UDC: 004.93:60:620 https://doi.org/10.15407/jai2023.01.147

# COLLECTING AND ANALYZING NEWS FROM NEWSPAPER POSTS IN FACEBOOK USING MACHINE LEARNING

**I. Mysiuk**[1], **R. Mysiuk**[2], **R. Shuvar**[3]
[1,2,3]Ivan Franko National University of Lviv, Ukraine
1, University st., Lviv, 79000
iruna.musyk8a@gmail.com; mysyukr@ukr.net.

[1]https://orcid.org/0000-0002-3641-4518
[2]https://orcid.org/0000-0002-7843-7646
[3]https://orcid.org/0000-0001-6768-4695

**Abstract.** Many people use social networks to spend their free time. News, especially at the time of great world changes, began to gain considerable popularity. Washington Post, New York Times, Time, Reuters, Forbes are among the most famous global newspaper publications. An average analyst can spend up to 40 hours a week collecting information about competitors and researching the most popular posts. According to the conducted research, an average of 40 new posts with news per day. The data processing process can be automated using modern information tools to facilitate the routine work of analysts. To analyze the target audience and reach, it is worth considering the text of the message, the number of likes, comments and links. This information was obtained using the Selenium automated web page testing tool using the Java programming language. The time spent on collecting data in the described way from four newspaper editions amounts to approximately 12 hours. The Tensorflow library using the JavaScript programming language is applied to the collected information. Based on information about the number of shares, comments, likes, frequency of news posts, an analysis was carried out using machine learning algorithms. Based on the clustering data, we can observe such a tendency that posts with a large number of likes receive a large number of comments and vice versa. An analysis of the most active hours of users in the network based on news posts is performed. As a result, the highest activity is observed at least three times a day, namely: in the morning hours from 9:00 to 11:00, in the lunch time of the day from 12:00 to 15:00 and in the evening time period from 18:00 to 20:00. This trend is due to the work schedule of most employees during the working week. The resulting statistical information in the work can be used for other content or user behavior in social networks.

**Keywords**: big data, machine learning, social networks, data analytics, automated data collection, data processing.

## Introduction

With the beginning of the "meta" era of information and data, social networks are gaining more and more popularity and becoming an integral part of life [1]. Especially news in social networks is becoming an extremely relevant and fast way of obtaining information. Among the most used social networks today are Facebook, Twitter, Instagram and TikTok. News information is mostly texts, and Facebook is the most convenient for adding such content. Usually, working with especially news data takes a lot of time. This process is monotonous and requires new research results for analysts. This approach can be automated using modern information technologies.

## Literature review

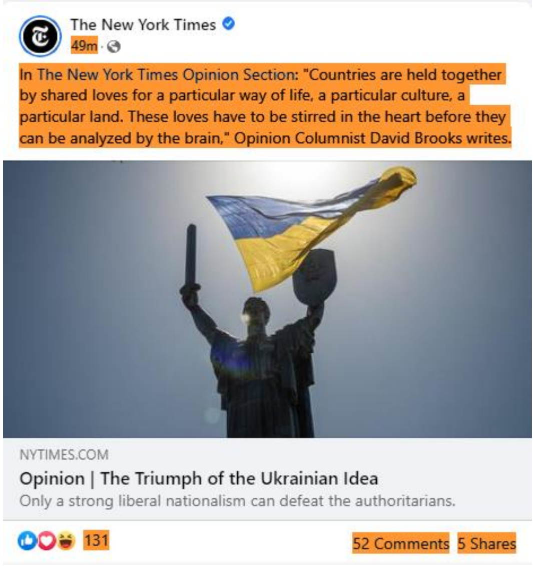
Machine learning is used in almost all spheres of life. Some aspects of application in the process of data analysis are described [1].

In Figure 1, the highlighted text shows which main parameters are read for analysis. This post is from the New York Times [2] Facebook page. In the same way, data is obtained from other investigated newspapers such as Washington Post [3], Reuters [4] and Forbes [5].

The work [6–8] describes the construction of a tool for collecting information based on Selenium, but not for social networks.

