Web Scraping and Review Analytics. Extracting Insights from Commercial Data

Andreea-Maria Tanasă
Simona-Vasilica Oprea
Adela Bâra
The Bucharest University of Economic Studies, Romania
tanasaandreea19@stud.ase.ro
simona.oprea@csie.ase.ro
bara.adela@ie.ase.ro

Abstract

Web scraping has numerous applications. It can be used complementary with APIs to extract useful data from web pages. For instance, commercial data is abundant, but not always relevant as it is presented on websites. In this paper, we propose the usage of web scraping techniques (namely, two popular libraries – BeautifulSoup and Selenium) to extract data from web and other Python libraries and techniques (vaderSentiment, SentimentIntensityAnalyzer, nltk, n consecutive words) to analyze the reviews and obtain useful insights from this data. A web scraper is built in which prices are extracted and variations are tracked. Furthermore, the reviews are extracted and analyzed in order to identify the relevant opinions, including complaints of the customers.

Key words: web scraping, price tracking, sentiment analysis, alerts
J.E.L. classification: Z13, C55, C81

1. Introduction

Web scraping has numerous applications for Human Resources (HR) companies, banks for their competitors, election, projects data to start with, marketing (Boegershausen et al., 2022), price comparison, public sentiment regarding Bitcoin, psychology (Landers et al., 2016), patterns of depression and suicidal thoughts (https://www.sas.com/en_ca/insights/articles/analytics/using-big-data-to-predictsuicide-risk-canada.html), etc. Data that can be extracted using web scraping may be included in the big data paradigm. A case study using Python and web scraping to extract big data in psychology filed is provided in (Landers et al., 2016). They also debate legal and ethical aspects related to web scraping projects. Nonetheless, some researchers focused on data validity, legal and ethical concerns regarding the collection of data from web. They reviewed more than 300 papers related to web usage in marketing. Several interesting questions were posed in this research, such as: "What information to extract? How to sample? At what frequency to extract information? How to process the information?" (Boegershausen et al., 2022).

Furthermore, there are many programming languages, techniques and tools that facilitate web scraping (Khder, 2021), (Saurkar and Gode, 2018). Web scraping is a technique to extract, parse, and organize data from the web in an automated way. Many webpages nowadays offer an API to access the data in a structured manner. Both APIs and web scraping are used and applied to investigate credibility of information on Twitter. They offer different views to extract data. On one hand, the study showed that some Twitter attributes cannot be extracted using only web scraping. Both methods required extensive pre-processing on tweets (e.g., normalization) and generate similar credibility values. Web scraping proved to be faster and more flexible than Twitter API. On the other hand, web scraping failed more often due to the webpage changes (Dongo et al., 2021).

However, APIs may not expose the desired functionality (Glez-Peña et al., 2013). The API may be rate limited: meaning that the data can be only accessed a number of certain times per second or per day. The API may not expose all the required data whereas the website does, therefore scaping is necessary in some cases. "Why is more efficient to combine BeautifulSoup and Selenium in scraping for data under energy crisis" was investigated by two authors (https://stec.univ-ovidius.ro/html/anale/RO/2022-issue2/Section%201%20and%202/19.pdf). They compared the scraping performance of the two Python libraries, demonstrating that the combination of BeuatifulSoup and Selenium is better when both data extraction and dynamic actions are required to obtain data.

2. Theoretical background

Several use cases were described in (vanden Broucke and Baesens, 2018). The authors presented cases when a researcher has to extract data that is not available on research platforms (such as Kaggle, ResearchGate). Scrape once and extract data is a case that is usually applied for proof-of-concept applications. Usually, data for new projects requires multiple data sources from where various time series can be merged into an updated data set. Another case can be to use scraping at each execution of an application to get data from invertors for daily forecast, for instance. The latter requires a more robust approach, to consider and treat more scenarios, and the availability of the webpage.

The so-called fair use of web scraping data is defined in the Copyright or Trademark Infringement of the United States. No explicit permission is required if the data is included in the academic research. Most commercial usage of copyrighted material requires explicit permission, however. Furthermore, in the Computer Fraud and Abuse Act (CFAA), it is stated that "intentionally accesses a computer without authorization… and as a result of such conduct recklessly causes damage" are considered an abuse, especially if the webpage owner can demonstrate damage or loss. Moreover, in Europe, there are regulations, directives and Computer Misuse Act and Trespass to Chattels. However, web scraping, especially on a large scale (also known as crawling when more sources are repeatedly scraped) for commercial purposes may lead to legal implications if permission is not obtained (Krotov, Johnson and Silva, 2020).

