ANALYSIS OF USER LOGS FROM THE E-LEARNING SYSTEM AND OPINION MINING IN THE FIELD OF PROGRAMMING

Katarina Karić[1], MSc; Marija Blagojević[1], PhD
[1] University of Kragujevac, Faculty of Technical Sciences, Čačak, SERBIA, katarina.karic@ftn.kg.ac.rs
[1] University of Kragujevac, Faculty of Technical Sciences, Čačak, SERBIA, marija.blagojevic@ftn.kg.ac.rs

***Abstract**: With the development of the Internet and information and telecommunication technologies, there is an increase of the amount of data stored daily in various databases. As there is a need to obtain different types of information from them, there is a development of technologies, i.e. tools for „mining“ according to the given data. Data Mining is the process of „mining“ large databases and extracting new and useful information. The goal is to discuss the information obtained by applying software tools, ie. by analyzing user logs on the course „Introduction to Programming“, they find and recognize certain predictions of some outcomes or users' behaviors in the field of usage mining, as well as to perform an analysis of opinions, ie. opinion mining on „Programming“.*

***Keywords:** Data mining, Usage mining, Log, Opinion mining, Programming*

1. INTRODUCTION

With the development of the Internet and information and telecommunication technologies, there is an increase in the amount of data stored daily in various databases, starting from government institutions, companies, and to various websites. As there is a need to obtain different types of information from them, there is a development of technologies, ie tools for "mining" according to the given data. Data Mining is a process of "mining" large databases and extracting new and useful information that can contribute to better and more successful business (if it is in this area) [1].
Web Mining, as a part of data mining, ie "mining" according to Internet data, especially on the World Wide Web, is an area that has been quite common lately and is attracting more and more attention. The goal is to use the collected data, ie information in the best possible way, and get results that will help in finding and recognizing certain patterns or patterns, predicting some outcomes or behaviors, and many other things in different spheres.
Web usage mining, as a part of web mining related to use, aims to identify patterns in user behavior in order to better understand and adapt the website to the user, through the analysis of logs as one of the techniques of Web usage mining. Logs can be defined as computer-generated records that record specific activities, which can provide a variety of information about users and their behavior, as well as security and many other issues.
Opinion Mining, is a concept of web content mining, ie. content mining and aims to find useful information on a particular topic (for example comments on Twitter, blogs, etc.) from users' opinions. An opinion is a personal belief or assessment of a subject. Opinion mining uses Data Mining, Artificial Intelligence and Natural Language Processing (NLP) techniques to find, interpret and extract data and opinions from information that can be found on the Internet [2].
The aim of the research conducted in this paper is to analyze user logs from the e-learning system, ie. from the course "Introduction to Programming", in order to identify patterns of user behavior in the course, as well as analyzing opinions, ie. mining opinions on the topic of "Programming".

2. RELATED RESEARCH

Many related research has been conducted in the areas of web usage mining and opinion mining in order to identify user behaviors and analyze opinions in different areas.
Through a related study [3] conducted at the University of Bangladesh, an analysis of the Learning Management System (LMS) for two subjects, conducted over six weeks, was conducted to find the relationship between student access to LMS and overall performance through log analysis technique and statistics. The results of the research show that students with low access received a bad grade, while access from student accommodation was higher than access from home.
In [4], the authors present some of the web usage tools, such as WebSIFT and WebLogMiner, which were used for analysis to improve the e - learning environment. The results showed that one of the main advantages of the research is

the ability of lecturers to better assess the process of mastering student material, which is extremely important in the field of education.
Through research [5], the authors used thought analysis techniques to analyze the concept of "Information Technology". The results indicate the possibility of successful application of the mentioned techniques, with the possibility of future research development in the domain of "neutral" attitudes, ie. opinions.
The authors in [6] analyze students' attitudes, by processing the text using the Rapid Miner tool. A related study [7], which deals with the analysis of unstructured Twitter data, was conducted to determine the accuracy of the results obtained with structured and unstructured data. It was found that the results of the analysis are more accurate, if the data for the analysis itself are structured.

