Analyzing Follower Data on Social Platforms Using Big Data Tools

Levente Füzér
John von Neumann Faculty of Informatics
Obuda University
Budapest, Hungary
fuzerl@stud.uni-obuda.hu

Miklós Sipos
John von Neumann Faculty of Informatics
Obuda University
Budapest, Hungary
sipos.miklos@nik.uni-obuda.hu

2025 IEEE 29th International Conference on Intelligent Engineering Systems (INES) | 979-8-3315-9771-9/25/$31.00 ©2025 IEEE | DOI: 10.1109/INES67149.2025.11078171

***Abstract*—Since their emergence, social platforms have rapidly evolved, with the internet playing an increasingly significant role in everyday life. Constant user activity generates vast amounts of data, making digital spaces not only sources of information but also valuable for analyzing human behavior and virtual social interactions. Understanding the underlying mechanisms of social platforms is essential, given their widespread influence. A key question arises regarding the factors contributing to the success and impact of online videos and posts. Additionally, effective strategies for maximizing content reach remain an area of interest for content creators and entrepreneurs. To address these aspects, an analytical system has been developed to automate large-scale data collection from social platforms. This system standardizes, processes, and analyzes the gathered data, providing insights through visualizations and statistical evaluations. As part of this study, data from over 20,000 short-form videos were analyzed and compared. Several notable patterns and regularities were identified, offering a deeper understanding of content performance dynamics on social platforms.**

***Index Terms*—social media analytics, big data, followers, metrics, web scraping, data visualization**

I. Introduction

The primary objective of this research is to uncover the factors that contribute to the success of short-form videos on social platforms such as TikTok, Instagram Reels, and YouTube Shorts. These platforms emphasize quick, engaging, and visually driven content. This study explores what influences viewership and engagement by identifying patterns—like the use of captions, hashtags, and other video traits—to help creators improve content performance.

Short-form videos have become a dominant force in digital media, offering creators new ways to connect with audiences. Yet with the sheer volume of content uploaded daily, standing out is a challenge. Factors like caption structure, trending hashtags, and background music can all impact visibility. Understanding these helps creators build thoughtful strategies to better engage their target audiences.

A mindful content strategy is essential in today's fast-paced digital environment. It's about more than visuals—it involves knowing audience habits and preferences. Using analytics to determine the best time to post or the most engaging content types can elevate a campaign, while authentic connections with viewers help build long-term engagement.

This research has strong practical relevance. Creators, marketers, and businesses can apply the findings to develop evidence-based strategies. Insights into metrics like likes, comments, and views support more targeted content planning and help build a loyal, engaged audience.

Beyond immediate applications, the study contributes to broader digital media knowledge. By analyzing large datasets and extracting actionable patterns, it paves the way for future work that bridges data-driven insights with creative content, helping creators thrive on dynamic social platforms.

II. Related work

A. Data Collection Systems

1) Web Scraping: Web scraping is used to extract data from websites and store it in a structured format. It is considered an efficient method for collecting large amounts of data that would be difficult to gather manually [1] [2].

2) Web Scraper: Web Scraper is offered as a free browser extension, allowing data extraction to be easily configured. Data can be exported in CSV or Excel format, while the paid version supports JSON and cloud storage services. Scheduled scraping and parallel data extraction processes are also supported [3].

3) Selenium: Selenium is recognized as an open-source automation tool that facilitates web scraping by interacting with dynamic web elements. JavaScript execution is handled efficiently, enabling the extraction of data that static parsers cannot access. Compatibility with multiple browsers is ensured, and parallel execution is supported [4].

4) Scrapy: Scrapy is widely used as a Python-based web scraping framework, known for its active developer community and broad operating system compatibility. It is commonly integrated into Python applications for structured data extraction [5].

5) Beautiful Soup: Beautiful Soup is employed as a Python package designed for parsing HTML and XML documents. It is optimized for tree-structured data and is compatible with various parsers, including the built-in Python HTML parser [6].

