ISSN 2224-087X. Електроніка та інформаційні технології. 2022. Випуск 19. С. 66–74
Electronics and information technologies. 2022. Issue 19. P. 66–74

УДК 004.67 DOI: https://doi.org/10.30970/eli.19.6

METHODS OF ANALYTICS OF BIG DATA OF POPULAR ELECTRONIC NEWSPAPERS ON FACEBOOK

I. Mysiuk[1], R. Mysiuk[1], R. Shuvar[1], V. Yuzevych[1,2]

[1]Ivan Franko National University of Lviv
50 Drahomanova St., UA-79005 Lviv, Ukraine
iruna.musyk8a@gmail.com

[2]Karpenko Physico-mechanical Institute of the NAS of Ukraine
5 Naukova St., UA-79601 Lviv, Ukraine

Due to the popularity of social networks, all famous brands use them to promote and support their product. For example, well-known foreign newspapers such as: Washington Post, New York Times, Time, Reuters, Forbes duplicate information about news on the Facebook social network for greater readership. Using methods of automated data collection from web pages, a list of posts is formed based on which data analysis is performed. Statistical results of frequency, popularity of certain articles, audience reach and people's reaction to posts are obtained from a large volume of data. In the Java programming language and using additional Selenium, JavaFX libraries, all processes for data normalization is developed and data visualization is used. In addition, the dependence of the post coverage of newspaper editions on the number of posts published during the day is investigated in Facebook social network. The work also examines the most popular posts and their topics. The relationship between keywords and real events is analyzed.

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

Overview

Social networks have become an integral part of our lives due to fast methods of communication and obtaining information. This trend has especially intensified during the pandemic. Newspapers, which reduced their circulation in paper form and began to increase the reach of the audience in the social network, were no exception. The Washington Post, New York Times, Time, Reuters and Forbes can be considered among the most authoritative world posts due to wide citations and verified information.

Modern technologies allow you to follow the news and exchange information in real time. Data accumulates over time and their number reaches billions every month. Processing such a large amount of data requires faster ways of searching and structuring them in some information system [1, 2].

4 newspapers on Facebook are selected for data analysis. As of the beginning of October 2022, the number of followers in the New York Times newspaper [3] in social networks reaches 18 million, in Reuters [4] - 6.6 million, in the Washington Post [5] ranges between 7.2 million and Forbes [6] is 7.5 million.

A whole team of specialists works to promote fresh news in social networks. First of all, the title of the post catches the audience's eyes, which should give an understanding and sum-

mary of the entire content of the post. The next factor for the promotion of the newspaper is the relevance and appropriateness of the information presented in social networks. Usually, such teams are based on information received from analysts and other experts in certain gas industries. Analysts, in turn, collect information from various sources, analyze competitors and, based on their experience, make certain conclusions and assumptions.

The problem is the long process of collecting and processing information, especially historical information. Today, there is another problem: the oversaturation of information on the Internet. Data redundancy is due to the existence of a large number of competing sources of information, which are often unverified.

Analytics should be conducted with as much data as possible. Usually, the most popular social networks include Facebook, Instagram, TikTok and Twitter. Each of the mentioned social networks has its own features of publishing news and other content. For example, Facebook and Twitter are similar in content type as they use mostly text content. But at the same time, there is a difference in the number of permissible characters for a certain post. On Twitter, posts are usually short and informative, for example in the form of headlines for the most part. Facebook allows you to include up to 5,000 characters in your posts by attaching photos or other interactive content. Instagram specializes mostly in photos, while TikTok specializes in short videos.

The purpose of the work is the analysis of automatically collected data, the search for certain trends and patterns in social networks based on newspaper posts.

The tasks that arise in accordance with the goal are the formation of statistics, periodicity, coverage, post preferences and the timeliness and relevance of information at a certain moment.

Application

The social network Facebook today is rapidly updated and developed. The main reason for this is its openness and high prevalence among people.

In connection with the transition from Russian social networks, people began to switch to Facebook and other social networks en masse. Every day, the news feed is viewed by millions of people around the world reading news, posting photos, running their own blog, and more.

Therefore, for data analysis, we took into account the parts of the post to which people most often pay attention and react according to Fig. 1.

First of all, people pay attention to the picture, if we are talking about cosmetics, clothes, or goods. When analyzing a page of news or informative posts, people pay attention to the text, the number of likes, comments and shares. In this case, it was the information channels of newspaper posts that were investigated, so the parts marked with red rectangles were taken into account.