3. RESEARCH TOOLS AND METHODOLOGY

As already mentioned in the introductory part, for the purposes of this research, and thus further improvement and refinement of the course itself, as well as its materials and topics, an analysis of the log file from the e-learning system, Faculty of Technical Sciences in Cacak (access address: https://eucenje.ftn.kg.ac.rs/course/view.php?id=116), course: Introduction to programming in .xlsx format. It stores data collected in the period from October 8, 2017 to July 8, 2020. The file size is almost 3MB, due to the long period during which the data was collected, as well as due to the large number of accesses. Initially, this data set contained information (presented in the form of columns) on: Date and time of access of course users (in the form of one column), then columns - "Full user name", "Affected user", "Component" and "Event context" "," Event name "," Origin "," Description "," IP address "and" Activity "column. By modifying the column "Full user name" in "Role of the user", so that instead of each name and surname of the user, his role is written in the form: student, teacher, associate or administrator, thus providing the possibility of analysis by roles in the tool. Also, the column "Date and time of access" is divided into two separate columns, because it allows obtaining information about the time of the most frequent access of users by hours or minutes, as well as about months, days and years. The columns "Description" and "IP address" were not of great importance for this type of analysis, so they were removed. The "Event Context" column has been modified to indicate only user access to the course, forum, directory, file, etc. The "Activity" column was added as a type of value for later analysis and indicates that each row is an action for both students and professors.
The "PowerBI" tool [8], which is a product of the company "Microsoft", was chosen for the analysis of the log file. Microsoft PowerBI is a business intelligence platform that provides business users with the ability to collect, analyze, visualize, and share data.
The analysis of opinions on the extremely popular topic of "programming" was conducted using the "RapidMiner" tool. The data were taken from the "Social Searcher" (social media web browser). Social Searcher is a free social media search Engine. It allows to search for content in social networks in real-time and provides deep analytics data in the some field such as in this paper - field of programming. Tool performs a search real-time social media (on Twitter, Google+, Facebook, Youtube, Instagram, Tumblr, Reddit, Flickr, Dailymotion and Vimeo) by keywords, posts history management, data access with pagination and comprehensive analytics reports for periods [9]. RapidMiner [10] is an open source software platform for Opinion Mining. Provides an integrated environment for machine learning, "Data Mining", "Text Mining", business analytics. It is used for business and commercial purposes, as well as for research, education, training, and supports all steps of the data mining process.

4. RESULTS AND DISCUSSION

This chapter is divided into two parts: first, it provides the results and discussion of user log analysis to determine the pattern of behavior in the e-course introduction to programming (Figures 1-6); the other provides the results of mining opinions on the topic of “Programming” (Figures 7-9).

4.1. Results of analysis and visualization of user logs

Figure 1 graphically shows the activities, ie use and attendance of the Introduction to Programming course, according to the roles of the course users (Students, Associates, Teachers, Administrator). The analysis of the obtained results concludes that the highest attendance of the course is by students (76,278), which is very positive because the system and the course are intended primarily for their teaching, acquiring knowledge, skills and achieving results in the field of programming. After that, the largest share of attendance is occupied by associates (3,047) and teachers (2,449).
To follow the time period in which the highest attendance on the course was expressed, it was necessary to perform an analysis of data by access time. The obtained results in Figure 2 show that the time of access of all users is most pronounced at 19 hours (highest presence) and 22 hours, and activity is also expressed in the time period between 12 hours and 15 hours. The information obtained can be important for publishing content, as well as defining certain activities and tasks just at the time when traffic is highest. Significantly lower attendance is expressed in the morning (from 7 am to 9 am).

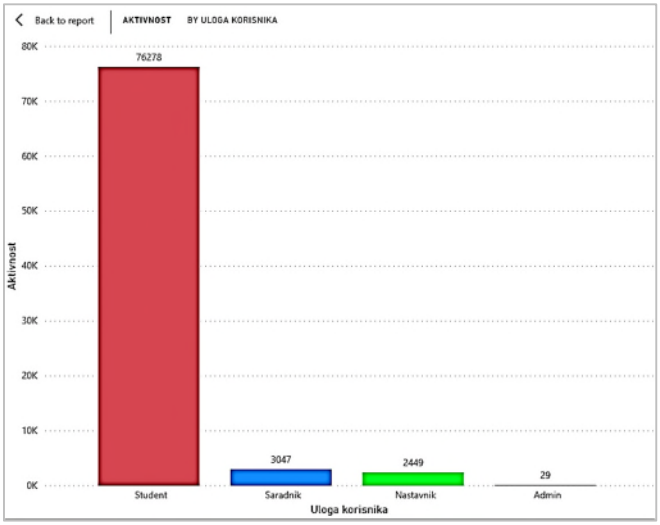

Figure 1: Total number of activities on the course Introduction to programming by user roles

