

SOCIAL MEDIA DATA PROCESSING AND ANALYSIS BY MEANS OF MACHINE LEARNING FOR RAPID DETECTION, ASSESSMENT AND MAPPING THE IMPACT OF DISASTERS

P. M. Kikin^{1*}, A. A. Kolesnikov², E. A. Panidi³

¹ Peter the Great St.Petersburg Polytechnic University (SPbPU), St.Petersburg, Russian Federation - it-technologies@yandex.ru

² Siberian State University of Geosystems and Technologies, Novosibirsk, Russian Federation - alexeykw@yandex.ru

³ Saint Petersburg State University, St.Petersburg, Russian Federation - panidi@ya.ru

KEY WORDS: Neural Networks, Social Media, Remote Sensing, Disaster Management, CNN, Machine Learning

ABSTRACT:

The main factor determining the possibility of using data obtained from social media as a source of information about the threat of emergencies is their relevance and accuracy. Thus, the important task is the determination of metrics for evaluating these parameters for a specific publication in a social media. It is worth noting the importance of this information channel as a source of eyewitness accounts from the scene. A comparison of social media data and official sources shows that social media contain a significant amount of unique information at different stages of emergency development. Also, when monitoring the situation for a specific event, social media allows to get more relevant information in comparison to official sources. Another important task is to search for emergency messages and their most accurate localization in space. A promising solution for the analysis and processing of social media data during emergency response is the application of artificial intelligence methods, and, particularly, machine learning techniques.

1. INTRODUCTION

Natural disasters as well as major manmade incidents are an increasingly serious threat for civil society. Effective, fast and coordinated disaster management crucially depends on the availability of a real-time situation picture of the affected area. Emergency managers require timely and accurate information on areas affected by disasters to prioritize relief efforts and plan mitigation measures against damage. Therefore, it is very important to provide emergency services quickly with necessary post-disaster maps, compiled on the principles of rapid mapping. Most of the traditional techniques supposed to use just one source of data, while in a time-critical disaster situation, utilization of multiple data sources is particularly desirable. Each additional data source provides extra features which can increase possible mapping efficiency and accuracy.

One of promising data source is social media. Increasing availability of smartphones together with a wide coverage of high-speed internet is leading to documentation of disaster events directly by individuals, with information shared in real-time. Social media data became widely explored just recently and largely due to the development of big data and natural language processing (NLP) technologies (Bruno, A. et al., 2019, Barozzi, S. et al., 2019, De Albuquerque, J.P. et al., 2015). However, despite much interest in using such sources in disaster management, exploiting this information is not trivial. The data usually have no validation or assessment of quality and may contain deliberate or unintended location bias (Hays, J., et al., 2008, Kumar S. S., et al., 2018, Middleton, S. E., et al., 2018).

The goal of represented research was to develop a new automated approach for rapid detection, assessment and mapping the impact of disasters employing social media data together with remote sensing data using various machine learning and GIS-based methods of spatial, image and texts analysis.

The complexity of emergencies monitoring task using additional information from social media is that the information flow of messages in social media contains:

- a huge amount of data that is constantly growing;
- duplicates;
- incomplete and inaccurate messages;
- false information;
- informal messages in which grammatical errors, shorthand, symbols (for example, emoticons), spelling errors, slang, irony, etc.

Proposed approach allows to achieve:

- very high accuracy of damage assessment and mapping;
- scalability of approach for various emergencies and territories;
- ability to work with various combinations of data sources;
- low dependency on data quality and quantity;
- high automatization and assessment speed ((Leichter A., et al., 2018, Migliaccio F., et al., 2019, Tavra M., et al., 2019)).

2. MATERIALS

To address these issues, it was necessary to make few general steps. Processing of social media data and extracting useful information from its publications are very challenging tasks, because crowdsourcing data, shared by civil users, is more difficult both in aggregation and in validation. This is due to multiple reasons including the challenge of extracting relevant information (e.g. identifying topicality) within unstructured or semi-structured web-harvested data, unknown quality (there may be little or no relevant metadata), and difficulty integrating it with other sources (which may, in turn, have their own issues of quality and uncertainty). When using social media, biases may also be presented, for example, due to a lack of digital engagement within certain demographics of the populations of

* Corresponding author

particular areas (Francalanci, C., et al., 2018, Houston, J.B., et al., 2015).

