Research and analysis of messages of users of social networks using BigData technology

I A Rytsarev[1,2], A V Kupriyanov[1,2], D V Kirsh[1,2] and R A Paringer[1,2]

[1]Samara National Research University, Moskovskoe Shosse 34A, Samara, Russia, 443086
[2]Image Processing Systems Institute of RAS - Branch of the FSRC "Crystallography and Photonics" RAS, Molodogvardejskaya street 151, Samara, Russia, 443001

e-mail: rycarev@gmail.com

Abstract. In this paper is dedicated to the World Cup held in the city of Samara from June 15 to July 15, 2018. As part of the work, a multithreaded collection in real time was organized, filtering and processing messages from users of the social network Twitter within the host city and its surroundings from May 15 to August 15, 2018. Then, a study was conducted of the texts of user messages on the subject of the popularity of topics and the construction of a "word cloud". The second study was the construction of a diagram of the dynamics of the number of messages in different languages. As part of the work, modules for collecting, filtering and processing data using BigData technology were implemented.

1. Introduction

Currently, social networks are booming: every day their users generate hundreds of terabytes of media content: images and video. The analysis of such content is of great importance for many areas of business. For example, it is impossible to overestimate the impact of Internet marketing on the promotion of goods and services. However, clear understanding of user requests is essential to use these mechanisms effectively. The source of such information can be the materials published by users of social networks, as well as the shares and reposts by users and the entire communities. But in the period of any major events the population of online communities can vary greatly. In this paper, a comparison is made between the flow of messages before the World Cup, during and after it.

The task considered in the framework of this work is undoubtedly an urgent task, the solution of which is also of great scientific importance in the field of data analysis. In the article [1], a large dataset of geotagged tweets containing certain keywords relating to climate change is analyzed using volume analysis and text mining techniques such as topic modeling and sentiment analysis. In the article [2], the local and global term frequencies are computed through a bag-of-words (BOW) model. To remove the negative impact of high dimensionality on the global term weighting, the principal component analysis is adopted; thereafter the fuzzy c-means algorithm is employed to retrieve the semantically relevant topics from the documents. In the article [3], examine the long-term relationship between signals derived from nine years of unstructured social media microblog text data and financial market developments in five major economic regions. Employing statistical language modeling techniques we construct directional sentiment metrics. In the article [4], the authors propose a background clustering technology for discussion. Compared with the traditional methods, background future clustering keeps the constrains caused by data sparseness and spatio-temporal dependence off, and can be used for unpredictable activities discovery.In the article [5], the method of applying cross-references was considered to improve the accuracy of providing dictionaries in the task of calculating distributions between social communities based on text messages. In the article [6], the technology of processing large-scale text data on data collected from a social network was tested. The article [7] proposed a mathematical model for calculating the activity of users of social networks. Article [8] proposed a technology for normalizing text data. To capture the contextual meaning of tokens, authors create a neural word embeddings using word2vec trained on over a million social media messages representing a mix of domains and degrees of linguistic deviations.

2. Social network data collection

The Twitter social network was selected as a data source for this study. The reasons for this choice are as follows:

- the network provides open access to its data (no restrictions on accessing the server data);
- Twitter is the second most popular social network (after Facebook, which does not provide open access to its data) among users all over the world;
- Twitter is not a specialized network, which means it reflects the public opinion of a wider range of users [9].

The data collection from the Twitter social network can be carried out using the software products Apache Ambari and Flume, this method is described in more detail in [10]. However, it is often more convenient to develop a dedicated software product using standard libraries (twitter4j, tweepy, etc.) to collect the data using a number of filters [11, 12].

As part of this study, a Python software package was developed, containing an authorization module, a data collection module, and a filtration module. This software package allows to collect data by geolocation, by keywords, by user. The Twitter social network has a restriction in the form of a message limit that a client can receive during real-time monitoring. According to the documentation, this limit is 60 messages per second (this is about 1% of the average rate of tweets). A network of computers located in different cities was set up and cloud services were involved in order to avoid interruptions in the operation of the software complex, and to minimize message loss. Multiple unique authorization keys have been implemented in each copy. The designed software complex operates in real-time monitoring mode, and can make requests to receive information located on servers.

The geolocation filtering parameters were the coordinates of the city of Samara (the host city of the World Cup) in the form of an extended geobox (48.9700523344,52.7652295668, 50.7251182524,53.6648329274), which includes not only the city of Samara, but also the city of Togliatti (the training base of football players and the city where the tourists lived), airport Kurumoch and the settlements nearby the city of Samara.

More than 1,200,000 user messages were collected during the operation of the distributed network of the software complex nodes.