Information is collected using the Selenium tool in the Java programming language [7]. In Java, it is possible to add additional libraries, in particular, when collecting data, a tool for automated testing was used, which allows direct access to elements on the web page. Moreover, the search for elements on the page is quite flexible, using various locators: xPath, css, id, class, tagName, etc. At least one of the named locators must identify a unique element on the web page. Therefore, even the deepest element in the DOM tree must be found. DOM (Document Object Model) contains a document based on an HTML page (HyperText Markup Language) and contains all the components that any page on the Internet.

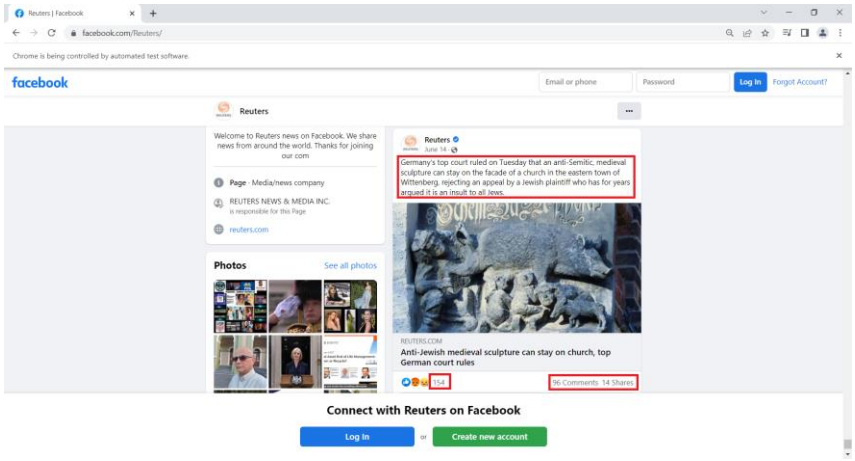

Fig.1. Displaying data on a web page

Selenium works using the JSON Wire Protocol, which has a set of methods for working with web elements such as click and getText(). Methods communicate with the web browser by executing certain events specified by the developer. All actions that the user can perform on web pages are contained in Selenium methods [7].

The principle of Selenium is to simulate real human behavior when working with a web browser. To begin with, you need to open a web page using the get method in a web browser, and then, sequentially describing each step with the help of element access locators, gradually find all the necessary components for the successful completion of a certain task. After the execution of the program, you can see the result visually, whether the output data is displayed in the console.

Collected data is stored in a csv (Comma-separated values) file. This format is used when writing data to a file where values are separated by commas. This file will be used for analytics. In this case, it will be most appropriate to use this format, since the data collected from the web pages of social networks of newspaper posts will be displayed in the file by separating each read value with a comma.

About 10,000 posts and related information from four newspaper posts are collected to research this topic.

According to Fig. 2 represents the visualization of the collected data takes place using the JavaFX library [8]. This library allows you to flexibly work with input data using graphs and has many useful functions for working with data [9].

Among the possible displays using this library are linear and columnar. In addition, it is possible to customize fonts, axis labels, change image scaling and display results in real time.

Each Facebook feed page has 5 news posts. Scrolling is used to move to the next page, which can be done using Selenium. Some posts were duplicated, so when processing the data there is a need to delete them.

The duration of real data collection took several hours. It is clear that it is necessary to prescribe some forced stops for uploading data. These pauses were set to 3000 milliseconds in this example. The delay is a relative value and may vary depending on the speed of the Internet

and the power of the computer. Therefore, this method of data collection can be considered quite fast not only from social networks, but also from other web pages.

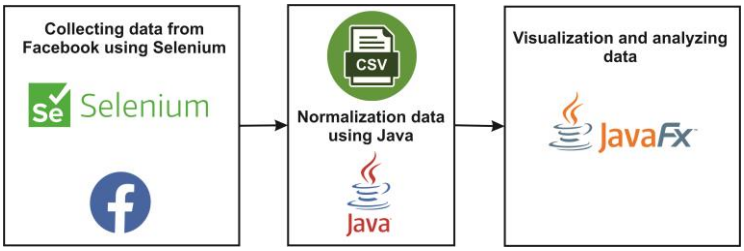

Fig.2. Flow chart of the process of collection and visualization of the developed functionality.