B. Data Processing Systems

1) ETL (Extract, Transform, Load): ETL processes are implemented to integrate data from multiple sources into a unified format. The data is extracted, transformed for consistency, and then loaded into a data warehouse or database for further analysis [7].

2) Jupyter: Jupyter is utilized as an interactive development environment for data processing and analysis. It enables Python-based scripting for cleaning, transforming, and visualizing data within a user-friendly notebook interface [8] [9].

3) Power BI: Power BI is applied as a business intelligence tool for data visualization and analysis. It is designed to process large datasets and generate insightful reports through interactive dashboards.

C. Specification of Analyzed Content and Introduction to Short-Form Videos

Several social media platforms were selected for analysis, focusing on videos, posts, and profiles. Priority was given to short-form videos — typically a few seconds to two minutes long — due to their rising popularity. The platforms examined include TikTok, Instagram, YouTube, and Facebook.

TikTok's rapid global rise since its 2016 launch is largely due to its short-form video format. Its algorithm-driven feed delivers personalized content instantly, with users scrolling through videos seamlessly. This success led platforms like YouTube (Shorts), Instagram (Reels), and Facebook (Reels) to introduce similar features, offering comparable data for analysis.

D. Analysis and Comparison of Collectible Data from Platform Profiles

Different social media platforms provide varying publicly accessible data. The types of data available on public user profiles and posts have been summarized:

- TikTok: username, follower count, following count, number of likes, biography, user-provided links.
- Instagram: username, follower count, following count, number of posts, biography, user-provided links.
- YouTube: username, subscriber count, number of videos, biography, user-provided links.
- Facebook: username, follower count, following count, biography, user-provided links, additional profile details (optional).

Most platforms provide similar profile data. Facebook allows additional profile details to be shared in the "About" section, however, these fields are optional and may not always be completed.

E. Analysis and Comparison of Collectible Data from Short-Form Videos

Only short-form video content has been considered, including Instagram Reels, Facebook Reels, and YouTube Shorts (where videos under one minute are automatically classified as Shorts).

- TikTok: uploader's name, upload time, view count, like count, comment count, number of saves, video description, hashtags, comment section content.
- Instagram Reels: uploader's name, upload time, view count, like count, comment count, number of saves, video description, hashtags, comment section content.
- YouTube Shorts: uploader's name, view count, like count, comment count, video description, hashtags, comment section content.
- Facebook Reels: uploader's name, view count, like count, comment count, number of shares, video description, hashtags, comment section content.

Summary: Most platforms provide similar metadata fields. However, Facebook does not make the upload time publicly available but includes the number of shares, which is valuable for measuring video success.

F. Key Factors Determining the Success and Effectiveness of Short-Form Videos

Traditional success indicators include views, likes, and comments, but deeper engagement metrics are also key.

One of the most important is average viewer retention rate. A low rate (e.g., two seconds) signals a loss of interest. This metric is often visualized as a graph showing how many viewers remain at each point in the video. Data from TikTok Analytics illustrates declining attention spans, a trend linked to fast-paced environments and sociological changes. Top-performing short-form videos use dynamic visuals, bright colors, and sharp, concise audio to keep viewers engaged and improve retention.

G. Importance of a Strategic Content Approach

A well-defined content strategy is essential for influencers, content creators, businesses, and organizations that rely on online campaigns to reach a broad audience effectively.

Strategic content creation involves understanding the target audience and intentionally producing content tailored to their preferences. Furthermore, content must be published on appropriate platforms at optimal times. Analytical tools play a crucial role in this process by providing real-time feedback on content performance, allowing creators to refine their strategies and maximize engagement.

H. Definition and Significance of Data Visualization

Data visualization refers to the graphical display of data, helping to identify trends, relationships, and outliers. It also supports clear communication with non-experts, reducing the risk of misinterpretation [10].

It plays a vital role across many fields, with tools like tables, charts, diagrams, and heatmaps used to simplify data and enhance understanding.