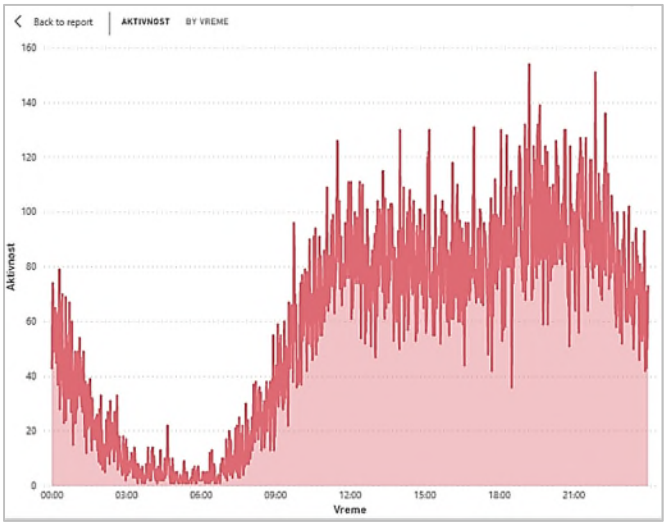

Figure 2: Course user activity by access time

Considering that the course Introduction to Programming is a subject from the winter semester at the Faculty of Technical Sciences in Čačak, an analysis of logs was conducted according to the date of user access, ie. months, to determine whether users are most present during the course period (Figure 3). The obtained results show that the highest activity was expressed in November (22,277) and December (17,908), and then in October (15,034), because this is the period when the material is constantly processed and published, as well as the colloquia and exam deadlines, ie checks. knowledge is in that period. The smallest approach is in August, since it is the time of collective vacation.

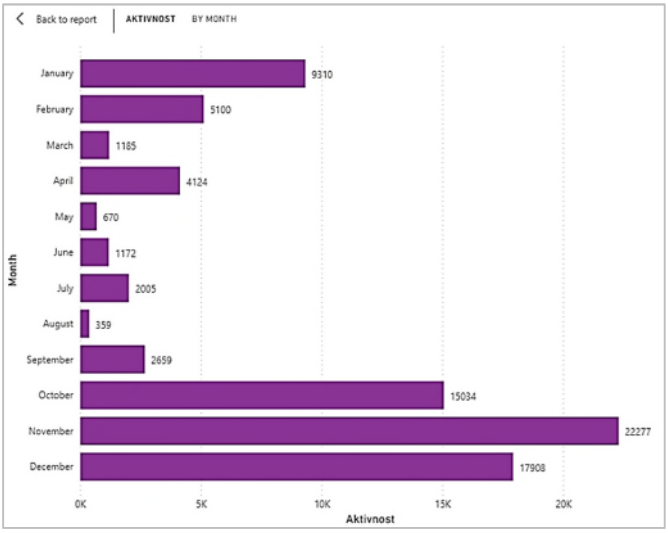

Figure 3: Total attendance at the course by date (months) of access

Figure 4 presents the percentage of access to the contents of the course during the 2019/2020 school year. The results have shown that the highest percentage of access is next to the course itself, on the Forum: Notices 2019/2020 (12.37%), and the lowest towards the Directory: Lectures 2019/2020 (8.57%). In order to increase the percentage of attendance at the lecture directory, one of the ideas would be that for each teaching unit, ie. lesson is posted by notices on the forum with web addresses (path) to a particular lecture, which leads to greater access to lectures, as well as to faster and better acquisition of knowledge and productivity in the field being studied

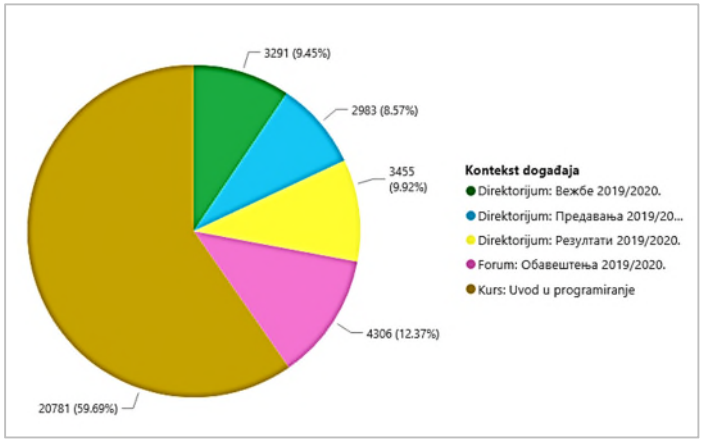

Figure 4: Percentage share of access to course content during the 2019/2020 school year

The results obtained in Figure 5 show that the time of student access is most pronounced at 7 pm (highest presence) and 10 pm, and activity is also expressed in the time between 12 and 3 pm.