So, to address this issue the following steps were taken:

- algorithms for collecting and initial filtering of relevant text and photo data were developed;
- an approach for extraction of features, important for damage assessment, from text information was developed;
- approach for extracting useful features from user-shared photo images was designed.

Since extracted features did not have any certain geometric form and strict spatial reference, it was necessary to design a way of how to order, represent and finally perform mapping of this data. Generally, the designed approach allows to understand what, where and which way is described in social media publications, and put it on a map (Krizhevsky, A., et al., 2012, Ravi Shankar A., et al., 2019).

Social media data was retrieved particularly through the public API with the help of the Tweepy open-source software, and by self-developed Python script. The queries were restricted to geotagged imagery and text publications falling within the study area and time frame. To detect relevant records special keywords and tags were designed with regard to the type of emergency. This words and tags were used while searching across certain defined fields of publications (Al-Rfou, R., et al., 2015, Havas, C., et al., 2015). This data was expanded by a simplified database of images derived from Flickr (NUS-WIDE-LITE), which contains, in addition to the images themselves, tag sets, EXIF and spatial references (Chua T., et al., 2009).

In addition, relevant posts from Facebook and the training dataset from the «Location» contest conducted by the Advanced Research Foundation were used (<https://fpi.gov.ru/tenders/799/>). The purpose of the competition is to determine the location by photo and recognition of objects on it, such as signs, license plates, information and advertising signs, attractions, as well as features of architecture and natural landscapes. The training dataset consists of 5,000 photos in Asia and Europe, scaled to 500 by 500 pixels, a set of tags and text messages (if any in the post).

3. METHODS

To further analyze the sample data, the exchangeable image file (EXIF) metadata associated with each image was retrieved. The EXIF provides metadata regarding each individual image, which include information recorded by the camera sensor or device used for capturing. These include potentially valuable geographic data such as the azimuth of the device during capturing.

In order to perform feature extraction from obtained images, Convolutional Neural Network (CNN) and the ImageNet pre-trained autoencoder were used. Several CNNs were estimated and compared to each other (Nguyen, D. T., et al., 2017, Ryabchenko, N. A., et al., 2016).

Results of feature extraction, both from imagery and text data, together with location data, and EXIF features (in case of availability for images) allowed to perform mapping of this features, applying traditional geocoding techniques to create point and linear objects and Kernel Density Estimation (KDE) to map distributed spatial phenomena.

To combine results of semantic segmentation, performed by various CNN models and results of feature extraction from social media data the probabilistic model was designed, which

evaluates quantitatively the contribution of each data source, and as a result forms a decision on class of damage a detected object belongs to.

To increase the accuracy of determining geographic coordinates by the presence of informative objects on the image, neural networks were trained (the Wide Residual Network (WRN-50-2) was chosen as the basic architecture), specializing in certain types of informative objects. Such objects were selected: road signs, flags, advertising (with the local language), inscriptions in the local language, car numbers, plants, roofs (material, color, structure, shape), bicycles, street material (granite, marble, asphalt - also the degree of its homogeneity), the proportions of streets (along windows, walls, etc.), the presence of the sea / ocean / lake / river, mountains / hills, the position of the sun (for orientation of photographs), the time of shooting (to determine the angle of incidence of the Sun), urns, traffic lights and lanterns, special vehicles, transport labels, public transport (with specific color / model), road markings, bridges, walls and fences (structure, ornament, color), patterns of buildings and ornaments, sewers (there is a certain color dependence on the country), sidewalk (borders), direction of travel (left / right), the symbol of the local currency / banknotes. The presence of each of these objects and its parameters formed additional features of each of the photographs. As for the task as a whole, the auto-tagging of photographs is a more complicated and complex task than just the classification of images, because it is necessary to determine the context, the presence of several objects and their interaction with each other. WRN-50-2 is an architecture trained on ImageNet, which is a ResNet-50 with double channel expansion. In addition to high accuracy, it is necessary to note an increase in the learning speed with the same number of features, which can be explained by the "width" of the Res-block and, therefore, great opportunities for parallelizing the calculations in it. As an alternative and a kind of base line, a combined model with two inputs (and data blocks, respectively) was used, using the pre-trained ConvNet VGG16 model for image processing and the NLP model analyzing combinations of pre-trained word sets using GloVe and Keras LSTM for receiving and processing tags words. These two input models first process their data block in parallel and then are combined with a fully connected output classification model that uses both the output of the image recognition model and the output of the NLP model to determine if the input image pair and set of word tags are a match and with what probability.

