ADV MATH SCI JOURNAL

Advances in Mathematics: Scientific Journal **9** (2020), no.6, 3703–3711
ISSN: 1857-8365 (printed); 1857-8438 (electronic)
https://doi.org/10.37418/amsj.9.6.48 Spec Issiue on ICAML-2020

DYNAMIC EXTRACTION AND ANALYTICS OF BIG DATA FROM CLOUD AND SOCIAL MEDIA INTEGRATED PLATFORMS

N. OBEROI[1], S. SACHDEVA, P. GARG, AND R. WALIA

ABSTRACT. The dynamic information extraction and examination is one of the key research segments in the real time applications including information investigation, criminal information examination, digital forensics, feelings mining, statistical surveying and numerous others. This methodology is also called web scratching and generally utilized in the prescient mining and information revelation progressively with the goal that the genuine information about the particular individual or item can be perceived from internet based life. Comparative kind of usage is finished by the ideological groups to get the resident audits about their gathering with the probabilities to win in the decisions. Furthermore, such approaches are additionally utilized by the corporate goliaths to get the criticism about their item from overall population. The presented work is depicting the utilization scenarios and extraction of dynamic information from the online networking stages so that the investigation should be possible adequately. The presented manuscript is depicting the usage patterns of big data obtained from social media and to have the predictive mining on the sentiments and emotions analytics using high performance libraries from the social media portals on cloud.

[1] *corresponding author*

2010 *Mathematics Subject Classification.* 68M11.

Key words and phrases. Big Data Analytics, Big Data Extraction, Big Data on Cloud, Big Data Analytics from Social Media.

1. Introduction

The usage of social media is having huge prominence from last decade and most of the users are getting benefits from such services. A number of social platforms are available in which the common people are sharing their views and messages. These are used for assorted applications [1]. The extraction of real time data from social media and online portals is done and the user reviews are extracted to check what the people are discussing about their products and services. This is more diverted towards user belief mining or sentiment mining [2].

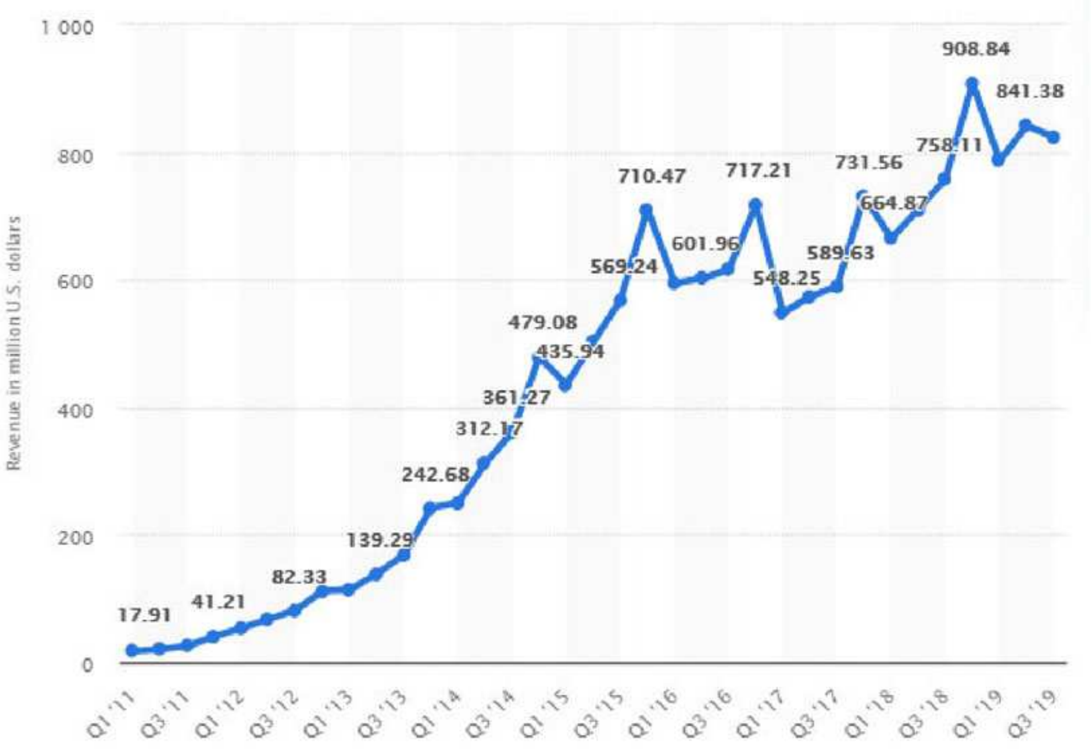


FIGURE 1. Twitter Quarterly Revenue

Key Social Media for Big Data Extraction and Usage Analytics include Facebook, Instagram, Tumblr, YouTube, Twitter, Spaces, VK, Wattpad, WeChat, TikTok, Reddit, QQ, Qzone, and many others.