## Methods and Materials

As a method of data collection, the method of reading them from a web page is chosen. Especially with code reuse, it is convenient to do this when the data has the same format. For example, as it is presented in the posts of the social network Facebook. Mostly, for better coverage, newspapers add news to social networks, as often as to their own web pages.



The New York Times
49m ·

In The New York Times Opinion Section: "Countries are held together by shared loves for a particular way of life, a particular culture, a particular land. These loves have to be stirred in the heart before they can be analyzed by the brain," Opinion Columnist David Brooks writes.

NYTIMES.COM
Opinion | The Triumph of the Ukrainian Idea
Only a strong liberal nationalism can defeat the authoritarians.

131 52 Comments 5 Shares

Fig.1. An example of the information that is used for the analysis is from the Facebook news page of the New York Times

Therefore, the Selenium automated testing tool is used to read the data and generate the test data set [6]. This framework allows you to simulate the user's work in web browsers, but performing monotonous actions in an automatic order [7]. In this work, the Google Chrome browser is selected.

The process of data normalization is performed using the Java programming language, to which the above-mentioned Selenium tool is added as a library and with a set of other built-in capabilities.

The last stage for the analysis is the visualization of the received information. The library linking approach is convenient to use. As you know, libraries are a set of ready-made functions. This allows you to reduce duplication and use ready-made parts of the implemented code [8].

Among the ones used in this work is the Tensoflow library, which is adapted to most programming languages. This tool is popular for machine learning. The browser is a universal and cross-platform tool for displaying data. And in recent years, websites have become particularly relevant due to the pandemic and the situation in the world.

The Javascript programming language allows you to link libraries. In particular, tensorflow.js is used in the work [9, 10].

The analysis process consists in reading data from an intermediate Comma Separated Values (CSV) file and displaying the results.

Data ready for processing is sent to the website for visualization using the Canvas component in HTML.

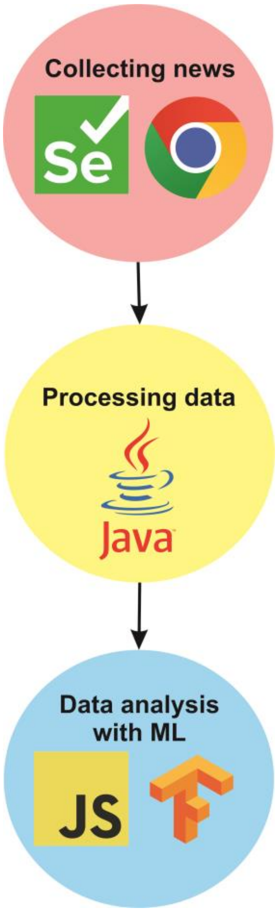


Fig.2. The process of processing news from posts with the tools used at each stage

Data analysis can be performed on the basis of certain statistical estimates and using existing algorithms [11].

From the received information about posts, it is possible to form certain trends and regularities [12]. The number of comments, shares and likes depends not only on the text of the news, but also on the time of its publication [13]. This is due to the active working hours and free time of readers.

### Problem Formulation

The main problem for news analysis is the large amount of information. On average, 40 news stories are distributed per day in each data source. Processing such a large amount of information requires a lot of time and routine work.

With the help of the implemented program, you can quickly analyze both current and historical information for a certain period of time in the past.

Obtaining data from one source in this way takes several hours. In addition, it is possible to parallelize information from several newspaper publications at the same time.

## Objective

The main goal of the work is to propose an approach for quick analysis of information over a certain period of time.

The main tasks in the work are the application of machine learning algorithms to analyze information about news and their popularity.

## Results

It is not always possible to evaluate the correspondence of the comments of certain posts to the number of likes. Usually, users leave comments under the most disturbing questions and the most relevant topics.