Several alternatives to BS and Selenium libraries do exist. They are PHP (curl package), R (rvest package), etc. Furthermore, Python libraries (like Scrapy) offer an alternative. However, a notable drawback of Scrapy is that it does not emulate a full browser, hence handling JavaScript could be difficult with this library. CatchControl is a tool offered by Python to avoid continuously stressing servers with requests, which is especially handy during development of scraping scripts where frequently a script is restarted to check if a bug has been fixed, the expected results are obtained, and so on. Moreover, there are interesting graphical scraping tools, such as: Portia, Parsehub, Kapow, Fminer, Dexi. Their main disadvantage is that there are scraping issues with heavy loaded JavaScript. Sometimes, these ready-made tools fail when the page is built in a less-straightforward manner. Web scraping can be a cat-and-mouse challenge as some webpage designers work hard to avoid or prevent scraping. For instance, Completely Automated Public Turing test to tell Computers and Humans Apart (CAPTCHA) designers are challenged by artificial intelligence and deep learning implementations (https://www.npr.org/sections/thetwo-way/2017/10/26/560082659/ai-model-fundamentally-cracks-captchas-scientists-say) that are used to pass such obstacles (Olufemi et al., 2021).

Sentiment Analysis or opinion mining is the process of assessing if a text is positive, neutral or negative. It extracts an opinion from texts that can be sometimes lengthy in order to clearly show the attitude of a customer. This process is applied especially in marketing, politics and public actions. Business companies are interested in finding out customers' opinions and feelings to craft their strategies. They are interested in understanding customers' sensitivity to campaigns and new products and finding out why some products are not desired. In politics, text mining is also relevant as it reveals the consistency and inconsistency of political opinions. Furthermore, public opinion can be analyzed, and estimations of election outcome may be designed.

Political statements are investigated to measure their impact on the economy and on particular assets, such as Bitcoin, electric cars, RES technologies, etc. Public actions can be monitored to understand social movements and identify moods and events that can emerge from public initiatives.

In this paper, the goal is to investigate how do the prices change for a popular category of tech products in a short period of time, as well as retrieving and studying the attitude of the customers towards that specific product. Only the products with a significant number of reviews will be used for analysis, reducing the large sample of products to a much smaller one, with more significant elements. A similar thinking will be applied for the price tracking over the selected period.

3. Research methodology

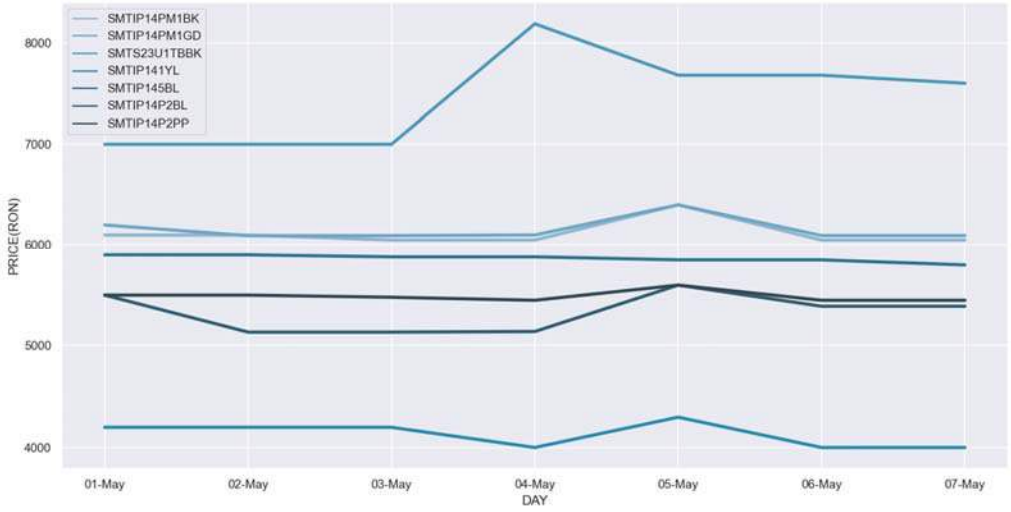
In this paper, a combination of Python libraries is proposed to extract and investigate data from the web. Python has a variety of libraries that interact with HTTP. For instance, *httplib2* is a small, fast HTTP client library. Originally it was developed by Googler Joe Gregorio, and now it is supported by the community. Another library is *urllib3* that is a powerful HTTP client for Python. *Requests* is also a popular library. *Grequests* extends requests to deal with asynchronous HTTP. *Aiohttp* is another library focusing on asynchronous HTTP. *Requests* and *get* method with a webpage string as parameter (url) create a page object. For static page elements, in order to obtain page's information, a BeautifulSoup (BS) object is created. This library is especially useful for finding html elements in the webpage source. For this purpose, the BS library relies on an HTML parser. In Python, multiple parsers do exist, such as: *html.parser* that is a built-in Python parser that is useful especially when using recent versions of Python 3 and requires no extra installations. *Lxml* is very fast parser, but it requires extra installations. *Html5lib* parses the webpage as a web browser, but it is slower. BS's main purpose is to transform the HTML into a tree-based representation. Extracting the content is facile using a BS object, and the two methods (find and findAll) fetching the data from the webpage. However, *requests* and BS are not enough to deal with script tags. Selenium is a powerful web scraping tool that was originally developed for the purpose of automated website testing. Selenium operates by automating various browsers to load a webpage, retrieve its contents, and control it like a user would when using the browser, clicking on buttons, on links, etc. Thus, it is a powerful tool for web scraping focusing on the dynamic actions that a user might do. Selenium can be used by several programming languages (Java, C#, PHP, and of course, Python). However, Selenium does not come with its own web browser and requires a WebDriver to interact. WebDrivers are available for browsers, including Internet Explorer, Chrome, Firefox, Edge, Safari, etc. With these WebDrivers, a browser window will open and simulate actions included in the Python code. The WebDriver has to be downloaded (https://sites.google.com/a/chromium.org/chromedriver/downloads) and its path inserted into Advanced System Setting Environment variables PATH or locate the WebDriver in the same directory as the Python scripts. Selenium finds elements and perform actions using several methods (like send_keys). Both libraries run on online platforms such as Google Collaboratory. Sentiment analysis using *vaderSentiment, SentimentIntensityAnalyzer* and *nltk* libraries are applied to identify whether the reviews are positive, neutral or negative. Moreover, the content of reviews is investigated measuring the combination of *n* consecutive words in order to identify the most frequent complaints.