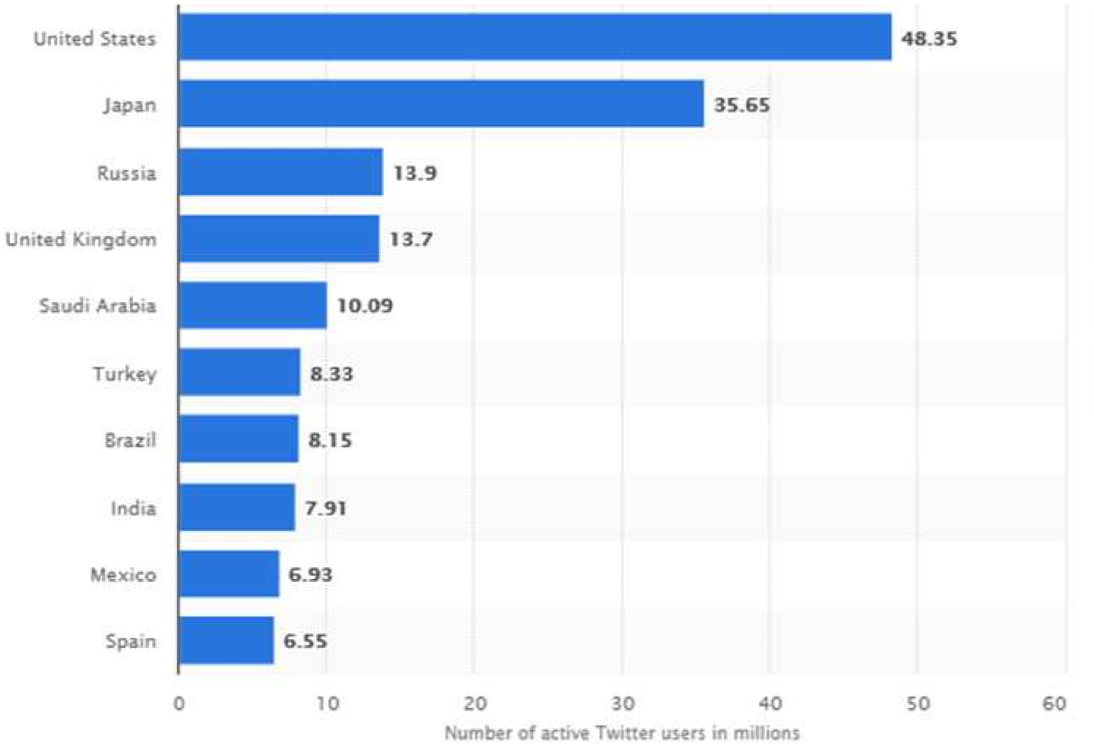


FIGURE 2. Twitter Usage Patterns in Global Perspectives

2. REVIEW OF LITERATURE

Fuchs et al. in 2017, in [1], presented the work on the big data analytics with the patterns analysis for the social media research. The work is having the key base on the cavernous analytics of datasets obtained from social platforms for predictive mining.

S. J. Miah et al. in [2], underlined the behavioral analytics on the datasets of tourism industry and that is quite effectual for the multidimensional big data evaluation patterns in assorted perspectives.

J. Li et al., [3], presented the approaches as well as the challenges with the feature points extraction and presentation to the front end so that the analysis can be done for assorted domains.

I. Lee et al. in [4], depicted the usage patterns with the impacts and evolution towards the big data analytics to have the effectual outcomes and knowledge discovery with the key points on the big data based evaluation.

Z. Xiang et al. in [5], undertaken the domain of smart tourism for the elevation of revenue and identification of the key basis for the dynamic data transformation obtained from social media.

Z. Lv et al. in [6], underlined the big data analytics with the key research aspects and dimensions towards the challenges so that the cumulative performance can be evaluated on different aspects of data fetched.

M. D. Lytras et al. in [7], presented the work associated with the human decision making process with the analytics of the datasets of big data based platforms and portals.

