JOURNAL OF EMERGING COMPUTER TECHNOLOGIES (JECT)

VOLUME: 2 ISSUE NO: 1 YEAR: 2022 PERIOD: JANUARY-JUNE

e-ISSN: 2757-8267

INTERNATIONAL, OPEN ACCESS, PEER-REVIEWED JOURNAL

http://dergipark.org.tr/ject

İZMİR AKADEMİ DERNEĞİ

JOURNAL OF EMERGING COMPUTER TECHNOLOGIES (JECT)

E-ISSN: 2757-8267

Volume :2 Number:1 Year:2022

JOURNAL OF EMERGING COMPUTER TECHNOLOGIES (JECT)
VOLUME: 1 ISSUE NO: 1 YEAR: 2021 PERIOD: JANUARY-JUNE
e-ISSN: 2757-8267
INTERNATIONAL, OPEN ACCESS, PEER-REVIEWED JOURNAL

Indexing and Abstracting

Index Copernicus
https://journals.indexcopernicus.com/search/details?id=68795

Academia.edu
https://gkgkgkg.academia.edu/JournalofEmergingComputerTechnologies

Google Scholer
https://scholar.google.com/citations?user=n986qsIAAAAJ&hl=tr

Academic Resource Index (Researchbib)
http://journalseeker.researchbib.com/view/issn/2757-8267

ROAD
https://portal.issn.org/resource/ISSN/2757-8267

Asos Index
https://asosindex.com.tr/index.jsp?modul=journal-page&journal-id=2594

JOURNAL OF EMERGING COMPUTER TECHNOLOGIES

Volume: 2 **No:** 1 **Year:** 2022 **Period:** January-June

Head Office	: İzmir
Type of Publication	: It is published electronically twice a year, in June and December.
Owner	: İzmir Akademi Derneği (İzmir Academy Association)

Editor	: Dr. Muhammet DAMAR, Dokuz Eylul University, İzmir, Turkey.

Journal of Emerging Computer Technologies

Address	: Büyükkale Mh. Kerimoğlu Sok. No:40-2 Tire / Izmir, TURKEY
Web	: https://dergipark.org.tr/en/pub/ject
Email	: journalect@gmail.com
Email	: izmirakademidernegi@gmail.com

Journal of Emerging Computer Technologies publishes scientific/original research articles. It is published electronically twice a year, in June and December. It is an **international scientific refereed journal** that publishes articles written in **English**, and includes acedemicians from different countries in its boards. Uses "Double-Blind Peer Review" in reviewing processes. It has adopted the open access principle. No fee is requested from the authors for open access, processing, publication fee or otherwise. It is totally **FREE.**

The responsibility of the articles published in this journal in terms of science and language belongs to the authors. The articles published in the journal cannot be used without reference.

CONTENT

Journal of Emerging Computer Technologies
Research Article
Received: *2022-01-20* | ***Reviewing:*** *2022-01-20 & 2022-03-19* | ***Accepted:*** *2022-03-20* | ***Online:*** *2022-03-20* | ***Issue Date:*** *2022-06-30*

Providing Priority Degrees for the Challenges of Cloud-Based Outsource Software Development Projects via Fuzzy Analytic Hierarchy Process Methods

Ayşe Övgü KINAY
Dokuz Eylul University
İzmir, Türkiye
ovgu.tekin@deu.edu.tr
0000-0001-9908-8652

Can ATILGAN
Dokuz Eylul University
İzmir, Türkiye
can.atilgan@deu.edu.tr
0000-0002-1680-6207