The normalization process includes several processes that are related to data preprocessing. It consists in removing duplicate data that may have remained during data collection, when reducing to a common format of time, text and marks [10, 11]. For example, a large number of likes is often indicated with the abbreviation 1.5K, that is, after translation, their number will be 1500. Also, the date format in the post depends on the time of post. Older posts contain only the month, while newer posts also contain the hour.

Analysis of results

Collected data can be analyzed using graphical representation of information. The graph shows the data of 4 newspaper editions for a period of three weeks from the middle of September to the beginning of October. For example, in Fig. 3 visualizes the New York Times has the most likes for certain posts over the course of three weeks.

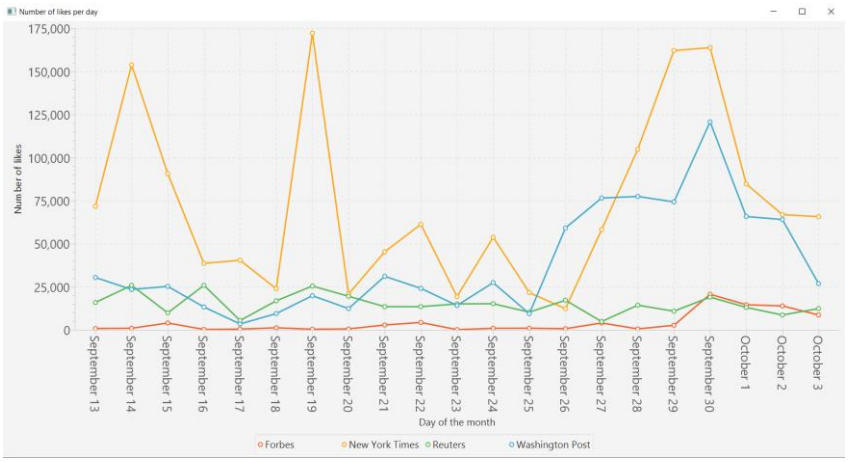

Fig.3. Displaying the number of likes for each of the newspaper editions

This is due to the fact that this particular newspaper has the most subscribers to this page on the network, so there is a large number of likes, in some cases reaching even 170,000 likes

for a certain post. In addition, you can see that the Washington Post also achieved significant results with 120,000 likes at the end of the month.

Of course, it is not advisable to conduct analytics based only on the number of likes or the number of subscribers of these newspaper posts [12]. After all, there are moments when a timely published post at the right time, even with small numbers of subscribers, can achieve unheard of results [13].

That is why, before each subsequent post, analysts review the current situation based on previous posts, taking into account the following aspects [14]:

- Relevance
- Expediency
- Periodicity
- Truthfulness, etc.

The second factor that affects the distribution of other content among strangers is the distribution of this post on the pages of social networks, in groups, conversations, etc.

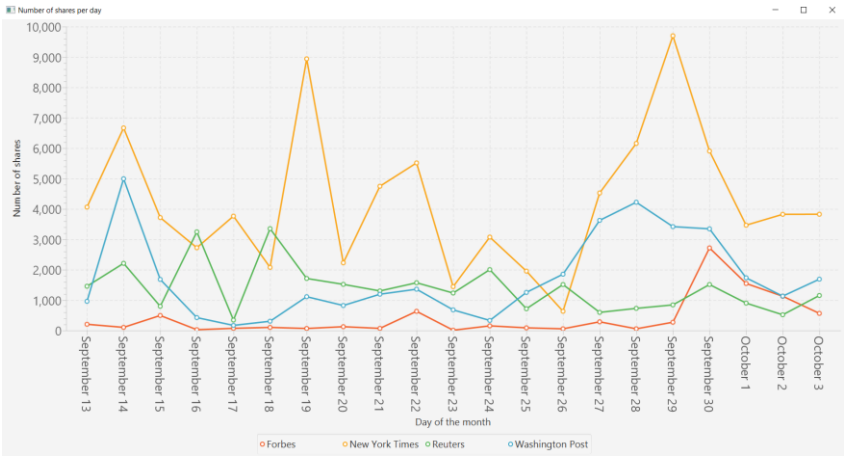

Fig. 4. The number of posts shared during a certain period

According to Fig. 4 it can be seen that the largest number of distributions is in the New York Times newspaper, followed by the Washington Post and Reuters. But these data cannot be considered a priority either. Because the topic of the post itself, the time of its post, is not taken into account. At the same time, such visualization is important in determining the final result of data analysis.