The strength of data visualization lies in its ability to make data accessible. It allows both simple and complex analyses to be understood by a wider audience, regardless of expertise.

These tools help interpret the massive flow of daily data. In business, they support strategic decisions; in government

and research, they clarify complex issues. Professionals across sectors (from executives to educators) benefit from effective data analysis and visualization.

I. The Most Common Types of Data Visualization

Various forms and tools are available for visualizing data, among which the most widely used are the following:

- Table: A structured format where data is presented in cells organized into rows and columns. This fundamental visualization format facilitates easy comparison and organization of data.
- Graph: Graphs consist of nodes and edges representing relationships between different entities. These relationships may be directed or undirected.
- Map: Maps are used to represent geographical data, such as countries, regions, or cities, and display relevant information. They enable analysis and comparison based on geographic location.
- Heatmap: Heatmaps are color-coded diagrams that depict data values across a specific space. Higher values are represented by warmer, more vibrant colors, while lower values appear cooler and less intense.
- Line Chart: Line charts illustrate changes in data over time or another variable. They are particularly useful for tracking trends and variations over time.
- Bar Chart: Bar charts use bars to represent data, where the height of the bars corresponds to data values. These charts are effective for comparing values and visualizing differences.
- Pie Chart: Pie charts utilize circular segments to represent data proportions. They are typically used to visualize simple ratios, parts, and whole relationships.
- Gantt Chart: Gantt charts are employed to display time-related tasks and events on a horizontal timeline. These charts are instrumental in project management and task scheduling.
- Venn Diagram: Venn diagrams use circular regions to illustrate relationships between data sets. Overlapping circles indicate shared attributes or connections among data points.

III. Methodology

The full system design can be seen on Figure 1., where each step (Figure 2.) of the procedure is basically represented with one specific component.

A. Data Collection

The data collection phase utilized automated tools, primarily Web Scraper and Octoparse, to systematically gather data from YouTube Shorts [1]. Web scraping is the automated collection of large amounts of data [2]. Before deciding to go with Web Scraper and Octoparse, several web scraping tools were researched and compared [3] [4] [5] [6]. The web scraping was conducted in an entirely legal manner, following a read-through of the specific platform guidelines [11] [12]. These tools were meticulously configured to navigate

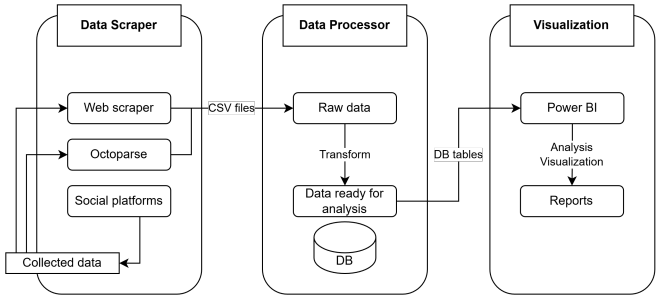

Fig. 1. Components of the system.

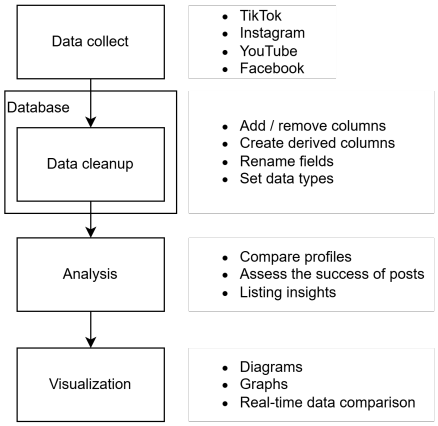

Fig. 2. High-level steps of the system.

through video pages, identify and capture relevant HTML elements, and extract critical details such as video views, likes, comments, captions, hashtags, and uploader information. The configuration process included creating specific selectors for each data type, ensuring that the scraping scripts accurately targeted the required elements on dynamically loaded web pages. Adjustments were made to handle potential challenges posed by JavaScript-heavy sites, including enabling headless browser modes and implementing scrolling actions to load hidden content.