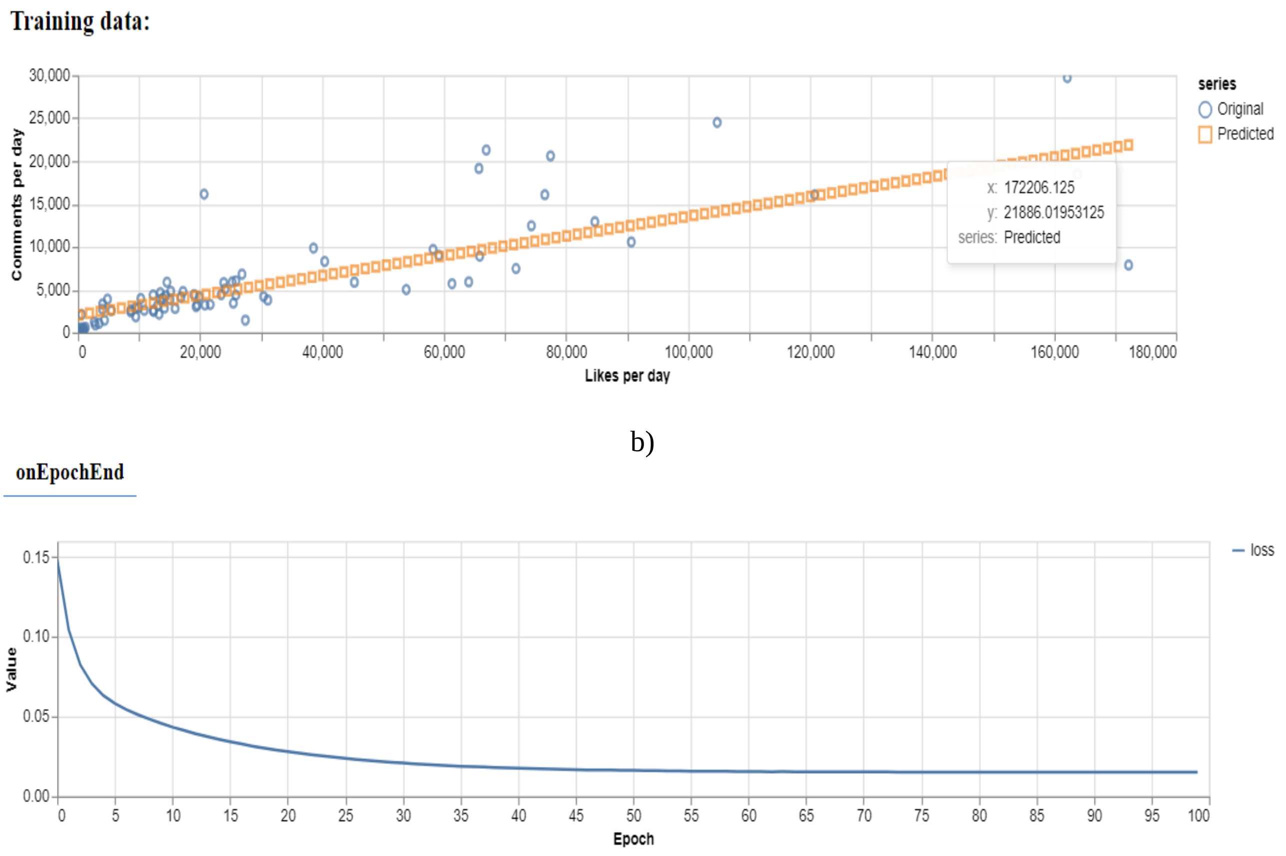
To predict statistics, you can use training based on a certain trend in the ratio of the number of likes to the number of comments.

The data is collected from different magazines from the same period of time. The number of input data is 100 records. Despite the rather small data set, certain trends can be seen.

The number of comments is less than the number of reactions to a post by an average of 5 times. Moreover, as the number of reactions increases, the number of comments decreases. This is probably related to people's tendency to express their preferences in the simplest way. In the case of a post, people's reactions are reflected most quickly through likes.

Figure 3 shows the expected ratio of the number of likes to the number of comments. But below the graph shows the learning process. To train the model, the batch size is set to 5, and the number of epochs to 100.

a)



b)

Fig. 3. The result of learning a) predicting the number of comments to the number of likes in a news feed in posts; b) process of study with batch size 5 and epochs 100

The k-means method can be used to cluster data. The mentioned algorithm is applied to this same set of data. As can be seen in Fig. 4, in general, it can be divided into 3 clusters. The first with him is the largest, it is where the number of comments and likes is small. The figure shows the coordinates of the center of the cluster. In the first case, the number of likes is equal to 15406, and the number of comments is 4320. The numbers were rounded in the output to whole values. The next one on the right is the average cluster by the amount of data that got into it. These are usually quite interesting posts among users. The last post on the right in the picture is the most popular and active post.