In the same way, you can describe the number of comments. In Fig. 5. it can be seen that between September 27 and October 1, users were most active by commenting or discussing the post. As can be seen from the graph, the number of commenters is much smaller than the number of likes. That is, users prefer to respond quickly to certain posts, using a variety of reactions, rather than writing text. The number of comments is approximately the same in all newspaper posts.

From the results obtained in terms of the number of shares and user preferences, it is possible to draw a conclusion about the most active days during the studied period: September 14, 19, and 29.

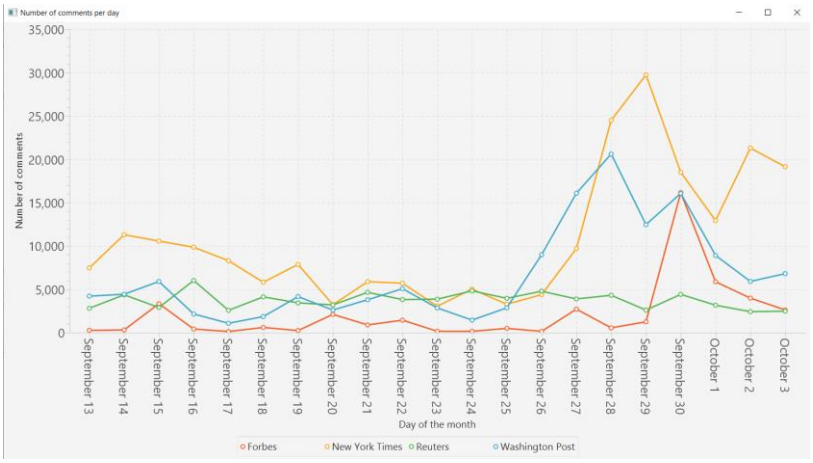

Рис.5. The number of comments under each post during a certain period

In Fig. 6. shows the average number of posts sprayed during the day reaches 25 on a page in the Facebook social network. The growth dynamics of the number of posts in the four editions is the same. The increase in the number of posts is noticeable at the beginning of October.

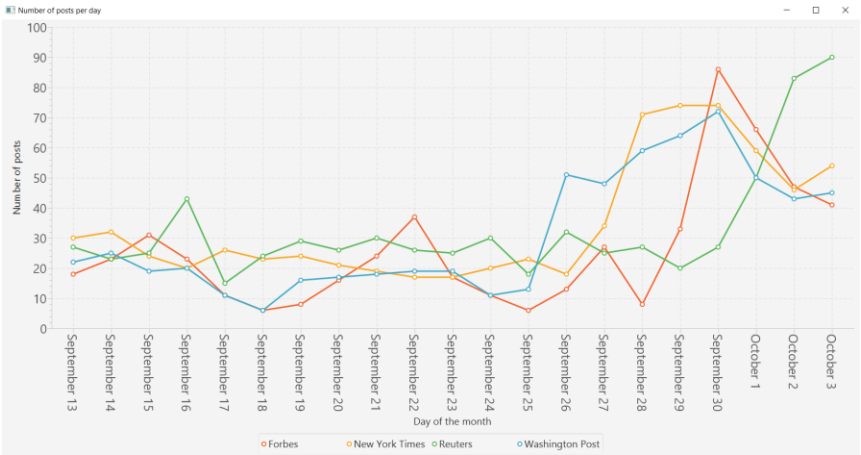

Fig.6. The number of posts under each post during a certain period

From this figure, we can conclude that the number of posts does not affect the number of likes. Despite the number of posts in a certain period of time from September 13 to September 25, the number of comments and preferences changed for each newspaper in different ways, but there is a synchronicity in the dynamics of reaching the target audience.

The period of the most popular posts among users was analyzed. For this purpose, from the total number of posts, only the text was left and divided into words. The number of words is counted during the execution of the program and the results are displayed graphically.

In Fig. 7. the number of mentioned words in the studied interval that had the greatest coverage in terms of the number of likes, distributions and comments among users is shown.