S. Stieglitz et al. [8], underlined the work with the challenges and the key perspectives of social media mining and presentation to the knowledge discovery applications for machine learning and further predictive mining.

3. Key Research Dimensions in Web Scraping

Cyber Patrol: Analysis of messages on social media about particular crime or discussion about particular crime Cyber Parenting: Extraction of data from web portals which are in discussion by the kids or teenagers so that their emotions and behavior can be analyzed. The Cyber Parenting Approaches to guard the kids against inappropriate portals can be integrated with these implementations.

Corporate Applications: Fetching the reviews of products and services from e-commerce portals so that the companies can improve their products as per the requirements and feedbacks of the users, [9].

3.1. **Prominent Tools for Web Scraping and Real Time Data Extraction.**

A number of tools and libraries are available for extraction of real time data from web pages. Following is the table in which assorted tools are mentioned with the associated URL from where these libraries can be installed.

3.1.1. *Python Based Tools for Web Scraping.*

From last few years, Python has gained huge prominence as a high performance programming language for multiple applications, [4]. Python is now days used in Cyber Security, Digital Forensics, Web Analytics, Deep Learning, Machine Learning, Grid Computing, Parallel Computing, Cloud Applications, Web Scraping and many other domains which needs higher degree of performance and accuracy with minimum error rate, [5]. Python is having more than 2 lac packages in its repository pypi.org to enrich the programmers, [6], and researchers so that they can work on diversified domains with huge flexibility, [7]. For Web Scraping and Real Time Data Extraction, [8], there are more than 10,000 projects and code libraries which can be used by the developers for their implementations, [9]. The main difference between web scraper and web parser is that the web scrapers are used to extract the data from web applications, [10] while the web parsers are used to break down and analyze the scraped data, [11]. In broader ways, the web scrapers are used and after that web parsers are integrated to analyze the extracted data for knowledge discovery and storage for particular applications.

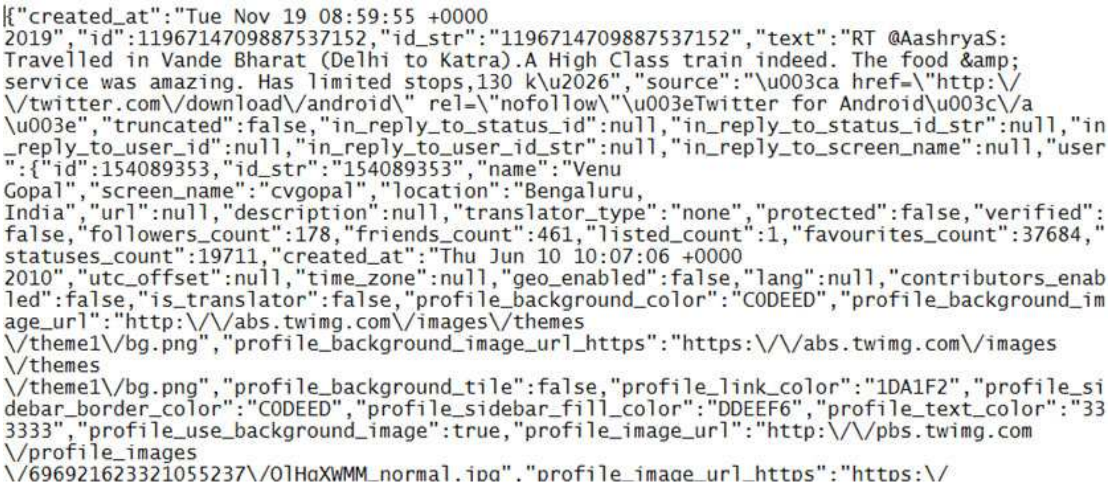
3.1.2. *BeautifulSoup.* URL: https://www.crummy.com/software/BeautifulSoup Python integrates enormous libraries and frameworks for the web scraping and data analytics, [12]. Besides the huge list of libraries of Python for data scraping, BeautifulSoup is widely used for the data extraction and logging of outcomes from multiple real time channels, [13]. The key benefits of BeautifulSoup, [14] involves that it is compatible with Python 2 as well as Python 3, Fast and High Performance Library, Dynamic Parsing and Extraction of Web Pages, Deep Searching, Navigation and Parsing of Content on Web Pages, Real Time Extraction of data without delay, Dynamic Storage of Web Data using HTTP Requests to the Web URL, Extraction of data and transformation in multiple formats including JSON, XML, CSV, TXT, XLS, XLSX and many others.