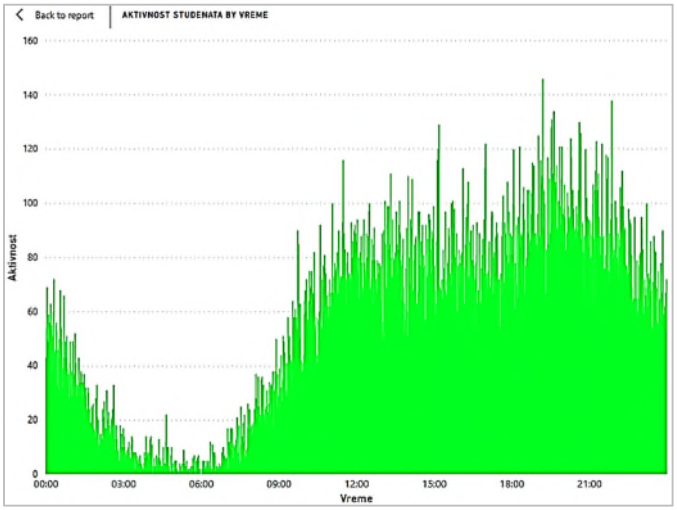

Figure 5: Student activity by course access time

The results obtained in Figure 6 show that the time of the lecturer's approach, ie. teachers and associates most pronounced in the period of 14 hours (maximum presence) and around 12 hours, activity is also expressed in the period of about 22 hours.

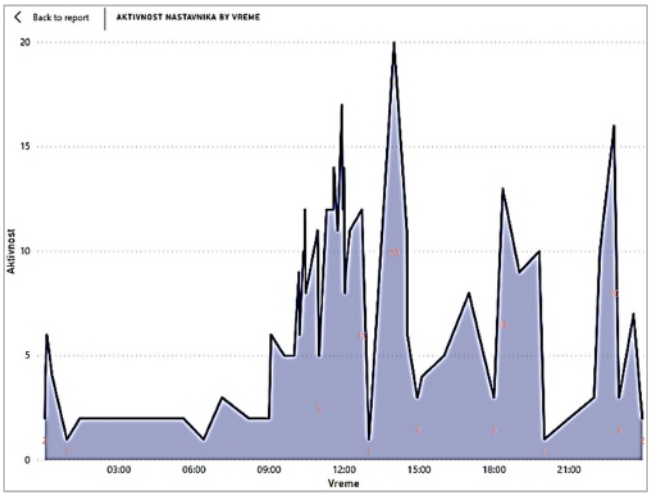

Figure 6: Teacher activity by course access time

Attendance times are largely the same, which is a positive outcome, however, in order to achieve a greater presence of students as the course itself, ie. contents and material, lecturers could define certain activities in the form of work assignments, quizzes, questionnaires, notices, etc. in the period around 7 pm (because then the most prominent presence is with students).

4.2. Results of the analysis of opinions on the topic "Programming"

Figure 7 illustrates the total number of positive (352), negative (8), neutral (8) comments related to the topic of “Programming”. It was determined that the largest number of neutral, then positive and negative comments - whose number is equal. As the analysis of the results shows that most of the comments was be neutral, there is a need to implement the sentiment analysis in future work. Sentiment analysis deals with the interpretation of sarcasm, irony, metaphors, i.e. the meaning of sentences, while opinion mining deals with separating the positive and negative parts of a sentence and opinion as a whole. So, the challenge in the analysis of sentiment is also the domain language dependence, so that identical expressions can carry a completely different sentiment, i.e. meaning in different environments / domains.

