing leveraging location-based social network data: a survey. ACM Comput. Surv. **52**(1), 1–39 (2019)

11. Hochman, N., Manovich, L.: Zooming into an Instagram City: reading the local through social media. First Monday (2013). https://doi.org/10.5210/fm.v18i7.4711

12. John, N.A., Nissenbaum, A.: An agnotological analysis of APIs: or, disconnectivity and the ideological limits of our knowledge of social media. Inf. Soc. **35**(1), 1–12 (2019). https://doi.org/10.1080/01972243.2018.1542647

13. Kapoor, K.K., Tamilmani, K., Rana, N.P., Patil, P., Dwivedi, Y.K., Nerur, S.: Advances in social media research: past, present and future. Inf. Syst. Front. **20**(3), 531–558 (2017). https://doi.org/10.1007/s10796-017-9810-y

14. Korobov, M.: Morphological analyzer and generator for Russian and Ukrainian languages. In: Khachay, M.Y., Konstantinova, N., Panchenko, A., Ignatov, D.I., Labunets, V.G. (eds.) AIST 2015. CCIS, vol. 542, pp. 320–332. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-26123-2_31

15. Lavanya, P.G., Kouser, K., Suresha, M.: Efficient pre-processing and feature selection for clustering of cancer tweets. In: Thampi, S.M., et al. (eds.) Intelligent Systems, Technologies and Applications. AISC, vol. 910, pp. 17–37. Springer, Singapore (2020). https://doi.org/10.1007/978-981-13-6095-4_2

16. Manikonda, L., Hu, Y., Kambhampati, S.: Analyzing user activities, demographics, social network structure and user-generated content on Instagram. arXiv preprint abs/1410.8 (2014)

17. Morstatter, F., Pfeffer, J., Liu, H., Carley, K.M.: Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's Firehose. In: Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013 (2013)
18. Mukhina, K.D., Visheratin, A.A., Nasonov, D.: Orienteering problem with functional profits for multi-source dynamic path construction. PLoS ONE **14**(4), e0213777 (2019). https://doi.org/10.1371/journal.pone.0213777
19. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: a data mining perspective. ACM SIGKDD Explor. Newsl. **19**(1), 22–36 (2017)
20. Singh, M., Bansal, D., Sofat, S.: Who is who on Twitter-Spammer, fake or compromised account? A tool to reveal true Identity in real-time. Cybern. Syst. **49**(1), 1–25 (2018). https://doi.org/10.1080/01969722.2017.1412866
21. Steiger, E., Westerholt, R., Resch, B., Zipf, A.: Twitter as an indicator for whereabouts of people? Correlating Twitter with UK census data. Comput. Environ. Urban Syst. **54**, 255–265 (2015). https://doi.org/10.1016/j.compenvurbsys.2015.09.007
22. Stringhini, G., Kruegel, C., Vigna, G.: Detecting spammers on social networks. In: Proceedings of the 26th Annual Computer Security Applications Conference, ACSAC 2010, pp. 1–9. ACM, New York (2010). https://doi.org/10.1145/1920261.1920263
23. Varol, O., Ferrara, E., Davis, C.A., Menczer, F., Flammini, A.: Online human-bot interactions: detection, estimation, and characterization. In: Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017, pp. 280–289 (2017)
24. Visheratin, A.A., Visheratina, A.K., Nasonov, D., Boukhanovsky, A.V.: Multiscale event detection using convolutional quadtrees and adaptive geogrids. In: 2nd ACM SIGSPATIAL Workshop on Analytics for Local Events and News, p. 10 (2018). https://doi.org/10.1145/3282866.3282867
25. Wilken, R.: Places nearby: Facebook as a location-based social media platform. New Media Soc. **16**(7), 1087–1103 (2014). https://doi.org/10.1177/1461444814543997
26. Yang, K.C., Varol, O., Davis, C.A., Ferrara, E., Flammini, A., Menczer, F.: Arming the public with artificial intelligence to counter social bots. Hum. Behav. Emerg. Technol. **1**(1), e115 (2019). https://doi.org/10.1002/hbe2.115
27. Yardi, S., Romero, D., Schoenebeck, G., boyd, D.: Detecting spam in a Twitter network. First Monday **15**(1) (2009). https://doi.org/10.5210/fm.v15i1.2793
28. Zhao, B., Sui, D.Z.: True lies in geospatial big data: detecting location spoofing in social media. Ann. GIS **23**(1), 1–14 (2017). https://doi.org/10.1080/19475683.2017.1280536