4. Findings

The purpose of this section is to provide the price tracking results and review analysis. The data analysed was extracted from an e-commerce website of a Romanian retailer which focuses on technology related products and the category chosen for the analysis is phones. Using *BeautifulSoup*, the following data was extracted and processed to match the purpose of the analysis: the product's code, the URL to the product's information page, details about the price (the initial price, discount and the final price), as well as data related to reviews which will be used later, such as the total rating from the reviews and the number of them.

The movement of the prices has been followed during two separate weeks: 15-April to 21-April and 01-May to 07-May. The number of products varied from one day to another, but after merging the tables, 386 products remained. Since there are over 385 products daily in the chosen category, the purpose of the analysis was to seek the products for which prices fluctuated the most in each of those two weeks. A sample was taken from the total number of products for each week, consisting in 11 products for April and 7 products for May, so as to clearly see the evolution for the selected products in the diagrams from Figure 1 and Figure 2.

Figure no. 1. Price fluctuation for the week 01 - 07 May



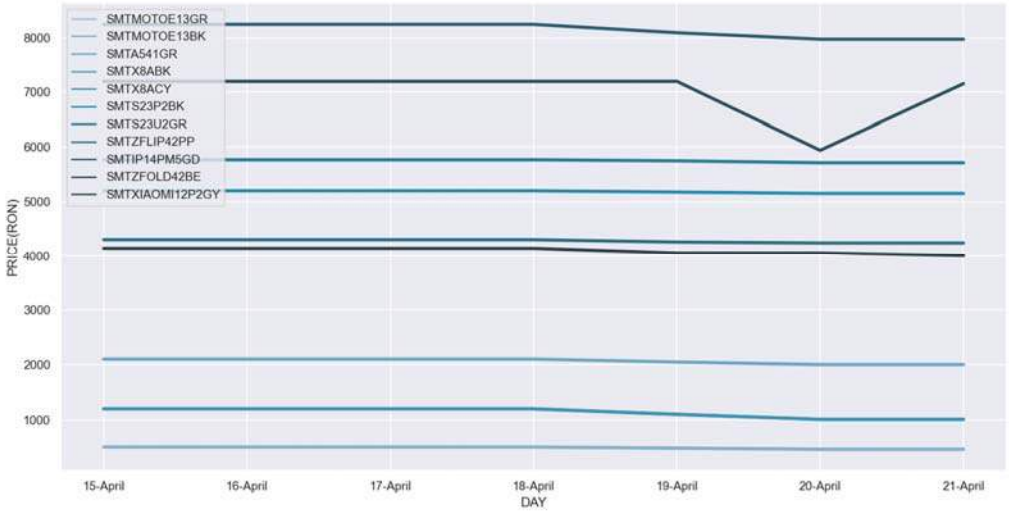
Source: the authors

Figure 1 represents the fluctuation in the prices for May, which, compared to Figure 2, representing the prices in April, is more dynamic. If we look at the names of the products corresponding to these codes, we observe the following connections:

- SMTIP14PM1BK → APPLE iPhone 14 Pro Max 5G, 128GB, Space Black
- SMTIP14PM1GD → APPLE iPhone 14 Pro Max 5G, 128GB, Gold
- SMTS23U1TBBK → SAMSUNG Galaxy S23 Ultra 5G, 1TB, 12GB RAM, Dual SIM, Phantom Black
- SMTIP141YL → APPLE iPhone 14 5G, 128GB, Yellow
- SMTIP145BL → APPLE iPhone 14 5G, 512GB, Blue
- SMTIP14P2BL → APPLE iPhone 14 Plus 5G, 256GB, Blue
- SMTIP14P2PP → APPLE iPhone 14 Plus 5G, 256GB, Purple

As we can see, the products which suffered multiple price changes within a single week were the latest phones from the brands Apple and Samsung. SMTS23U1TBBK, the Galaxy S23 Ultra 5G, had the most sudden change, a raise of 17%, followed by a fall in the next two days. For this week the prices had a predominantly upward trajectory, with visible changes from one day to another. For the week 15 April to 21 April, there were not many sudden price fluctuations, but more subtle differences suggesting that mostly the prices went down. We see that the most visible change was for the product with the code SMTZFOLD42BE, which had a dramatic decrease on 20 April, then went back up a day later. The name of the product impacted by this change was "SAMSUNG Galaxy Z Fold4 5G, 256GB, 12GB RAM, Dual SIM, Beige", another one of most recent Samsung releases.

Figure no. 2 Price fluctuation for the week 15 - 21 April



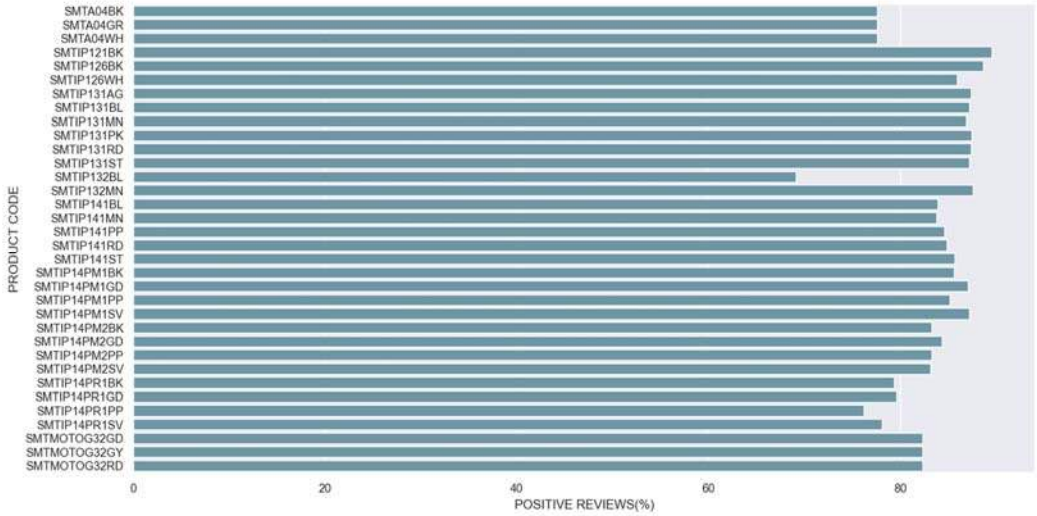
Source: the authors

It can be concluded that the products which are most prone to price changes from one day to another during the mentioned two-week period are, more often than not, the latest released phones.

A sample of 392 tech products was extracted from the website, only from the phones category. From each device, all the reviews were extracted, alongside the review score and the number of reviews, which will be later used to compare the response received from the sentiment analysis to the overall opinion of the product. The products were filtered by the number of reviews, excluding those with no reviews (6,12%) and those which received less than 100 reviews (75,77%). All the reviews were translated to English using *googletrans Translator*, a free Python library which uses the Google Translate API to be able to translate at once pieces of text with a maximum of 15000 characters (https://py-googletrans.readthedocs.io/en/latest/). The reviews were selected and analysed with the purpose of establishing the customer's feeling towards the acquired product.

Sentiment analysis was used to determine the opinion of the buyer by placing it in the pertaining category: positive, negative or neutral. Each review returned a scoring system which consisted of three values: negative, neutral and positive, all of which should add up to 1 to create a whole. The last component of the scoring system is called 'compound', which was used as the overall score for each review (https://vadersentiment.readthedocs.io/en/latest/pages/about_the_scoring.html). The review was considered positive if it had a compound of at least 0.5 and negative if it was less than or equal to -0.5. Anything in between was considered neutral. Only the devices with more than 100 written reviews were selected, the percentage of the positive reviews was calculated for each one of them and the results are presented in Figure 3.