To avoid detection by anti-bot systems, the scraping process used safeguards like random time intervals and limited simultaneous requests. This helped prevent platform bans. Extracted data was exported to CSV or Excel for further processing. Datasets were reviewed for completeness, and issues like partial records were corrected through script adjustments. The final result was a robust dataset with thousands of data points, offering a solid base for analysis.

B. Data Processing

Raw data was imported into Microsoft SQL Server for structured storage and efficient processing. The focus was on cleaning and standardizing data — removing non-numeric characters, unifying date formats, and normalizing text fields.

Custom scripts in Visual Studio automated the transformations, handling edge cases like missing values and calculating

key metrics such as engagement rate (likes + comments ÷ views) to compare video performance.

The scripts were built for reuse and scalability, enabling smooth processing of new data. Each step was validated with sample checks to catch and correct issues before analysis.

C. Analysis

The processed and enriched dataset was then imported into Microsoft Power BI, a platform chosen for its advanced analytical and visualization capabilities [7]. After the selection of the visualization tool, visualization principles and guidelines were researched, to ensure the quality and clarity of the visualizations [13] [14].

Interactive dashboards were built in PowerBI to explore how video attributes relate to performance. Bar charts showed links between caption length and engagement, while scatter plots examined hashtag count versus viewership.

Advanced filters and drill-downs allowed deeper analysis of specific video groups, with statistical tools validating trends. Key metrics—like average engagement by caption length or view distribution for trending hashtags—were computed and visualized.

The full pipeline, from data collection to visualization, was iteratively refined for accuracy and clarity. This approach demonstrated the value of data-driven strategies in understanding and optimizing video content.

IV. Evaluation

A. Optimal Upload Frequency

The analysis examined how upload frequency affects viewership on YouTube Shorts. Results showed a strong positive correlation—channels uploading daily saw higher view counts. Frequent uploads boost visibility and signal activity to algorithms, often leading to greater promotion. Still, quality must be maintained to avoid diminishing returns.

Figure Figure 3. illustrates this correlation. Blue bars show total views for the top 15 cooking-related channels over two years, while the red line indicates daily upload rates. With few exceptions, more frequent uploads aligned with higher views, reinforcing the value of consistent content.

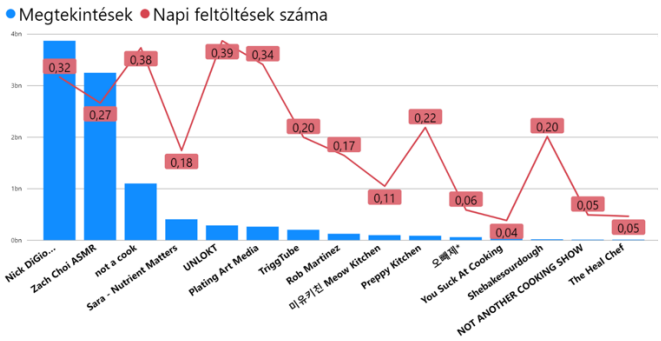

Fig. 3. View and upload frequency correlation (blue: view count, red: daily upload count).

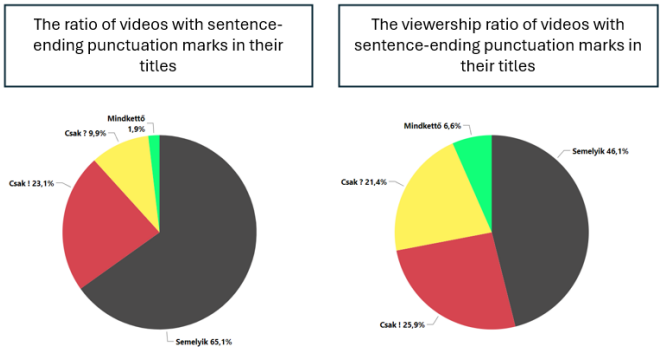

Fig. 4. Sentence-ending punctuation mark distribution. Color meanings: grey = no characters, red = only exclamation mark (!), yellow = only question mark (?), green = both.