3.2. **Scraping Messages and User Timeline from Twitter Social Media.** Many times it is required to analyze the user feedbacks and reviews from social media about the particular products or services. Generally, the people share their views and feedback about products on Twitter, Facebook, Instagram and other similar social platforms, [15]. The use of web scraping can be done to extract these messages and timelines so that the users' sentiments can be evaluated for improving the services or product features, [16].

3.3. **Scraping Images and Timelines from Instagram.** Instagram is nowadays very popular platform for sharing the multimedia content using mobile based platform. The user broadcasted content on Instagram can be fetched using BeautifulSoup library. One of the key applications of this implementation can be to evaluate the current status of a particular user so that the behavior and interests of that particular user can be identified. The following code snippet can be used to extract the images and content associated with the specific hashtag of Instagram and it can be beneficial for the researchers working on specific news coverage linked with the particular hashtag. In addition, the criminal investigation authorities can extract all the messages or files linked with the specific hashtag with the timeline of users.

```
html = urllib.request.urlopen(url,
context=this.ctx).read()
    bsoup = BeautifulBsoup(html, ’html.parser’)
    myscript = bsoup.find(’myscript’, text=lambda t: \
                       t.startswith(’window._sharedData’))
    page_json = myscript.text.split(’ = ’, 1)[1].rstrip(’;’)
    data = json.loads(page_json)
    for instapost in data[’entry_data’]
    [’TagPage’][0][’graphql’
            ][’imyghashtag’][’edge_imyghashtag_to_media’]
            [’edges’]:
        image_src = instapost[’node’]
        [’thumbnail_resources’]
        [1][’src’]
        myhs = open(imyghashtag + ’.txt’, ’a’)
        myhs.write(image_src + ’\n’)
        myhs.close()
def main(this):
    this.ctx = ssl.create_default_context()
    this.ctx.check_hostname = False
    this.ctx.verify_mode = ssl.CERT_NONE
    with open(’imyghashtagdata.txt’) as f:
        this.content = f.readlines()
    this.content = [x.strip() for x in this.content]
    for imyghashtag in this.content:
        this.getlinks(imyghashtag,
        ’https://www.instagram.com/explore/tags/’+
        imyghashtag + ’/’)
```

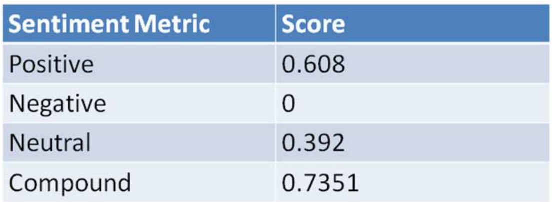
In addition, Twitter also provides a specific library titled Tweepy for web scraping with the focus of research and development. By this package, all the tweets can be extracted from social media about a particular search keyword mentioned in the code.



```
{"created_at":"Tue Nov 19 08:59:55 +0000
2019","id":1196714709887537152,"id_str":"1196714709887537152","text":"RT @AashryaS:
Travelled in Vande Bharat (Delhi to Katra).A High Class train indeed. The food &amp;
service was amazing. Has limited stops,130 k\u2026","source":"\u003ca href=\"http:\/
\/twitter.com\/download\/android\" rel=\"nofollow\"\u003eTwitter for Android\u003c\/a
\u003e","truncated":false,"in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in
_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_reply_to_screen_name":null,"user
":{"id":154089353,"id_str":"154089353","name":"Venu
Gopal","screen_name":"cvgopal","location":"Bengaluru,
India","url":null,"description":null,"translator_type":"none","protected":false,"verified":
false,"followers_count":178,"friends_count":461,"listed_count":1,"favourites_count":37684,"
statuses_count":19711,"created_at":"Thu Jun 10 10:07:06 +0000
2010","utc_offset":null,"time_zone":null,"geo_enabled":false,"lang":null,"contributors_enab
led":false,"is_translator":false,"profile_background_color":"C0DEED","profile_background_im
age_url":"http:\/\/abs.twimg.com\/images\/themes
\/theme1\/bg.png","profile_background_image_url_https":"https:\/\/abs.twimg.com\/images
\/themes
\/theme1\/bg.png","profile_background_tile":false,"profile_link_color":"1DA1F2","profile_si
debar_border_color":"C0DEED","profile_sidebar_fill_color":"DDEEF6","profile_text_color":"33
3333","profile_use_background_image":true,"profile_image_url":"http:\/\/pbs.twimg.com
\/profile_images
\/696921623321055237\/OlHqXWMM_normal.jpg","profile_image_url_https":"https:\/
```

FIGURE 3. Twitter Usage Patterns in Global Perspectives

The evaluations of the tweets are done using sentiment mining approach with the Natural Language Toolkit in Python as well as associated plotting libraries.