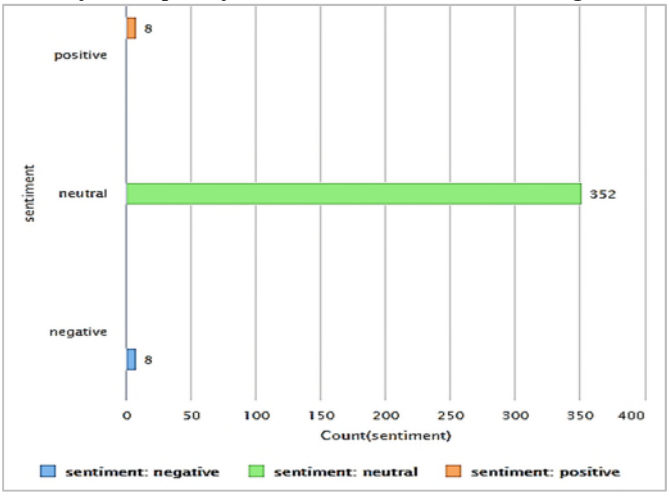

Figure 7: Total number of neutral, positive and negative comments

Figure 8 presents the total number of types of comments, ie posts grouped by social networks. It was determined that the largest number of comments in the form of photos is present on the "Flickr" network, then comments via the link are most present on the "Reddit" network, while opinions are set in the form of the most represented status on the social network "Twitter", which was expected. As Instagram is one of the most popular networks today, the promotion of this area - programming in the form of sponsored posts/comments by type of photo and video would significantly contribute to a larger number of interested users, whose area of interest is programming.

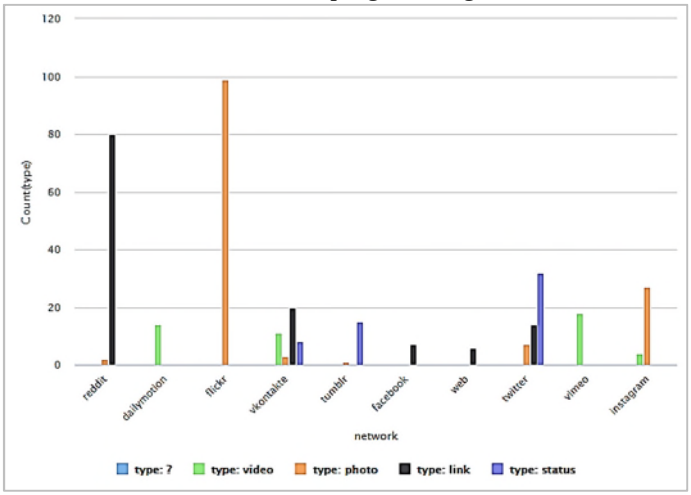

Figure 8: The total number of comment types grouped by social networks

Figure 9 shows the number of all comments: positive, negative and even neutral grouped by social networks: The largest number of positive comments on the topic of programming is on the network "Vkontakte" (4), while the number of negative (5) on the network "Reddit". The largest number of neutral comments (97) is present on the "Flickr" network. The results obtained by this analysis also indicate the analysis of sentiment in future work, in order to analyze the comments interpreted as neutral in detail, which would lead to obtaining more information about the opinion of users on this topic.

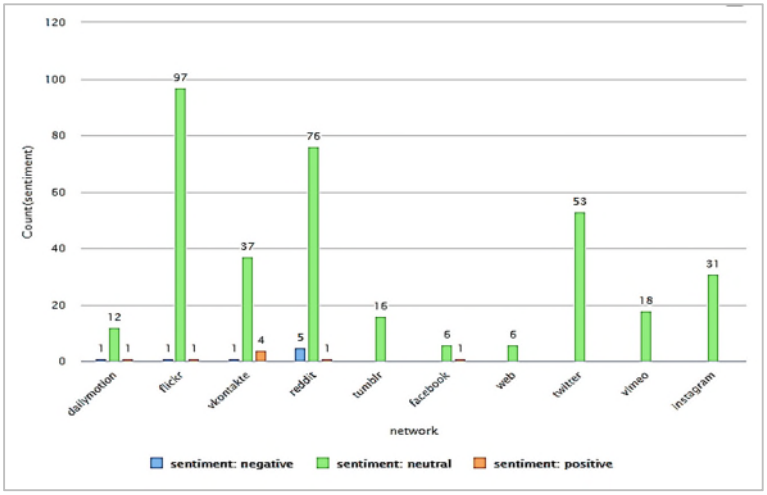

Figure 9: The total number of neutral, positive and negative comments on social networks