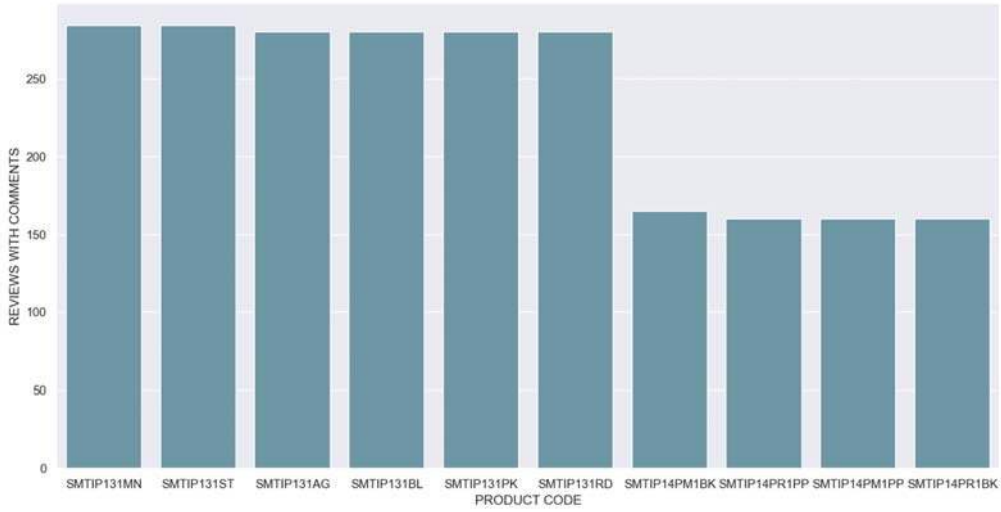
Figure no. 3 Percentage of positive reviews



Source: the authors

In Figure 3, it is clearly seen that the percentage of positive reviews is well over 65%. The lowest one (69.15%) belongs to code SMTIP132BL, corresponding to the product "APPLE iPhone 13 5G, 256GB, Blue" and the biggest score (89.57%) belongs to product SMTIP121BK which is the product "APPLE iPhone 12 5G, 128GB, Black ". Product SMTIP121BK had an overall score of 4.87 out of 115 reviews, while product SMTIP132BL had an overall score of 4.90 out of 295 reviews. By looking at the difference between the grade-only reviews and the written ones, for SMTIP121BK there are 113 written reviews and 2 grade-only, whereas for SMTIP132BL there are 110 written reviews and 185 grade-only. From this, it can be deducted that the higher grade for the device with the lowest percentage of positive scores comes from the grade-only reviews which could not be included in the sentiment analysis. The top 10 devices with the most written reviews are represented in Figure 4. Upon searching the product codes, it was observed that all the products correspond to the latest iPhones models from Apple: iPhone 13, iPhone 14 Pro and iPhone 14 Pro Max, with varying specifications.

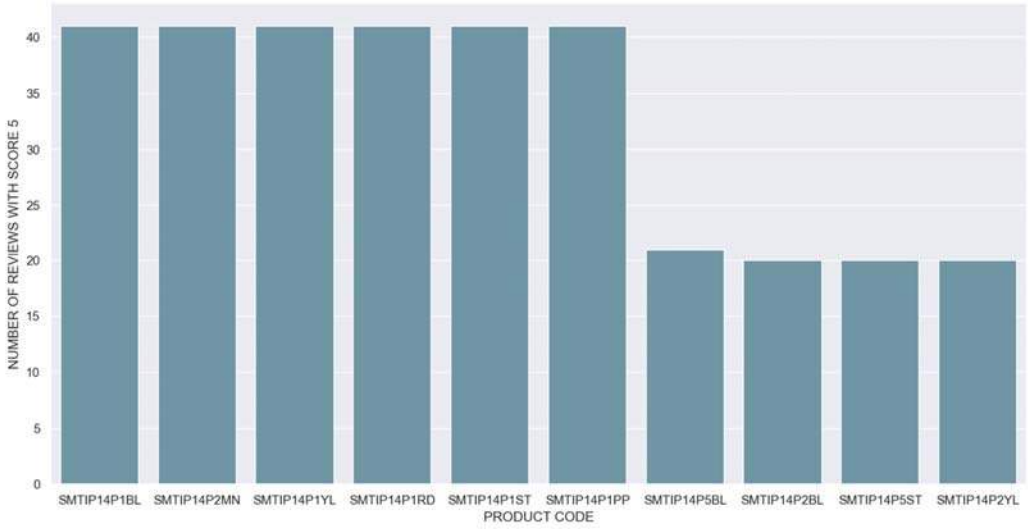
Figure no. 4 Top 10 devices based on written reviews