B. Impact of Video Title Length

The study examined the optimal length for video titles, considering that most users view videos on mobile devices. Analysis showed that titles under 40 characters performed better in terms of views and likes, accounting for 67% of total views and 71% of total likes. This aligns with mobile display constraints, as longer titles are often truncated, reducing their visibility and effectiveness. [15] Only a portion of viewers read the titles of short-form videos, and only a fraction of those actually take the effort to expand a truncated title. Shorter titles also tend to be more engaging and easier for viewers to process, making them more likely to click on videos.

C. Content of Video Titles

The analysis explored how title elements like punctuation and emotive language affect engagement. Titles with question marks or exclamation points saw higher interaction rates, as these elements trigger curiosity and emotional responses. Questions, in particular, encouraged comments, boosting engagement and platform promotion.

Another benefit of using questions is community building. While watching a video is a fleeting, one-sided interaction, engaging in conversation with a creator fosters a deeper, longer-lasting connection, increasing the likelihood of followers and further interactions. Results are shown in Figure 4.

D. Engagement Rate Significance

The engagement rate (calculated as the percentage of views resulting in likes and comments) was analyzed as a critical performance metric. A higher engagement rate was positively correlated with better algorithmic promotion, as platforms tend to prioritize content that generates meaningful interactions. For example, an engagement rate increase of just 1% could translate into thousands of additional interactions for high-view videos, underscoring the importance of encouraging viewer interaction through compelling content and calls-to-action.

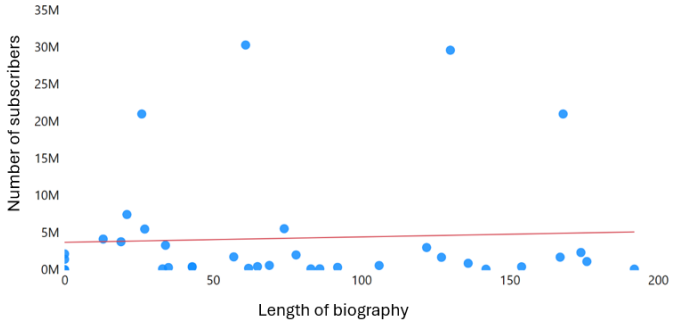

Fig. 5. Biography length and subscriber count correlation.

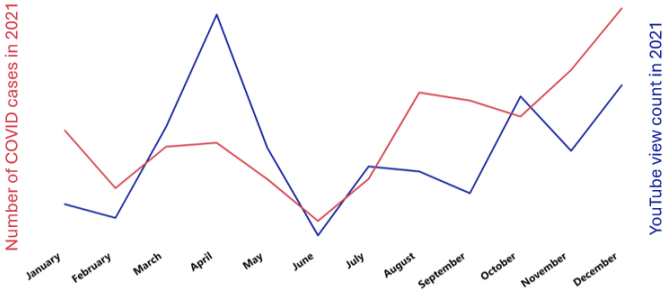

Fig. 6. The correlation of YouTube views and COVID cases in 2021.

E. Relevance of Biography Length

The study evaluated whether the length of a creator's biography influenced their video success. Data indicated no significant correlation between biography length and subscriber count or video performance. This finding suggests that short-form video viewers rarely engage with channel pages, focusing instead on content directly surfaced by platform algorithms. While biographies or video descriptions can be useful for branding or linking to other platforms, they play a negligible role in determining video success, as Figure 5. shows.

F. Use of Sounds and Music

The analysis highlighted the role of auditory elements in capturing and retaining viewer attention. Videos featuring trending sounds or music often performed better, as familiar audio can evoke emotional connections or nostalgia, enhancing viewer engagement. Additionally, soundtracks that align with video themes contributed to improved retention rates. Creators are advised to integrate relevant sounds or music strategically, as this can amplify the emotional impact and memorability of their videos.