3. Analysis of the collected data using the BigData technology

The merging of the collected data, data processing and analysis using traditional approaches requires huge computational resources and takes a long time. For this reason, it was decided to use the BigData technology and the computing cluster for processing extra-large data available at the Samara University.

First of all, the data collected had to be merged. For this purpose, a data merging module was implemented using MapReduce technology. As a result of the module operation, we received more than 170,000 unique user messages.

The second task was the primary data processing. Streaming data obtained from social networks contains a lot of service information. Only the relevant data is important for further analysis; therefore, it is necessary to separate the service information from the relevant data. For this purpose, a json-response processing module has been implemented. This module uses the MapReduce technology for data structuring by way of arranging the data and excluding non-relevant and service data.

The third task was to analyze the data collected. The first study was the construction of a "tag cloud" for each of the three months separately. The results of the study are provided in Fig. 1, 2 and 3 respectively.

Figure 1. "Tag Cloud" for the period 15.05-14.06, 2018.

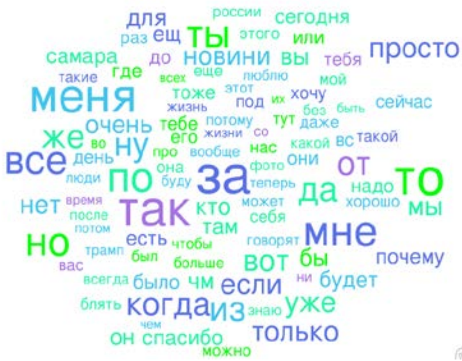

Figure 2. "Tag Cloud" for the period 15.06-14.07, 2018

Figure 3. "Tag Cloud" for the period 15.07-14.08, 2018.

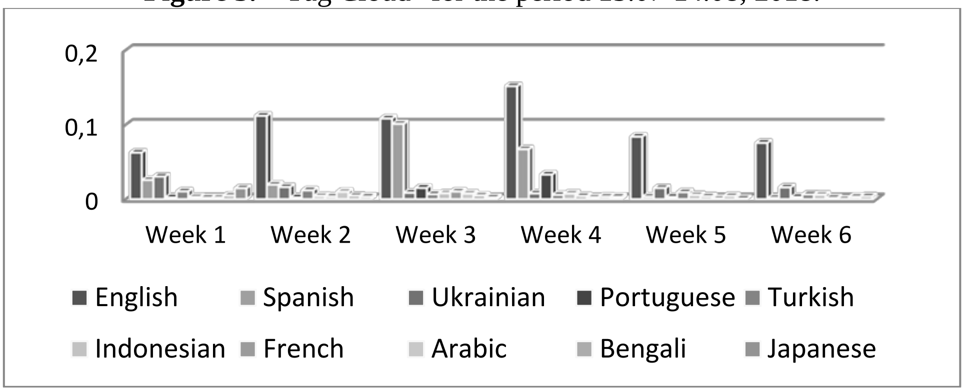

Figure 4. Distribution of messages by language for the period 11.06-22.07.2018.

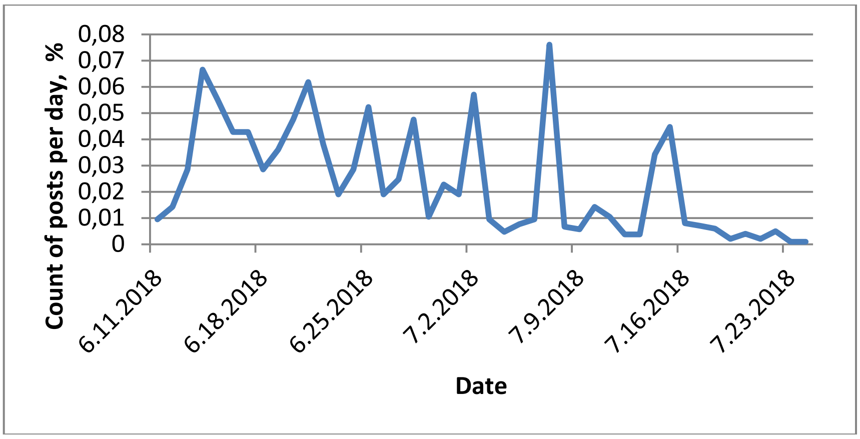

Figure 5. The distribution of the count of messages by day for the period 11.06-24.07 2018.