Sentiment Metric	Score
Positive	0.608
Negative	0
Neutral	0.392
Compound	0.7351

FIGURE 4. Scoring of Tweets in Different Classes

This type of scripts can be used to identify the discussion of a particular keyword on social media. The police investigation teams and cyber forensic departments can analyze the tweets and messages which are related to a particular crime so that the deep analytics of the involved persons can be done. These tweets are extracted in JavaScript Object Notation (JSON) format which is further transformed to Comma Separated Value (CSV) or any other format so that deep evaluation can be done.

4. Conclusion

The continuous web information extraction is required for some, applications including criminal information investigation, statistical surveying, resident inputs, customer audits and numerous others. The statistical surveying master and specialists can extricate the information about specific subjects or items from different entryways for information disclosure for client notion conduct examination. Numerous corporate associations gather these kinds of datasets from online gateways with the goal that they can improve their administrations and item includes.

ACKNOWLEDGMENT

We are thankful to the research and development cell of our university for their support

References

[1] C. Fuchs: *From digital positivism and administrative big data analytics towards critical digital and social media research*, European Journal of Communication, **32** (2017), 37 – 49.

[2] S.J. Miah: *A big data analytics method for tourist behaviour analysis*, Information and Management, **54** (2017), 771 – 785.

[3] J. Li, H. Liu: *Challenges of feature selection for big data analytics*, IEEE Intelligent Systems **32** (2017), 9 – 15.

[4] I. Lee: *Big data Dimensions, evolution, impacts, and challenges*, Business Horizons, **60** (2017), 293 – 303.

[5] Z. Xiang, D. R. Fesenmaier: *Big data analytics, tourism design and smart tourism*, Analytics in smart tourism design Springer, (2017), 299 – 307.

[6] Z. Lv, H. Song, P. Basanta-Val, A. Steed, M. Jo: *Next-generation big data analytics: State of the art, challenges, and future research topics*, IEEE Transactions on Industrial Informatics, **13** (2017), 1891 – 1899.

[7] M.D. Lytras, V. Raghavan, E. Damiani: *Big data and data analytics research: From metaphors to value space for collective wisdom in human decision making and smart machines*, International Journal on Semantic Web and Information Systems, **13** (2017), 1 – 10.

[8] S. Stieglitz, M. Mirbabaie: *Social media analytics–Challenges in topic discovery, data collection, and data preparation*, International journal of information management, **39** (2018), 156 – 168.

[9] N. Koseleva, G. Ropaite: *Big data in building energy efficiency: understanding of big data and main challenges*, Procedia Engineering, **172** (2017), 544 – 549.

[10] W. He, F.K. Wang, V. Akula: *Managing extracted knowledge from big social media data for business decision making*, Journal of Knowledge Management, **21** (2017), 275 – 294.

[11] E. Ahmed, I. Yaqoob, I.A.T. Hashem: *The role of big data analytics in Internet of Things*, Business Horizons, **129** (2017), 459 – 471.

[12] L. Richardson: *Beautiful Soup Documentation*, Official Documentation, **4** (2007)

[13] T. D. Smedt, W. Daelemans: *Pattern for python*, Journal of Machine Learning Research, **13** (2012), 2063 – 2067.

[14] R. Lawson: *Web scraping with Python*, Packt Publishing Ltd, (2015)

[15] L. C. Dewi, A. Chandra: *Social Media Web Scraping using Social Media Developers API and Regex*, Procedia Computer Science, **157** (2019), 444 – 449.

[16] L. Molyneux, R. R. Mourão: *Political journalists normalization of Twitter: Interaction and new affordances*, Journalism Studies, **20** (2019), 248 – 266.

MMEC, MM(DU)
M. M. University, Mullana, Haryana
Email address: `neelamoberoi1030@mmumullana.org`

MMEC, MM(DU)
M. M. University, Mullana, Haryana
Email address: `sakshisachdeva26@gmail.com`

MMEC, MM(DU)
M. M. University, Mullana, Haryana
Email address: `prachigargji@gmail.com`

MMICT&BM,MM (DU)
M. M. University, Mullana, Haryana
Email address: `ahluwalia.rubika@gmail.com`