5. CONCLUSION

The use of data mining techniques, i.e. web mining, has numerous advantages, from obtaining information relevant to the read content and structure in certain areas of research, to forming a "pattern" of behaviour of users who access them and use certain systems. This type of analysis can help a lot in improving the functioning of the site, understanding the behaviour and needs of visitors, i.e. users, as well as improving the services or products offered, whether it is an "online" or "offline" business. The use of the "Usage Mining" technique is one of such techniques that aims to obtain information of importance related to the use, i.e. user interaction on the Web, i.e. when it comes to the analysis conducted in this paper, determining user behaviour in the course "Introduction to Programming". The aim of the analysis of user logs was to determine how often course participants access the course, at what time and with what content. Such information can be useful to adjust both the publication of the content and the notification of the participants to the period when the highest presence was determined (19 hours in the research conducted in this paper).

Future innovations of the course would be that teachers encourage participants / students to interact and attend the course, both by adding new, current, and interactive content in the field of programming in the form of a file, and by starting topics related to course issues in forums provided for and to more frequent activities in the form of homework, quizzes, etc., to significantly increase the percentage of course attendance and the number of participants.

Organizations can, based on the analysis of public opinion, find out what is the placement of their products and services, what are the effects of promotional campaigns, what needs to be changed, improved, innovated ... When it comes to the research conducted in this paper, the obtained data are relevant to the current topic - programming, it was concluded that for this topic most of the comments are neutral, which leads to future research development in the field of sentiment analysis, ie. analysis of opinions that may contain some form of irony or sarcasm.

The opinion of users is very important, and it is often needed because it is already much appreciated and implemented in the topic and content, whether it is a course within the site, or a product or service, as users or visitors are on which the success of all the above depends.

ACKNOWLEDGEMENTS

This study was supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia, and these results are parts of the Grant No. 451-03-9/2021-14/200132 with University of Kragujevac - Faculty of Technical Sciences Čačak.

REFERENCES

[1] Blagojević, M. Primena Veb majninga u obrazovanju: 3. Međunarodna konferencija Tehnika i informatika u obrazovanju, 2010: 739-742.

[2] Opinion mining for social media, [Online]. Available at: https://www.slideshare.net/dianamaynard/prague2012-opinionmining . [Accessed 01.08.2021.]

[3] Chanchary F. H., Haque I., Khalid M. S. Web usage mining to evaluate the transfer of learning in a web-based learning environment: First International Workshop on Knowledge Discovery and Data Mining (WKDD 2008): IEEE, 2008., Available at: https://ieeexplore.ieee.org/document/4470388

[4] Zaiane O., Web usage mining for a better web-based learning environment, report, Alberta, Canada, Department of Computing Science, University of Alberta Edmonton, 2001. , Available at: https://era.library.ualberta.ca/items/0a182195-ce39-4b5d-a1c1-291ed91a0f36/view/336c2a34-b149-4d9d-95a1-ad78c08ee35c/TR01-05.pdf
[5] Jovičić A., Plašić J., Blagojević M., Ranković A. Analysis of term in information technologies using sentiment mining techniques: 6th International conference on Knowledge management and informatics, 2020: 1-6. Available at: http://kmi.vtsns.edu.rs/KMI_2020/radovi/1-KMI_Informatika/KMI_informatika-1.pdf
[6] Blagojević M., Kuzmanović B. Text processing in analysis of students' attitudes: International conference on information technology and development of education (ITRO 2016), 2016: 97-100. Available at: http://www.tfzr.uns.ac.rs/itro/Zbornik%20ITRO%202016.pdf
[7] Kharde V., Sonawane, S. Sentiment analysis of Twitter data: A survey of techniques: International Journal of Computer Applications, Volume 139, No.11; 2016: 5-15.
[8] Microsoft PowerBI, [Online]. Available at: https://searchcontentmanagement.techtarget.com/definition/Microsoft-Power-BI . [Accessed 01.08.2021.]
[9] Socialsearcher, [Online]. Available at: https://www.social-searcher.com/ . [Accessed 01.08.2021.]
[10] Wikipedia. Rapidminer, [Online]. Available at: https://en.wikipedia.org/wiki/RapidMiner . [Accessed 01.08.2021.]