Well-developed algorithmic methods were used to recognize inscriptions and individual characters in the study. To detect areas of possibly containing letters and numbers, voting was used based on the Haar cascade that had been trained and a histogram analysis of the regions. Voting was organized as follows. Pixels assigned to inscriptions simultaneously by the Haar cascade and histogram analysis were moved to the next processing stage. Other pixels that were selected just by only one of the algorithms were proceed for the further processing only if they formed areas with a continuous area of at least 100 pixels. It was necessary to try to normalize the areas selected at the first stage to increase the quality of character recognition of the last stage. Since most of the inscriptions we are interested in are presented in the form of information tablets, the most effective was the use of the Hough transformation. This transformation allows to quickly select two main lines and crop the image by them. In addition, a contrast increase filter (OCR 1 block on figure 1) was used for the obtained image. Since part of the inscriptions still does not have a strictly rectangular shape and some important characters could be eliminated, the initial set of pixels obtained after the first stage of processing was sent for recognition to the OCR 2 block. Thus,

at the last stage of recognition, two images were used, which were processed in Tesseract OCR. The two summary lines were further compared. The comparison was performed using the Tanimoto coefficient. With similarities of more than 70%, priority was given to the line obtained based on the processed image. If in the raw image the number of characters was more than 2 times higher than the results in the processed image, then the lines were combined through a space, otherwise the first option was used. The final processing scheme for text blocks is shown in Figure 1. The recognition results became an additional feature for other parameters of the photo. The presented algorithm has satisfactory performance and well interpretability, but its accuracy on complex inscriptions is insufficient. Therefore, in the future it is planned to expand the intermediate processing steps with neural network tools.

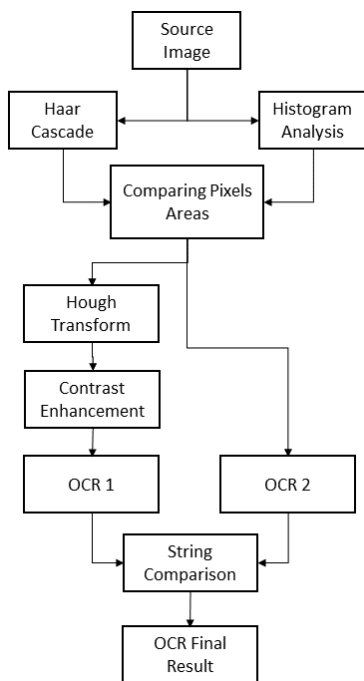


Figure 1. Photo processing algorithm

The Tesseract OCR was used in the study to recognize labels and individual characters / numbers. Recognition results were used as additional features to other parameters of photographs.

When assessing the accuracy of the obtained mathematical model, the importance levels of the forecast quality were identified. It was necessary because an error of several tens of kilometers in a sparsely populated territory is not so significant compared with an error of a kilometer in the center of a large city. To calculate the coefficients of importance levels, we used the normalized population density at the location of the survey according to the formula:

$$pd_i = \frac{PD_i}{PD_{max}} \quad (1)$$

where pd_i = normalized population density at the location of the i -th image

PD_i = population density at the place of shooting images i

PD_{max} = Earth's maximum population density.