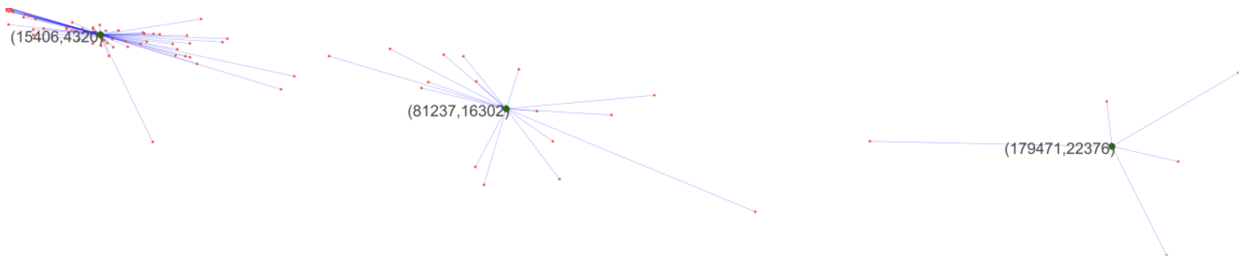


Fig.4. The result of clustering the number of likes and comments

The most popular posts in Forbes are related to internal events: Hayes Barnard and green energy (3900 likes) likes, Amazon show and Nielsen chart (3400 likes), Mark Zuckerberg and 3d on Forbes (2400 likes). In the newspaper edition for this same period in the New York Times, publications related to Queen Elizabeth and traditional lament (78,000 likes), Volodymyr Zelensky and reclaimed Izium (33,000 likes), James Webb Space Telescope and Neptune (25,000 likes) received the most coverage in terms of likes ). In turn, at Reuters, readers reacted the most with likes to posts related to Russia's aviation and Boeing and Airbus (6,500 likes), Joe Biden and combat racism (5,200 likes), Trevor Noah and leave the program (4,700 likes). And in the Washington Post newspaper, it is related to Jimmy Carter and 98th birthday (44,000 likes), closing the statue of Tubman (32,000 likes), Virginia high schools and transgender students (26,000 likes).

Not all posts contain dates as well as hours. Since the data is collected on October 4, only the posts between September 13 and September 28 contained such information. Therefore, the results regarding hours are limited to this time period.

According to Fig. 5. you can see the number of likes of a certain publication during the day from the newspaper edition of the Washington Post. Each user can see such statistics on their page and track the activity of a certain post over time. Such statistics give an understanding of the times when there is the greatest interaction of users with posts and when it is worth publishing a post on social networks. It can be seen from this figure that from 2 to 3 p.m. this is the time when there are the most users on the network. The reason for such activity is lunch time, when people have time to rest.

The next active phase can be considered the evening time from 22:00 to 23:00. It is clear that there are also more users on social networks after the working day. Also, it is worth noting that the statistics differ depending on the day of the week.

Fig. 6. shows the received active hours of a particular Reuters newspaper publication. As you can see, there is a change in time compared to the Washington Post newspaper. At Reuters, the graph shows that the active hour is 9:00. This is probably due to a different target group of users on the Facebook social network. After that, the next active hour is from 13:00 to 14:00, which is also lunch time. The last active hour during the day is from 19:00 to 20:00 in the evening.