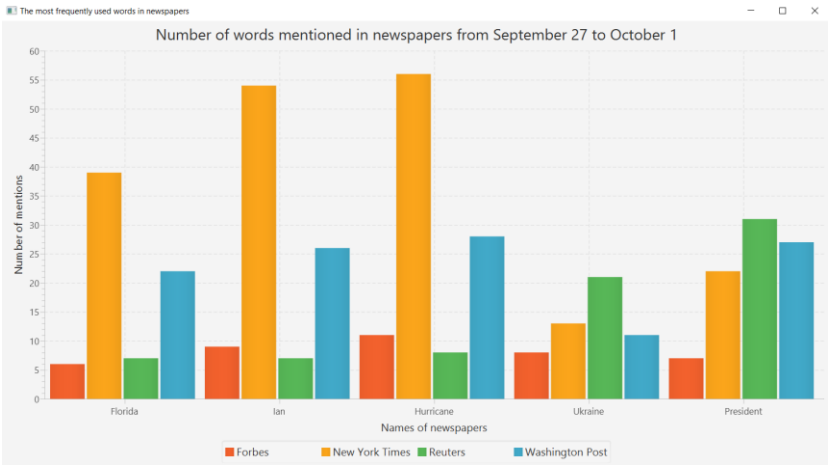

Fig.7. Histogram of the most frequently mentioned words in newspaper posts during the period of highest popularity

From the bar chart, we can see that in this period from September 27 to October 1, the words Ian, Florida, Hurricane, Ukraine and President are mentioned. All the listed words really had something to do with these events. For example, there were frequent presidential addresses mentioning Ukraine and the tragic events in Florida due to Hurricane Ian. However, there were some differences in the amount of use of certain keywords within each journal. For example, Elon Musk was often mentioned in the Forbes newspaper (7 times), in Reuters - the Nord Stream (15 times), pipelines (14 times) and gas (26 times). Confirmation of this is the gas leak in the North Sea, which caused a lot of publicity in social networks.

Conclusion

The amount of information is constantly increasing, and an important aspect today is its form of presentation and its content. It is important to accurately and compactly form the submitted marial in the title of the post. Especially in social networks, where there is a lot of unnecessary content such as advertising, offered content. The idea of social networks is to use content to keep the user interested in social networks for as long as possible. Recently, the amount of news has steadily increased and the impact of events on our daily life is constantly growing.

An analysis of the posts of the world famous magazines Forbes, Washington Post, Reuters and New York Times in the Facebook social network was carried out. The data format in this case is the same when examining such journals and access to the main elements on the web page is shared. That is, with the help of the used Java programming language and the auxiliary tools JavaFx and Selenium, it is possible to simplify the work when collecting and

normalizing data. Collected information about the date, likes, text, shares and comments allows you to estimate the reach of a certain post and the interest of users in it.

As a result of the analysis, the dependence of the number of likes, comments, distributions and posts on the date of post was obtained. From graphic representations, it is possible to draw conclusions about the impact of certain events and people's reactions to them. For example, in the period from September 27 to October 1, there is a special activity of users related to external events in the world.

In addition, the main keywords for the most active time period were formed and the affiliation of those words to certain events was determined. The following words were among the most used in the text of posts during the studied time period: Ian, Ukraine, President, Hurricane, Florida. Moreover, the tendency of using these words is similar in all editions.

Machine learning algorithms and more detailed analysis of a longer period of time will be applied to the existing results in future works. Often, additional information such as hashtags and direct links to the website are added to the posts, which can also be used for analysis.

This approach to data analysis using information technologies can be used not only for the chosen topic, but also for other areas of our life.

References

[1] *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. Vol. 41, No 4: P. 175-186. doi: https://doi.org/10.1177/2057158521991949

[2] *Domingue J., Lasierra N., Fensel A., van Kasteren T., Strohbach M., Thalhammer A.* // Big Data Analysis. In: Cavanillas, J., Curry, E., Wahlster, W. (eds) New Horizons for a Data-Driven Economy. Springer, Cham. 2016. doi: https://doi.org/10.1007/978-3-319-21569-3_5.

[3] New York Times [Online]. URL: https://www.facebook.com/nytimes/

[4] Reuters [Online]. URL: https://www.facebook.com/Reuters/

[5] Washington Post [Online]. URL: https://www.facebook.com/washingtonpost/

[6] Forbes [Online]. URL: https://www.facebook.com/forbes/

[7] Selenium automates browsers. [Online]. URL: https://www.selenium.dev/

[8] JavaFX. [Online]. URL: https://openjfx.io/

[9] *Crews B., Drees J., Greene D.* Data-driven quality assurance to prevent erroneous test results // Critical Reviews in Clinical Laboratory Sciences, 2020. Vol. 57. No 3, P. 146-160, DOI: https://doi.org/10.1080/10408363.2019.1678567

[10] *Singh D., Singh B.* Investigating the impact of data normalization on classification performance // Applied Soft Computing, Volume 97, Part B, 2020, 105524, ISSN 1568-4946, doi: https://doi.org/10.1016/j.asoc.2019.105524.