Source: the authors

Product code SMTIP131MN, which is one of the devices with the most written reviews, corresponds to "APPLE iPhone 13 5G, 128GB, Midnight" with a review score of 4.91 from a total of 290 reviews. Out of these 290 reviews, 284 consisted of written reviews with 252 having a positive sentiment (86.9%) and 33 having a negative sentiment (11.38%). On the other side of this top, product SMTIP14PR1BK corresponds to "APPLE iPhone 14 Pro 5G, 128GB, Space Black" and has 160 written reviews out of 160 reviews in total. The written reviews were analyzed, which led to the conclusion that 127 reviews were found with a positive sentiment (79.38%) and 25 (15.62%) with a negative sentiment, having a 4.89 review score. It can be concluded that the latest iPhones were the most requested, leading, thus, to a great number of written reviews, whereas for many other devices the review was only resumed to the grade itself.

Figure no. 5 Devices with a perfect score rated by number of reviews

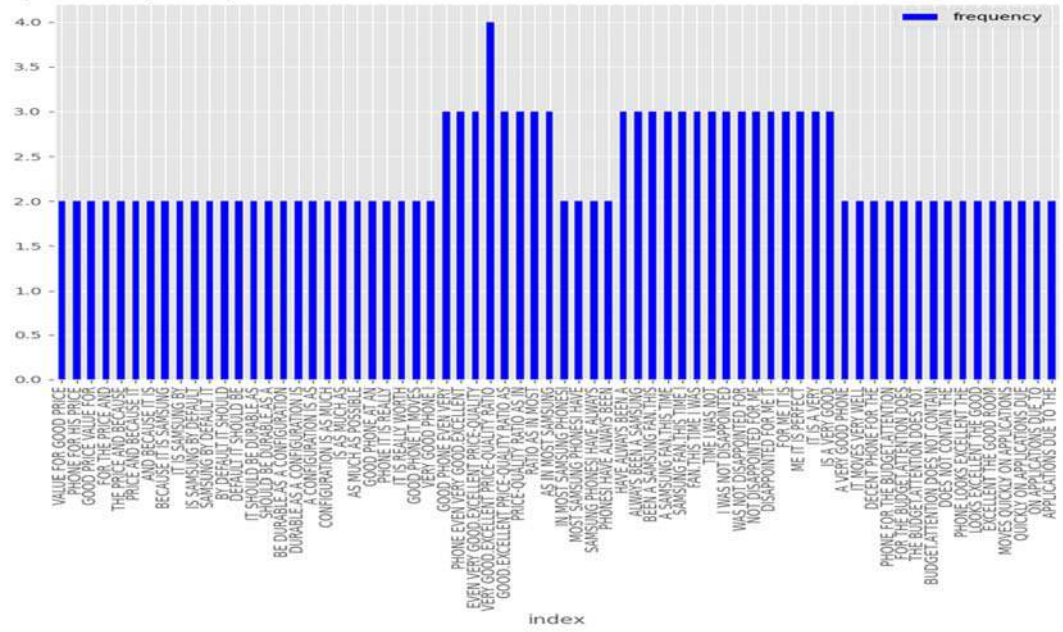


Source: the authors

Figure 5 aims to highlight the devices with a perfect review score, 5 out of 5, taking into account only the phones which have at least 15 reviews, sorted in a descending order by the number of reviews. As we can see, the first 6 devices have the same number of reviews, 41, while the remaining 4 have a significantly lower number, 21 and 20. All the devices present in the above chart are different versions of the same product, iPhone 14 Plus 5G, which have a few distinctions such as color and storage capacity. Just as in Figure 4, it seems that the latest versions of iPhones have the highest ratings, being among one of the most popular devices.

A n-gram analysis is performed for three devices in order to identify common segments of words based on their frequency. The first device is Phone: SAMSUNG Galaxy A04s, 32GB, 3GB RAM, Dual SIM, Black; Review score 4.73; Number of reviews: 116; Code: SMTA04BK. The total number of 4-grams is 1243. Analyzing the frequencies that are higher than 2, we obtained the following text elements that all reveal positive opinions of customers. A selection of opinions is displayed in Figure 6.