The error in determining the coordinates was calculated by the following formula:

$$error = \sqrt{(1 + 0.1pd_i) * \Delta d^2} \quad (2)$$

where d = distance between the predicted and true geographic coordinates of the image (meters).

The final image processing algorithm is as follows. The first step is to extract information from EXIF. If there are shooting coordinates in it, then depending on the settings for checking the correctness, either the transition to the next image is carried out, or an additional check of the extracted coordinates is performed based on the following steps. The second step is to extract any textual information using Tesseract OCR. Next, the extracted text is used to query the geocoder and to create a separate feature containing the source text. If the coordinates were obtained as a result of geocoding, then they are used for basic positioning or verification of data extracted from EXIF. In the third step, the image content is classified and a list of tags is obtained, corresponding to the previously described list of informative objects. In the absence of parameters with coordinates in EXIF and after the geocoding procedure, the list of tags obtained in the third step is used to find the most similar images and the resulting coordinates are calculated based on the proximity parameter. These steps can be represented in the form of the following diagram, presented in Figure 2.

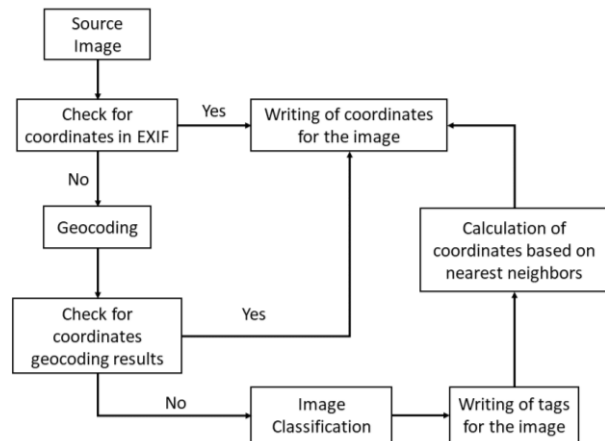


Figure 1. Photo processing algorithm

Thus, each photo after processing is characterized by the following features:

- coordinates from EXIF;
- list of tags from the post of a social network;
- coordinates based on geocoding results;
- coordinates based on other, most similar photos;
- text string from a social network message;
- a text string formed from characters recognized using OCR;
- a list of tags indicating the presence and parameters of informative objects extracted using a neural network and encoded using the One-Hot-Encoding method;
- date and time of shooting from EXIF;
- date and time based on the position of the sun (if possible).

At the moment, in the proposed algorithm there is no mutual verification of the reliability of some information on the basis of others from the same post. For example, there is no mutual check:

- text of the message;
 - the results of the OCR (for example, the name of an institution);
 - tags;
 - time taken from the metadata of the photo and time of the event.
- Such checks are planned to be implemented in subsequent studies as additional experiments are required. Difficulties arise with options for indicating the degree of certainty. If for the date, time of shooting and coordinates you can specify a single number showing the coefficient of similarity (the difference between dates or points in space), then for text fields options are possible. A simple indication of the number of identical words (tokens) can be applied, but it will not be a full indicator of similarity. It is also necessary to take into account the total length of the text, the types of words themselves (named entity), that is, it is more important if the matching words are designations of geographical objects, names of organizations, etc.

4. EXPERIMENTAL RESULTS

To assess the impact of criteria on the quality of determining coordinates, all available 5000 objects were divided into groups according to population density, as well as additional criteria. Photos containing any inscriptions were evaluated separately (table 1).

Normalized population density	For photos with text data	For photos without text data
< 0,2	5110	12780
> 0.2 and <0.5	2250	5160
> 0.5 and <0.8	1580	3490
> 0.8	1310	3100

Table 1. Coordinate prediction results for photos containing captions

The next group was identified by the criterion for the presence of tags to the post of a social network (table 2).