[11] *Mysiuk R., Yuzevych V., Mysiuk I.* Api test automation of search functionality with artificial intelligence // Stuc. intelekt. 2022. Vol. 27, No 1. P. 269-274 doi: https://doi.org/10.15407/jai2022.01.269

[12] Hızal, A. Frequency domain data merging in operational modal analysis based on least squares approach // Measurement, 2021. Vol. 170, 108742. https://doi.org/10.1016/j.measurement.2020.108742

[13] *Marino C, Gini G, Vieno A, Spada M.* A comprehensive meta-analysis on Problematic Facebook Use // Computers in Human Behavior, 2018. Vol. 83, P. 262-277, ISSN 0747-5632, https://doi.org/10.1016/j.chb.2018.02.009.

[14] *Ostic Dragana, Qalati Sikandar Ali, Barbosa Belem, Shah Syed Mir Muhammad, Galvan Vela Esthela, Herzallah Ahmed Muhammad, Liu Feng.* Effects of Social Media Use on Psychological Well-Being: A Mediated Model // Frontiers in Psychology. 2021. Vol. 12. ISSN 1664-1078. doi: https://doi.org/10.3389/fpsyg.2021.678766

[15] *Ji Changqing & Li Yu & Qiu Daowen & Jin Yingwei & Xu Yujie & Awada Uchechukwu & Li Keqiu & Qu Wenyu*. Big data processing: Big challenges. Journal of Interconnection Networks. 2013. doi: https://doi.org/10.1142/S0219265912500090.

МЕТОДИ АНАЛІТИКИ ВЕЛИКИХ ДАНИХ ПОПУЛЯРНИХ ЕЛЕКТРОННИХ ГАЗЕТНИХ ВИДАНЬ У FACEBOOK

Ірина Мисюк [1], Роман Мисюк [1], Роман Шувар [1], Володимир Юзевич [1,2]

[1]Львівський національний університет імені Івана Франка,
вул. Драгоманова, 50, м. Львів, 79005, Україна

iruna.musyk8a@gmail.com

[2] Фізико-механічний інститут ім. Г. Карпенка НАН України,
вул. Наукова, 5, м. Львів, 79601, Україна

У зв'язку із набуттям популярності соціальних мереж усі відомі бренди використовують їх для просування та підтримки свого продукту. Газетні видання активно використовують соціальні мережі як альтернативу паперовим варіантам. Найбільш популярними соціальними мережами є Facebook, Instagram, Twitter та TikTok. До прикладу відомі іноземні газетні видання такі, як: Washington Post, New York Times, Time, Reuters, Forbes дублюють інформацію про новини в соціальній мережі Facebook для більшого охоплення читачів. Перед початком збирання необхідно здійснити пріоритезацію ключових елементів публікації та порівняти їх присутність з іншими газетними виданнями. Методами автоматизованого збирання даних з веб сторінок сформовано список публікацій на основі яких проведено аналітику даних. З великого об'єму даних було отримано статистичні результати періодичності, популярності певних статтей, охоплення аудиторії та реакції людей на публікації. Проведено нормалізацію даних для спільного формату всіх зібраних даних. Цей процес розділений на кілька етапів: видалення дублікатів, переведення схожих даних до єдиного формату та заміна допоміжних символів на зрозумілі для комп'ютера мову. Використано лінійне та стовбчасте відображення результатів, завдяки якому найкраще відображено різницю між значеннями в кожен період часу.

Мовою програмування Java та з використанням додаткових Selenium, JavaFX бібліотек розроблено всі процеси для нормалізації даних та використано візуалізацію даних. У розробленій програмі зібрано всі етапи для аналізу та відображення результатів. Крім того було досліджено залежність охоплення публікації газетних видань від кількості публікації виставлених протягом дня. У роботі також досліджено найбільш популярних публікації та їх тематику. Проаналізовано звязок ключових слів з реальним подіями. Отримані результати по найвживаніших словах в певному періоді часу можуть бути використані для визначення акцентів певних газетних видань у соціальних мережах. Такий підхід може бути застосований для аналізу, порівняння чи прогнозування даних з використанням інформаційних технологій базуюючись на тенденції та відслідковувати їх тенденції з часом.

Ключові слова: великі дані, соціальні мережі, аналітика даних, автоматизоване збирання даних, опрацювання даних.

Стаття надійшла до редколегії 12.10.2022
Прийнята до друку 18.10.2022