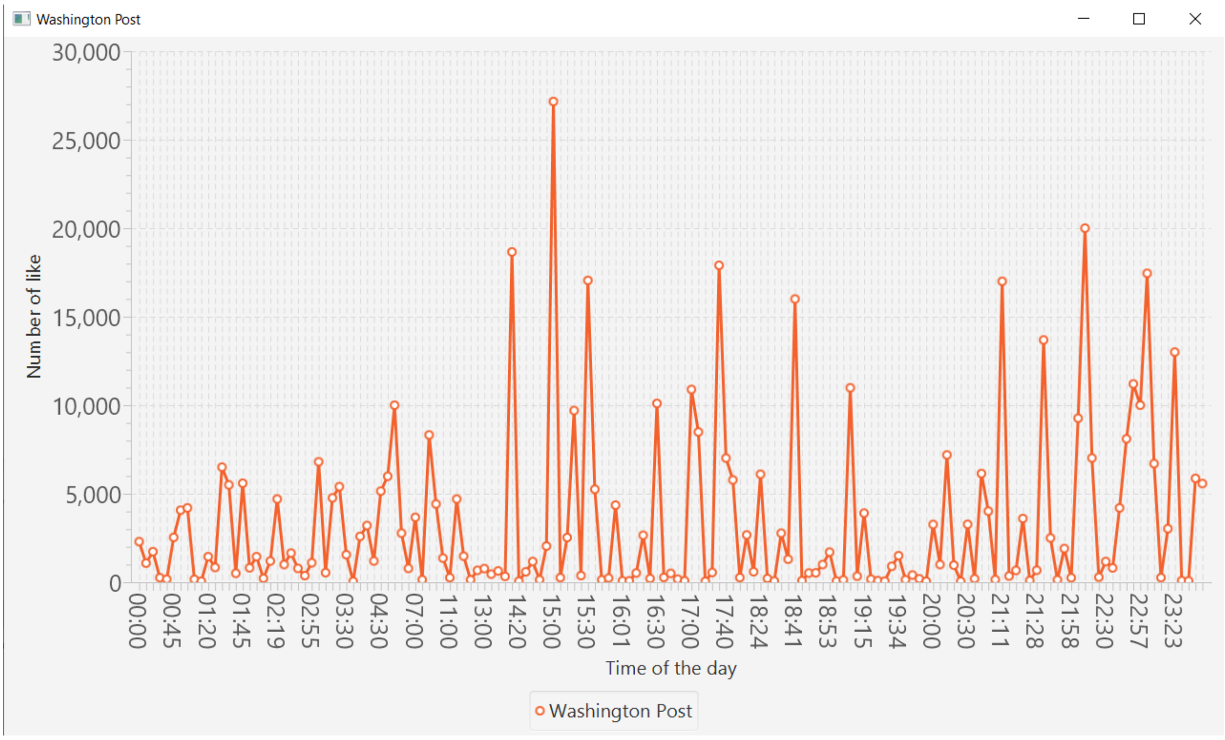


Fig. 5. Analysis of the statistical results of the Washington Post newspaper by the number of post likes per hour