G. Purpose and Advantages of Hashtags

Hashtags were identified as an effective tool for categorizing content and reaching specific audiences. [16] Videos using hashtags relevant to their content generally achieved higher discoverability, as hashtags help algorithms classify and recommend videos to interested viewers. [17] However, it is important to use hashtags in moderation, as using three to five targeted hashtags yielded the best results, while excessive or irrelevant hashtags could reduce a video's performance by confusing algorithms.

H. Limitations of Hashtag Usage

Conversely, the analysis explored potential downsides of hashtags. Overuse of hashtags, particularly those exceeding the recommended limit of five, often diluted their effectiveness. Videos with an excessive number of hashtags experienced lower average views and likes. By filtering data to exclude videos with more than five hashtags, the performance of hashtagged videos improved significantly, suggesting that relevance and restraint are key to leveraging hashtags effectively.

I. The Role of Timeliness and Trends

Timeliness emerged as a pivotal factor in content success. Videos aligned with current trends or significant events within their niche consistently outperformed those that were not. For instance, videos responding to popular challenges or incorporating trending themes gained more visibility due to increased viewer interest and algorithmic promotion. This analysis reinforces the value of monitoring industry trends and tailoring content to capitalize on moments of peak relevance.

J. Segment Specific Formats

Beyond trend-driven content, each segment has proven, long-standing formats. While originality is key, using established formats is also important. For example, in tech videos, unboxing and comparative reviews are popular. Unboxing involves setting up devices, while comparative reviews evaluate specs to determine the superior product. Using these formats effectively targets specific audience segments.

K. The Effect of COVID pandemic On YouTube Shorts

The global pandemic in 2020 led to a rise in mobile and computer use, which drove traffic to social media platforms [18]. A dataset on 2021 COVID-19 infections was compared with YouTube viewership data from the same year, analyzed monthly [19]. Although YouTube data from the 2020 lockdown period was unavailable (as YouTube Shorts launched in September 2020), the 2021 trends showed similar patterns. The line chart shows a correlation between viewership and COVID-19 infection rates, with lockdowns increasing free time and boosting YouTube consumption as more people stayed at home. This analysis (showed on Figure 6.) serves to illustrate that online content viewership and success can often be influenced by unexpected and seemingly unrelated external factors.

V. Conclusion

Success on social platforms depends mainly on consistency. Uploading high-quality content frequently helps build an audience faster than any guidelines.

Keep video titles under 40 characters to avoid truncation on mobile screens and to encourage clicks, as titles that give away too much may reduce user curiosity.

Adding questions, exclamation marks, and emojis to titles increases attention. Questions engage viewers, boosting interaction, which signals success to the platform's algorithm, expanding reach. Engaging with followers regularly helps build a community and makes interactions more memorable.

Effective use of sounds and visuals is key to maintaining attention. A video with lively visuals, subtitles, and background music keeps viewers more engaged than a plain version with just images and narration.

Hashtags should be relevant to the content and used sparingly — no more than five to reach a specific audience. Irrelevant hashtags won't improve visibility.

Certain factors, like YouTube biographies and video descriptions, didn't show a direct impact on video success, though they can help direct traffic elsewhere, such as to online stores or other profiles.

Staying current is vital. For example, game review channels must cover new releases, and cooking channels should target recipes for major holidays to stay relevant.

Proven formats in each segment can also boost success. While originality matters, utilizing established formats effectively targets specific audiences.

Content preferences vary by platform. Users on TikTok (ages 13-24) prefer different content than those on Facebook (ages 30-45). Leveraging platform-specific demographic data is crucial for tailoring content strategies.

References

[1] U. Shreya et al. "Articulating the construction of a web scraper for massive data extraction". In: (Feb. 2017). [Date of access: 10 04 2024].

[2] B. Zhao. *Web Scraping*. [Date of access: 10 04 2024]. Oregon State University, Corvallis, OR: Springer International Publishing, 2017.