Normalized population density	For photos with tags	For photos without tags
< 0,2	4510	12780
> 0.2 and <0.5	1988	5160
> 0.5 and <0.8	1432	3490
> 0.8	958	3100

Table 2. Coordinate prediction results for posts containing tags

Another group of objects was identified and evaluated according to the quality criterion for automatic tagging using a neural network (table 3).

Normalized population density	For photos with automatically extracted tags	For photos without automatically extracted tags
< 0,2	3921	11391
> 0.2 and <0.5	1463	4269
> 0.5 and <0.8	1175	2741
> 0.8	810	2691

Table 3. Coordinate prediction results for photos with tag extraction using a neural network

An analysis of the results showed that the procedure for the automated generation of tags using a neural network has the greatest impact on the quality of determining coordinates. The rather large errors obtained when processing photos with captions are most likely due to the imperfection of the basic geocoding

algorithms and can be corrected to the smaller side by using natural language processing algorithms. Also, considering the large difference between the error values for the normalized intervals, less than 0.2 and in the range from 0.2 to 0.5, in the future it is expected to expand the number of intervals and shift the threshold values of the intervals.

For additional validation of the information extracted from the message, it makes sense to expand the list of attributes based on a structural analysis of the relationship between the author of the message and the photo (tools similar to Twitter Alerts functionality). Using this approach, each message can be characterized not only on the basis of available text, photos and tags, but also on the basis of information about the author. In turn, this information can be divided into two categories. The first is the author's attributes:

- name;
 - age;
 - status;
 - places of study and work;
 - etc.
- And the second is social relations:
- number of connections;
 - their density;
 - thematic and social orientation;
 - etc.

In addition, it is possible to additionally evaluate the distances (in social connections) between users of the same settlement and thereby indirectly talk about the proximity of a real location (Soufiene, J., et al., 2019).

Another area of research will be the use of a similar algorithm for automatic tagging of videos. Using videos to assess the environment will significantly increase the coverage of the territory under consideration, the accuracy of observer positioning (if GNSS coordinates are not specified) due to the possibility of recognizing a much larger number of objects, the reliability of information (creating fake videos is much more complicated than photographs).

From the technical point of view, promising results for automated tagging of photos are shown by the architecture of the neural network PyramidNet. The main idea of this CNN is to gradually increase the number of channels in the convolution, and not several times, as happens in regular ResNet. Such an approach, theoretically, with the same model complexity as ResNet, should provide quality improvement (Han, D., et al., 2017).

The procedure of extracting / forecasting and checking the coordinates presented in this article allowed to increase the reliability of the extracted data from the posts of social networks and thereby ensure a higher quality of decision-making in emergency situations. However, it is worth noting the lack of versatility of this approach in relation to automatic tagging using neural networks. For maximum unification and full use of the proposed approach, it is necessary to significantly expand the database of photographs with images of objects from different continents and countries.

5. CONCLUSION

Any data obtained during and after an emergency can reduce all types of losses. The proposed two-step verification procedure makes it possible to assess the accuracy of the position of available data from social networks for subsequent mapping and spatial analysis. The proposed procedure can be generalized and

applied for almost any type of emergency, as well as for the usual refinement of cartographic information. Today, the interpretation of the semantic content of messages from social networks is a complex process and cannot be performed exclusively automatically with high accuracy and versatility, but it is fundamental for obtaining more reliable cartographic materials.