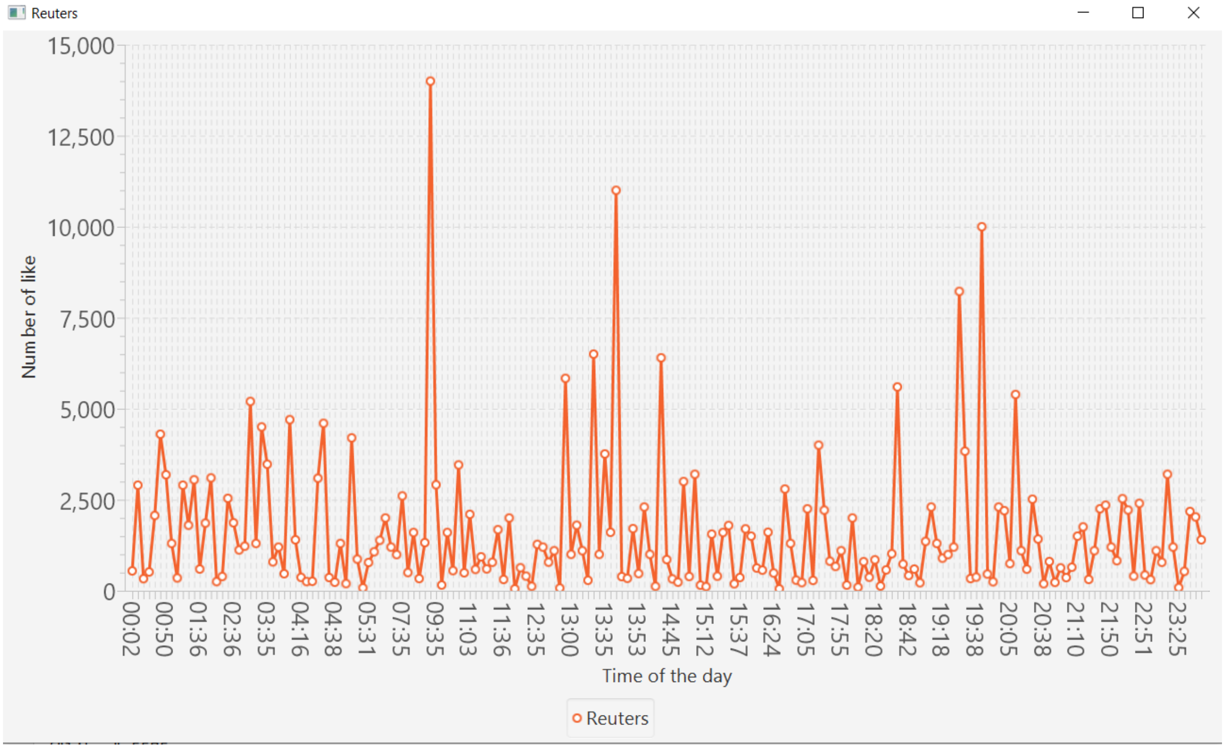


Fig. 6. Analysis of the statistical results of the Reuters newspaper by the number of post likes per hour.

According to Fig. 7. you can see that the schedule is quite different compared to other newspapers on the Facebook social network. But since the New York Times has the largest number of subscribers on its page from different parts of the world and is one of the most popular newspapers today, such statistics are justified.

As a result, the New York Times newspaper can consider the active hours to be from 1:00 to 3:00 in the morning, that is, for users from Europe, it will be the usual evening time they spent reading the news. After that, the next active time is from 11:00 to 13:00, followed by lunch. Then in the evening active hour, which is clearly visible from Fig. 7. is from 17:00 to 18:00 in the evening. The last active phase of this newspaper edition is from 21:00 to 23:00, which is obvious due to the after-work hours of the American target audience, which, of course, is the largest.

From Fig. 8. you can see the statistics of active hours of the Forbes newspaper. The active phase of this newspaper can be started from 1:00 to 3:00. The next active hour for users is from 11:00 to 13:00. After that, the next active phase is between 15:00 and 16:00, after that at 17:00 and 18:30. In addition, a high number of likes is noticeable at 20:00 and 22:00.

Of course, depending on the topic and relevance of the post in social networks, statistics may differ depending on the target audience as well. From the graphs described above, it can be stated that the number of likes directly depends on the period of posting posts in social networks.

The described method allows for a more detailed analysis of user behavior in social networks. In addition to evaluating the number of likes at certain hours in news posts, it is possible to obtain certain trends by other criteria, such as the number of comments or shares. Considering the approximate similarity of the number of comments to the number of likes obtained by data clustering, it can be assumed that similar results can be obtained. Also, there is a similar dependence in the predicted number of likes and comments.