Thu 14 Jun 16:00	Russia	-	Saudi Arabia	Moscow
Fri 15 Jun 13:00	Egypt	-	Uruguay	Yekaterinburg
Fri 15 Jun 16:00	Morocco	-	Iran	Saint Petersburg
Fri 15 Jun 19:00	Portugal	-	Spain	Sochi
Sat 16 Jun 11:00	France	-	Australia	Kazan
Sat 16 Jun 14:00	Argentina	-	Iceland	Moscow
Sat 16 Jun 17:00	Peru	-	Denmark	Saransk
Sat 16 Jun 20:00	Croatia	-	Nigeria	Kaliningrad
Sun 17 Jun 13:00	Costa Rica	-	Serbia	Samara
Sun 17 Jun 16:00	Germany	-	Mexico	Moscow
Sun 17 Jun 19:00	Brazil	-	Switzerland	Rostov-on-Don
Mon 18 Jun 13:00	Sweden	-	South Korea	Nizhny Novgorod
Mon 18 Jun 16:00	Belgium	-	Panama	Sochi
Mon 18 Jun 19:00	Tunisia	-	England	Volgograd
Tue 19 Jun 13:00	Colombia	-	Japan	Saransk
Tue 19 Jun 16:00	Poland	-	Senegal	Moscow
Tue 19 Jun 19:00	Russia	-	Egypt	Saint Petersburg
Wed 20 Jun 13:00	Portugal	-	Morocco	Moscow
Wed 20 Jun 16:00	Uruguay	-	Saudi Arabia	Rostov-on-Don
Wed 20 Jun 19:00	Iran	-	Spain	Kazan
Thu 21 Jun 13:00	Denmark	-	Australia	Samara
Thu 21 Jun 16:00	France	-	Peru	Yekaterinburg
Thu 21 Jun 19:00	Argentina	-	Croatia	Nizhny Novgorod
Fri 22 Jun 13:00	Brazil	-	Costa Rica	Saint Petersburg
Fri 22 Jun 16:00	Nigeria	-	Iceland	Volgograd
Fri 22 Jun 19:00	Serbia	-	Switzerland	Kaliningrad
Sat 23 Jun 13:00	Belgium	-	Tunisia	Moscow
Sat 23 Jun 16:00	South Korea	-	Mexico	Rostov-on-Don
Sat 23 Jun 19:00	Germany	-	Sweden	Sochi
Sun 24 Jun 13:00	England	-	Panama	Nizhny Novgorod
Sun 24 Jun 16:00	Japan	-	Senegal	Yekaterinburg
Sun 24 Jun 19:00	Poland	-	Colombia	Kazan
Mon 25 Jun 15:00	Uruguay	-	Russia	Samara
Mon 25 Jun 15:00	Saudi Arabia	-	Egypt	Volgograd
Mon 25 Jun 19:00	Iran	-	Portugal	Saransk
Mon 25 Jun 19:00	Spain	-	Morocco	Kaliningrad
Tue 26 Jun 15:00	Denmark	-	France	Moscow
Tue 26 Jun 15:00	Australia	-	Peru	Sochi
Tue 26 Jun 19:00	Nigeria	-	Argentina	Saint Petersburg
Tue 26 Jun 19:00	Iceland	-	Croatia	Rostov-on-Don
Wed 27 Jun 15:00	South Korea	-	Germany	Kazan
Wed 27 Jun 15:00	Mexico	-	Sweden	Yekaterinburg
Wed 27 Jun 19:00	Serbia	-	Brazil	Moscow
Wed 27 Jun 19:00	Switzerland	-	Costa Rica	Nizhny Novgorod
Thu 28 Jun 15:00	Japan	-	Poland	Volgograd
Thu 28 Jun 15:00	Senegal	-	Colombia	Samara
Thu 28 Jun 19:00	England	-	Belgium	Kaliningrad
Thu 28 Jun 19:00	Panama	-	Tunisia	Saransk
Fri 29 Jun				
Sat 30 Jun 15:00	C1	-	D2	Kazan
Sat 30 Jun 19:00	A1	-	B2	Sochi
Sun 01 Jul 15:00	B1	-	A2	Moscow
Sun 01 Jul 19:00	D1	-	C2	Nizhny Novgorod
Mon 02 Jul 15:00	E1	-	F2	Samara
Mon 02 Jul 19:00	G1	-	H2	Rostov-on-Don
Tue 03 Jul 15:00	F1	-	E2	Saint Petersburg
Tue 03 Jul 19:00	H1	-	G2	Moscow
Wed 04 Jul				
Thu 05 Jul				
Fri 06 Jul 15:00	W49	-	W50	Nizhny Novgorod
Fri 06 Jul 19:00	W53	-	W54	Kazan
Sat 07 Jul 15:00	W55	-	W56	Samara
Sat 07 Jul 19:00	W51	-	W52	Sochi
Sun 08 Jul				
Mon 09 Jul				
Tue 10 Jul 19:00	W57	-	W58	Saint Petersburg
Wed 11 Jul 19:00	W59	-	W60	Moscow
Thu 12 Jul				
Fri 13 Jul				
Sat 14 Jul 15:00	L61	-	L62	Saint Petersburg
Sun 15 Jul 16:00	W62	-	W61	Moscow

Figure 6. World Cup World Cup schedule 2018.