REFERENCES

- Bruno, A. & Xia, Junshi & Baier, Gerald & Yokoya, Naoto & Koshimura, Shunichi. 2019. Multi-Source Data Fusion Based on Ensemble Learning for Rapid Building Damage Mapping during the 2018 Sulawesi Earthquake and Tsunami in Palu, Indonesia. *Remote Sensing*. 11. 886. 10.3390/rs11070886.
- Al-Rfou, R., Kulkarni, V., Perozzi, B., Skiena, S., 2015. Polyglot-NER: Massive multilingual named entity recognition. *Proceedings of the 2015 SIAM International Conference on Data Mining, SIAM*, pp. 586–594.
- Barozzi, S., Marquez, J. L. F., Ravi Shankar, A., Pernici, B. et al., 2019. Filtering images extracted from social media in the response phase of emergency events. *16th Conference on Information Systems for Crisis Response and Management*, pp. 1–12.
- De Albuquerque, J.P., Herfort, B., Brenning, A., Zipf, A., 2015. A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. *International Journal of Geographical Information Science*, 29(4), pp. 667-689. doi.org/10.1080/13658816.2014.996567
- Francalanci, C., Pernici, B., Scalia, G., Zeug, G., 2018. Talking about places: Considering context in geolocation of images extracted from tweets. *GI-Forum 2018, Issue 1, Salzburg, July 2018, Short paper*, pp. 243–250.
- Havas, C., Resch, B., Francalanci, C., Pernici, B., Scalia, G., Fernandez-Marquez, J. L., Achte, T. V., Zeug, G., Mondardini, M. R. R., Grandoni, D., Kirsch, B., Kalas, M., Lorini, V., Ruping, S., 2017. E2mC: Improving Emergency Management Service Practice through Social Media and Crowdsourcing Analysis in Near Real Time. *Sensors*, 17(12), p. 2766.
- Hays, J., Efros, A. A., 2008. Im2gps: estimating geographic information from a single image. *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE*, pp. 1–8.
- Houston, J.B., Hawthorne, J., Perreault, M.F., Park, E.H., Goldstein Hode, M., Halliwell, M.R., Turner McGowen, S.E., Davis, R., Vaid, S., McElderry, J.A., Griffith, S.A., 2015. Social media and disasters: a functional framework for social media use in disaster planning, response, and research. *Disasters*, 39(1), pp. 1-22. doi.org/10.1111/disa.12092
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, pp. 1097-1105.
- Kumar S. S., Reddy S., Saran S. and Kocaman S., 2018. Design and Development of Trendyinsight Application Software for Location Based Social Media Trending. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLII-5, pp. 809–812, 10.5194/isprs-archives-XLII-5-809-2018
- Leichter A., Wittich D., Rottensteiner F., Werner M. and Sester M., 2018. Improved Classification of Satellite Imagery Using Spatial Feature Maps Extracted from Social Media. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLII-4, pp. 335–342. 10.5194/isprs-archives-XLII-4-335-2018
- Middleton, S. E., Kordopatis-Zilos, G., Papadopoulos, S., Kompatsiaris, Y., 2018. Location Extraction from Social Media: Geoparsing, Location Disambiguation, and Geotagging. *ACM Trans. Inf. Syst.*, 36(4), pp. 40:1–40:27.
- Migliaccio F., Carrion D., and Ferrario F., 2019. Semantic Validation of Social Media Geographic Information: A Case Study On Instagram Data for Expo Milano 2015. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLII-2/W13, pp. 1321–1326. 10.5194/isprs-archives-XLII-2-W13-1321-2019
- Nguyen, D. T., Ofli, F., Imran, M., Mitra, P., 2017. Damage assessment from social media imagery data during disasters. *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, ASO-NAM '17, ACM, New York, NY, USA*, pp. 569–576.
- Ravi Shankar A., Fernandez-Marquez J. L., Pernici B., Scalia G., Mondardini M. R. and Serugendo G., 2019. Crowd4EMS: A Crowdsourcing Platform for Gathering and Geolocating Social Media Content in Disaster Response. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLII-3/W8, pp. 331–340. 10.5194/isprs-archives-XLII-3-W8-331-2019
- Ryabchenko, N. A., Miroshnichenko, I. V., Gnedash, A. A., Morozova, E. V., 2016. Crowdsourcing systems on facebook platform: Experiment in implementation of mathematical methods in social research. *Journal of Theoretical Applied Information Technology*, 85(2).
- Tavra M., Racetin I. and Peroš J., 2019. Combining Social Media and Authoritative Data for Crisis Mapping: A Case Study of a Wildfire Reaching Croatian City Of Split. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLII-3/W8, pp. 415–420, 10.5194/isprs-archives-XLII-3-W8-415-2019