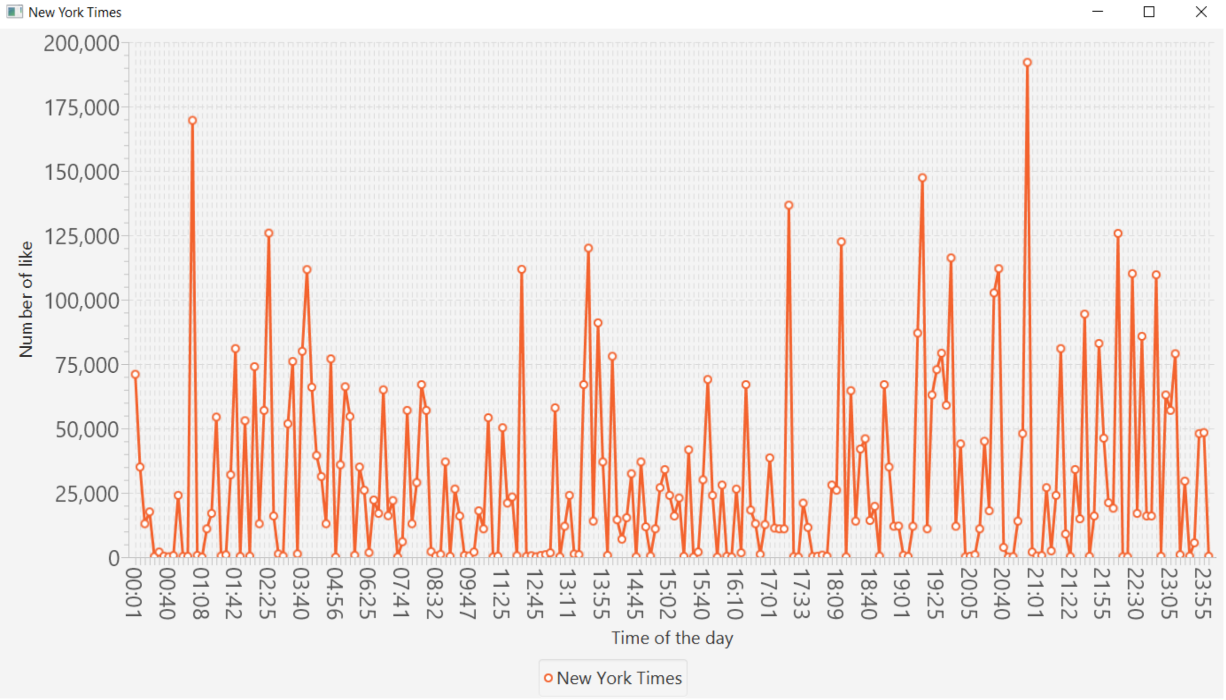


Fig. 7. Analysis of the statistical results of the New York Times newspaper by the number of post likes per hour

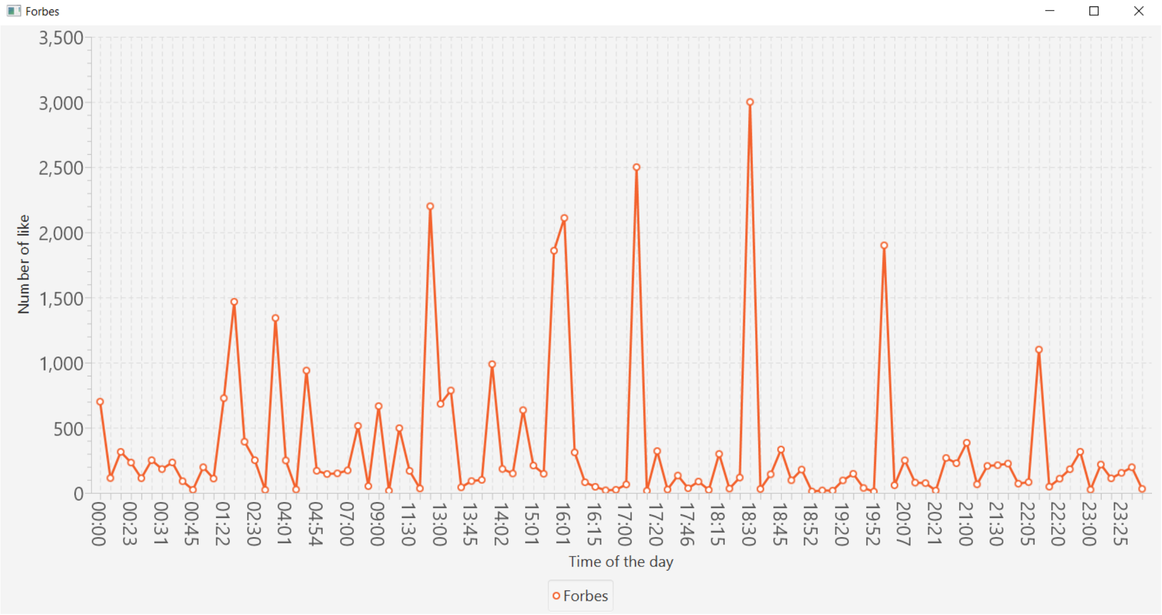


Fig. 8. Analysis of the statistical results obtained by the Forbes newspaper by the number of post likes per hour

Depending on user time zones, the active hours may shift in time, but the general trend is that there are at least three active phases: morning, noon, and evening.