It can be seen in Figures 1, 2 and 3, that the filling of the “clouds” changed dramatically with the beginning of the World Cup in the Samara Region. Taking into account the results of the previous study, the decision was taken to look at the dynamics of changes in the number of messages in different languages in the next study. The analysis of the language of writing a message was carried out on the basis of data provided by the Twitter social network in json-response. A 7-day period was selected as an analysis period. The results are provided in Fig. 4.

As it can be seen from Fig. 4, the number of messages in the languages other than Russian varied in accordance with the football games held in the city of Samara. It started to increase a week before the beginning of the tournament, then the number of messages remained at the same level throughout the tournament and then dropped to the values close to zero due to the departure of delegations.

Additionally, we construct a graph of user activity by day (Figure 5) and relate it to the schedule of games (Figure 6).

As can be seen on the graph, the peak of user activity fell on the days of the games at the Samara Arena stadium. On the days of the games at other stadiums, user activity was lower than on the days of matches in Samara. On the other days, the activity did not exceed 0.01 percent (the exception was the match days for the third place and the final). The peak of activity came on 07.07.18 when the matches at the Samara Arena and Russia - Croatia took place. After the end of the World Cup, user activity has declined sharply.

4. Conclusion

In this paper, a study was conducted of the activity of the users of the social network Twitter of the Samara region, as well as the activity of the guests of the 2018 World Cup who came to support the national teams in the city of Samara. The study showed that a major event can drastically change the main subjects of messages and dictionaries of frequently used words in social networks. From this it follows that when analyzing social network data in the period of any major events, it is necessary to apply methods of reactive data analysis, as well as take into account user profile information for correct data processing (collect separate statistics, since it will be completely different from statistical data which was collected before the event).

5. References

[1] Dahal B, Kumar S A P and Li Z 2019 Topic modeling and sentiment analysis of global climate change tweets *Social Network Analysis and Mining* **9(1)** 24

[2] Rashid J, Shah S M A and Irtaza A 2019 Fuzzy topic modeling approach for text mining over short text *Information Processing & Management* **56(6)** 102060

[3] Groß-Klußmann A, König S and Ebner M 2019 Buzzwords build Momentum: Global Financial Twitter Sentiment and the Aggregate Stock Market *Expert Systems with Applications*

[4] Zhu C, Du J 2018 Background feature clustering and its application to social text *Information Processing Letters* **136** 44-48

[5] Rytsarev I A, Kupriyanov A V, Kirsh D V and Liseckiy K S 2018 Clustering of social media content with the use of BigData technology *Journal of Physics: Conference Series* **1096(1)** 012085.

[6] Blagov A, Rytcarev I, Strelkov K and Khotilin M 2015 Big Data Instruments for Social Media Analysis *Proceedings of the 5th International Workshop on Computer Science and Engineering* 179-184

[7] Rytsarev I, Blagov A 2017 Creating the Model of the Activity of Social Network Twitter Users *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)* **9(1-3)** 27-30

[8] Kusumawardani R P, Priansya S and Atletiko F J 2018 Context-sensitive normalization of social media text in bahasa Indonesia based on neural word embeddings *Procedia computer science* **144** 105-117

[9] Rytsarev I A, Blagov A V 2017 Development and research of algorithms for clustering data of super-large volume *CEUR Workshop Proceedings* **1903** 80-83

[10] Mikhaylov D V, Kozlov A P and Emelyanov G M 2016 Extraction of knowledge and relevant linguistic means with efficiency estimation for the formation of subject-oriented text sets *Computer Optics* **40(4)** 572-582 DOI: 10.18287/2412-6179-2016-40-4-572-582
[11] Rytsarev I A, Kirsh D V and Kupriyanov A V 2018 Clustering of media content from social networks using BigData technology *Computer Optics* **42(5)** 921-927 DOI: 10.18287/2412-6179-2018-42-5-921-927.
[12] Kropotov Y A, Proskuryakov A Y and Belov A A 2018 Method for forecasting changes in time series parameters in digital information management systems *Computer Optics* **42(6)** 1093-1100 DOI: 10.18287/2412-6179-2018-42-6-1093-1100

Acknowledgments
This work was financially supported by the Russian Foundation for Basic Research under grant # 19-29-01135, # 18-37-00418, # 17-01-00972 and by the Ministry of Science and Higher Education within the State assignment to the FSRC “Crystallography and Photonics” RAS No. 007-GZ/Ch3363/26 (theoretical results).