[3] Scrapy. *Scrapy main page*. https://scrapy.org/. [Date of access: 10 04 2024].

[4] B. Pfalzgraf. *How to Use Selenium to Web-Scrape with Example*. https://towardsdatascience.com/how-to-use-selenium-to-web-scrape-with-example-80f9b23a843a. [Date of access: 10 04 2024]. Apr. 2020.

[5] B. Soup. *Beautiful Soup 4.12.0 Documentation*. https://www.crummy.com/software/BeautifulSoup/bs4/doc/. [Date of access: 10 04 2024].

[6] Web Scraper. *Web Scraper*. https://webscraper.io/. [Date of access: 10 04 2024].

[7] T. f. Salesforce. *What Is Data Visualization? Definition, Examples, And Learning Resources*. https://www.tableau.com/learn/articles/data-visualization. [Date of access: 01 05 2024].

[8] Eszter Lukács, Renáta Levendovics, and Tamas Haidegger. "Enhancing Skill Assessment of Autonomous Robot-Assisted Minimally Invasive Surgery: A Comprehensive Analysis of Global and Gesture-Level Techniques applied on the JIGSAWS Dataset". In: *Acta Polytechnica Hungarica* 20 (Jan. 2023), pp. 133–153. DOI: 10.12700/APH.20.8.2023.8.8.

[9] Csanád Ferencz and Máté Zöldy. "Neural Network-based Multi-Class Traffic-Sign Classification with the German Traffic Sign Recognition Benchmark". In: *Acta Polytechnica Hungarica* 21 (Jan. 2024), pp. 203–220. DOI: 10.12700/APH.21.7.2024.7.12.

[10] Peter Bednár, Juliana Ivaničáková, and Martin Sarnovský. "Semantic automatization of the data-analytical processes". In: *2022 IEEE 16th International Symposium on Applied Computational Intelligence and Informatics (SACI)*. 2022, pp. 000239–000242. DOI: 10.1109/SACI55618.2022.9919438.

[11] C. Dilmegani. *Is Web Scraping Legal? Ethical Web Scraping Guide*. https://research.aimultiple.com/web-scraping-ethics/. [Date of access: 04 12 2024]. Jan. 2024.

[12] DataDome. *Is Web Scraping Illegal?* https://datadome.co/guides/scraping/is-it-legal/. [Date of access: 04 12 2024]. June 2024.

[13] A. C. Telea. *Data Visualization principles and practice - second edition*. [Date of access: 10 04 2024]. CRC Press, 2007.

[14] S. R. Midway. *Principles of Effective Data Visualization*. https://www.cell.com/patterns/pdf/S2666-3899(20)30189-6.pdf. [Date of access: 01 05 2024].

[15] H. V. *7 Awesome Tips to Write YouTube Titles That Generate Views*. https://medium.com/illuminations-mirror/youtube-titles-e4f92125ded6. [Date of access: 02 12 2024]. July 2022.

[16] SproutSocial. *Hashtags: What they are and how to use them effectively*. https://sproutsocial-com.translate.goog/insights/what-is-hashtagging/?_x_tr_sl=en&_x_tr_tl=hu&_x_tr_hl=hu&_x_tr_pto=sc. [Date of access: 12 11 2024]. Mar. 2023.

[17] J. Zote. *How hashtags on Facebook still work for businesses in 2024*. https://sproutsocial-com.translate.goog/insights/hashtags-on-facebook/?_x_tr_sl=en&_x_tr_tl=hu&_x_tr_hl=hu&_x_tr_pto=sc. [Date of access: 23 01 2024]. Jan. 2024.

[18] Abida Sultana et al. "Digital screen time during the COVID-19 pandemic: a public health concern". In: *F1000Research* 10 (Feb. 2021), p. 81. DOI: 10.12688/f1000research.50880.1.

[19] Edouard Mathieu et al. "Coronavirus (COVID-19) Cases". In: *Our World in Data* (2020). https://ourworldindata.org/covid-cases.