## Conclusion

Social networks are no less a tool for promotion than ordinary advertising on a billboard. Today, many brands and bloggers invest heavily in advertising to promote some information. News publications are no exception and often advertise themselves. The greatest attention is paid to the veracity of the information, relevance and timeliness, design of the submitted post.

The analysis of this kind of information allows you to perform various data manipulations and research them on the basis of real and actual data. Any kind of relationship allows a client or analyst to draw conclusions about a particular page and help promote a particular post without buying advertising directly.

The paper analyzes time intervals during the day for the following newspaper publications: Reuters, Washington Post, Forbes and New York Times. The most active phase among users can be considered from 9:00 to 11:00, 12:00 to 15:00 and 18:00 to 20:00.

The TensorFlow library in the JavaScript programming language is used for data analysis with machine learning tools. The obtained results allow you to see the average statistics of the number of likes to the number of comments. The described process of data collection and processing allows working with a large amount of data from several sources at once.

## References

1. Mikolajewicz Nicholas, Komarova Svetlana V. (2019). Meta-Analytic Methodology for Basic Research: A Practical Guide: Frontiers in Physiology. 10. doi: https://doi.org/ 10.3389/fphys.2019.00203.
2. New York Times [Online]. Available: fahttps://www.facebook.com/nytimes/
3. Washington Post [Online]. Available: https://www.facebook.com/washingtonpost/
4. Reuters [Online]. Available: https://www.facebook.com/Reuters/
5. Forbes [Online]. Available: https://www.facebook.com/forbes/
6. Henrik Wendt, Matteus Henriksson. (2020). Building a Selenium-based data collection tool: Linköping University. Department of Computer and Information Science. 46. Available: https://www.diva-portal.org/smash/get/diva2:1464404/FULLTEXT01.pdf
7. Mysiuk R., Yuzevych V., Mysiuk I. (2022) Api test automation of search functionality with artificial intelligence. Stuc. intelekt. 27(1), 269-274. doi: https://doi.org/10.15407/jai2022.01.269.
8. Satish Gojare, Rahul Joshi, Dhanashree Gaigaware (2015), Analysis and Design of Selenium WebDriver Automation Testing Framework,

Procedia Computer Science, 50, 341-346, ISSN 1877-0509, https://doi.org/10.1016/j.procs. 2015.04.038.

9. Daniel Smilkov, Nikhil Thorat, Yannick Assogba, Ann Yuan, Nick Kreeger, Ping Yu, Kangyi Zhang, Shanqing Cai, Eric Nielsen, David Soergel, Stan Bileschi, Michael Terry, Charles Nicholson, Sandeep N. Gupta, Sarah Sirajuddin, D. Sculley, Rajat Monga, Greg Corrado, Fernanda B. Viégas, Martin Wattenberg (2019) TensorFlow.js: Machine Learning for the Web and Beyond. Proceedings of the 2 nd SysML Conference, Palo Alto, CA, USA. doi: https://doi.org/10.48550/arXiv.1901.05350.
10. Tensorflow [Online]. Available: https://blog.tensorflow.org/2021/01/custom-object-detection-in-browser.html
11. Popenoe R, Langius-Eklöf A, Stenwall E, Jervaeus A. A practical guide to data analysis in general literature reviews. Nordic Journal of Nursing Research. 2021; 41(4):175-186. doi:10.1177/2057158521991949.
12. Cheonsoo Kim, Sung-Un Yang, Like, Cheonsoo Kim, Sung-Un Yang, Like, comment, and share on Facebook: How each behavior differs from the other, Public Relations Review, Volume 43, Issue 2, 2017, Pages 441-449, ISSN 0363-8111, https://doi.org/10.1016/j.pubrev.2017.02.006.
13. Ying Shin Chin & Hasmah Zanuddin (2022) Examining fake news comments on Facebook: an application of situational theory of problem solving in content analysis, Media Asia, 49:4, 353-373, DOI: 10.1080/01296612.2022.2067945.

The article has been sent to the editors 24.10.22.

After processing 10.11.22.