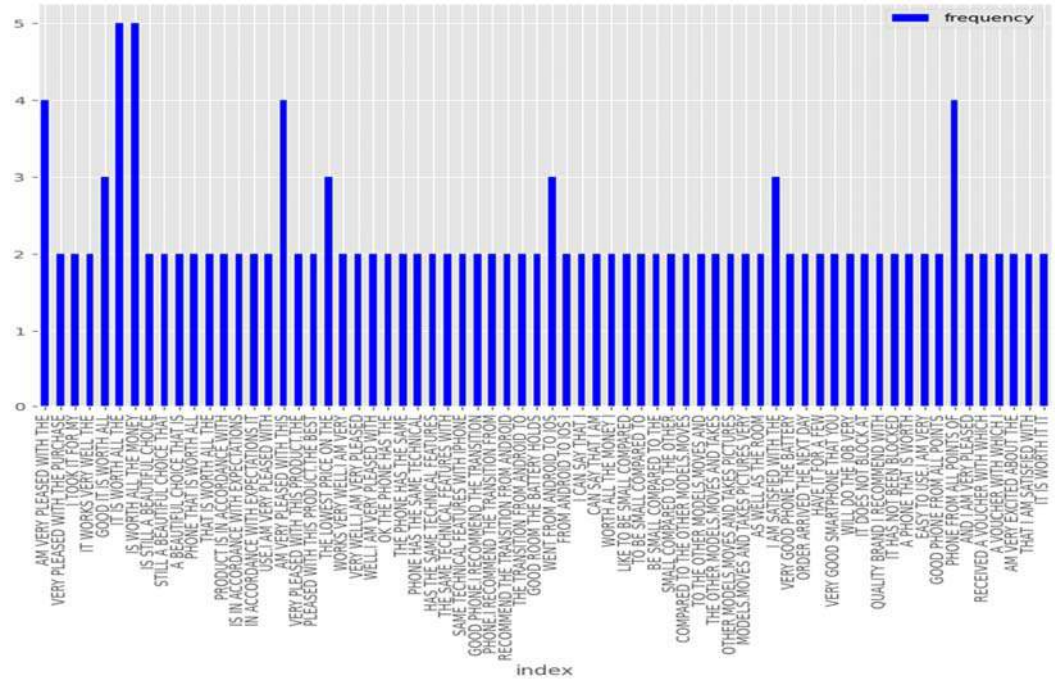
Figure no. 6 Opinions for SMTA04BK



Source: the authors

The second device is Phone: APPLE iPhone 14 5G, 128GB, Midnight; Review score 4.82; Number of reviews: 101; Code: SMTIP141MN. The total number of 5-grams is 5770. The higher number compared to the previous device convinced us to increase the analysis from 4 to 5-grams. Analyzing the frequencies that are higher than 2, we obtained the following text elements that again reveal positive opinions of customers. A selection of opinions with the highest frequencies is displayed in Figure 7.

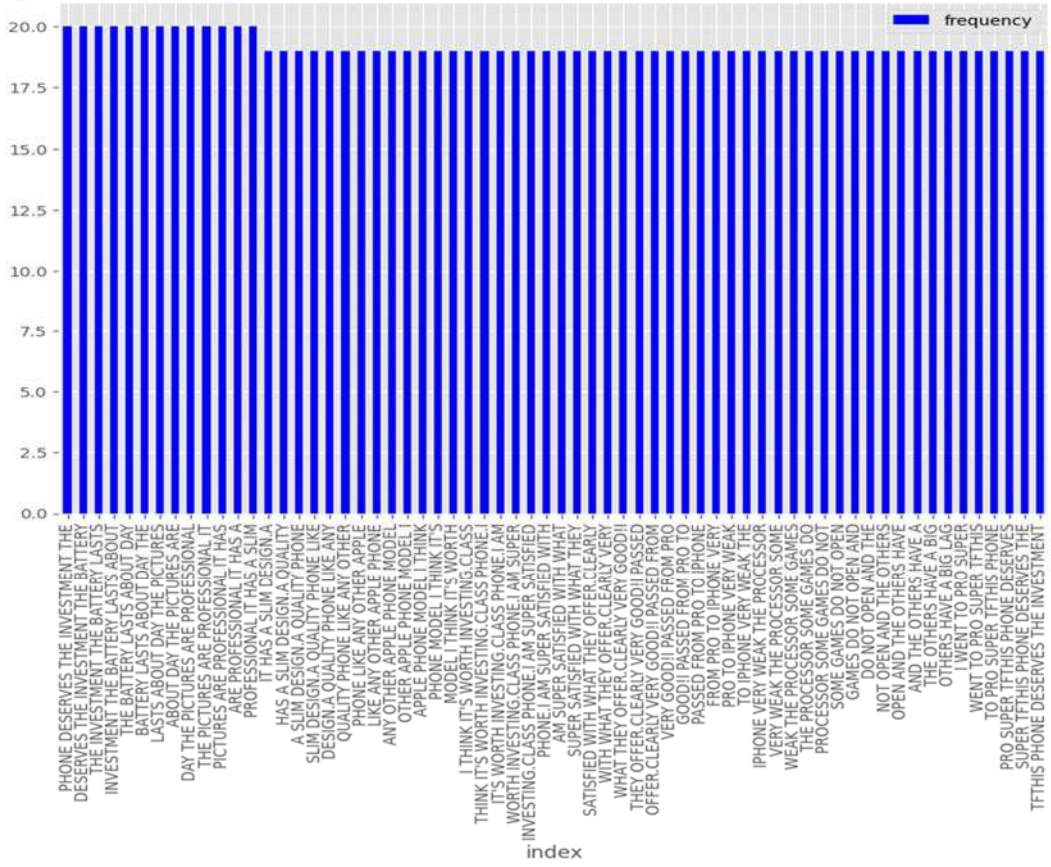
Figure no. 7 Opinions for SMTIP141MN



Source: the authors

The third device is Phone: APPLE iPhone 14 5G, 512GB, Purple; Review score 4.82; Number of reviews: 101; Code: SMTIP145PP. This phone is indicated to have a lower positive opinion. The total number of 5-grams is 119. Analyzing the frequencies that are higher than 2, we obtained the following text elements.

Figure no. 8 Opinions for SMTIP145PP



Source: the authors

Analyzing the higher frequencies, one can notice negative opinions, such as IPHONE VERY WEAK THE PROCESSOR; OTHERS HAVE A BIG LAG, etc.

5. Conclusions

In this paper, we scraped and tracked changes in prices for a popular category of tech products in a short period of time. The attitude of the customers towards that specific product was analyzed using sentiment analysis techniques. Only the products with a significant number of reviews were used for analysis, reducing the large sample of products to a much smaller one, with more significant elements.

We analyzed the prices over two weeks in April and May and we found out that the products which suffered multiple price changes within a single week were the latest released phones from the brands Apple and Samsung.

Sentiment analysis using *vaderSentiment, SentimentIntensityAnalyzer* and *nltk* libraries were applied to identify whether the reviews are positive, neutral or negative. Moreover, the content of reviews was investigated measuring the combination of *n* consecutive words in order to identify the most frequent complaints.

6. Acknowledgement

This work was supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS-UEFISCDI, project number PN-III-P4-PCE-2021-0334, within PNCDI III.

7. References

- Boegershausen, J., Datta, H., Borah, A. and Stephen, A.T., 2022. Fields of Gold: Scraping Web Data for Marketing Insights. *Journal of Marketing*. 86(5), pp. 1-20, https://doi.org/10.1177/00222429221100750.
- vanden Broucke, S. and Baesens, B., 2018. *Practical Web Scraping for Data Science*. *Practical Web Scraping for Data Science*, https://doi.org/10.1007/978-1-4842-3582-9.
- Dongo, I., Cardinale, Y., Aguilera, A., Martinez, F., Quintero, Y., Robayo, G. and Cabeza, D., 2021. A qualitative and quantitative comparison between Web scraping and API methods for Twitter credibility analysis. *International Journal of Web Information Systems*. 17(6), pp. 580-606, https://doi.org/10.1108/IJWIS-03-2021-0037.
- Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M. and Fdez-Riverola, F., 2013. Web scraping technologies in an API world. *Briefings in Bioinformatics*. 15 (5), pp. 788–797, https://doi.org/10.1093/bib/bbt026.
- Khder, M.A., 2021. Web scraping or web crawling: State of art, techniques, approaches and application. *International Journal of Advances in Soft Computing and its Applications*. 13(3), pp. 145-168, https://doi.org/10.15849/IJASCA.211128.11.
- Krotov, V., Johnson, L. and Silva, L., 2020. Tutorial: Legality and ethics of web scraping. *Communications of the Association for Information Systems*. 47(1), pp. 555-581, https://doi.org/10.17705/1CAIS.04724.
- Landers, R.N., Brusso, R.C., Cavanaugh, K.J. and Collmus, A.B., 2016. A primer on theory-driven web scraping: Automatic extraction of big data from the internet for use in psychological research. *Psychological Methods*. 21(4), pp. 475-492, https://doi.org/10.1037/met0000081.
- Olufemi, I.E., Adebiyi, A.A., Ibikunle, F.A., Adebiyi, M.O. and Oludayo, O.O., 2021. Research trends on CAPTCHA: A systematic literature. *International Journal of Electrical and Computer Engineering*. 11(5), pp. 4300-4312, http://doi.org/10.11591/ijece.v11i5.pp4300-4312.
- Saurkar, A. V and Gode, S.A., 2018. An Overview On Web Scraping Techniques And Tools. *International Journal on Future Revolution in Computer Science & Communication Engineering*, https://api.semanticscholar.org/CorpusID:198993824.