***Abstract*— Cloud-based outsource software development (COSD) is a fairly new and popular software development methodology, which is enabled by the enormous growth of the cloud computing services in the last decade. The key idea of the methodology is to support software development processes of companies having software development team members from all around the world work collaboratively via cloud services. While there are quite some benefits a company could draw, there are also some challenges associated with the execution of a COSD project. It is intuitively essential to have a reliable way to assess a COSD project for its success. In this study, using Magnitude Based Fuzzy Analytic Hierarchy Process (MFAHP) as a method to prioritize and weight the challenges of a COSD project is presented. MFAHP is a fuzzy extension of the classical AHP which is shown to produce comparable results to other Fuzzy AHP (FAHP) methods with much smaller number of computations. The performance of the suggested methodology is evaluated and compared to Chang's Fuzzy Extent Analysis on AHP (FEA) and Geometric Mean (GM) Methods, which are two other established FAHP methods. The results show that MFAHP and GM perform quite similar, whereas FEA gives inconsistent outputs. Among 21 different subcategories of COSD project challenges determined, "compatibility issues" are anticipated to have the highest weight individually while "organization management" is the most important of 4 main categories.**

Keywords*— *criteria prioritization, magnitude based fuzzy analytic hierarchy process, challenges of cloud-based outsource software development projects.

I. INTRODUCTION

The computer technologies advance very rapidly and virtually all kinds of modern business processes require high-end IT capabilities to be successful in a competitive market. With newer generations of computer hardware are released every year and software even more frequently, maintaining a secure and up-to-date IT infrastructure is likely one of the most critical challenges an organization inevitably faces. Cloud computing offers an appropriate solution to this problem with little to no risk. Using cloud computing services, IT resources are purchased in a pay-as-you-use fashion. Also, the users are given the freedom to expand or shrink the size of their resource usage and change service configurations effortlessly.

A very rapid growth is observed in the cloud computing marketplace, as using a cloud service is such a convenient and low-risk way to satisfy computing needs of a business process. Cloud-based outsource software development (COSD) is a fairly new and popular methodology that allows companies to reap full benefits of cloud computing services while outsourcing development processes [1]–[3].COSD also enables a continuous and productive development process as it allows hiring skilled developers at a global scope, who can work together on the same project at a variety of time zones [4]. On the other hand, sharing cloud infrastructure while developing a software project has its exclusive challenges [5]. In summary, for the successful execution of COSD projects, various challenges including geographical, temporal, and intercultural differences should be well evaluated and processes should be controlled accordingly [6]–[8]. Prioritizing and weighting those challenges could be of critical importance for the success of a COSD project.

The analytic hierarchy process (AHP) is a well-studied method used to assign weights to a set of criteria, which then can be used to make a decision. Incorporating fuzzy sets with AHP helps to improve the method to better deal with the inaccuracy of individual judgments. In fact, a fuzzy AHP (FAHP) method was applied to the problem of determining COSD challenges in [5]. However, the method used in that study, which is Chang's Fuzzy Extent Analysis on AHP (FEA) [9], has some flaws that could be detrimental to the process of assessing a COSD project. In this study, Magnitude Based Fuzzy Analytic Hierarchy Process (MFAHP) is used to present a refined and reliable method to prioritize and weight the challenges of a COSD project, while dealing with the shortcomings of the previous studies on COSD. MFAHP is a fuzzy extension of the classical AHP which is shown to produce comparable results to other FAHP methods with much smaller number of computations [10]. Application of the presented methodology with MFAHP as well as two other FAHP, i.e. FEA and the Geometric Mean method (GM), using the data laid out in [5] are realized.

The following section provides the definitions of fundamental concepts of FAHP and the algorithms used in the methods. The details of the applications of the methods and a discussion on their results are given in the third section. Finally, the fourth section concludes the paper.

II. FUNDAMENTAL CONCEPTS AND ALGORITHMS

The fuzzy set theory, which allows better expression of the uncertainties in the data, was proposed by Zadeh in 1965 [11].

Cite (APA): Kınay, Ö. A., Atılgan, C. (2022). Providing Priority Degrees for the Challenges of Cloud-Based Outsource Software Development Projects via Fuzzy Analytic Hierarchy Process Methods. *Journal of Emerging Computer Technologies*, 2(1), 1-6.

It is the expression of the data with the degree of belonging in the range [0,1] instead of belonging to a certain cluster or not.

The triangular membership function, which is one of the membership functions frequently used in fuzzy set theory, is as follows.

Definition 1. $A = (l, m, u)$ on $U = (-\infty, \infty)$ is expressed as a triangular fuzzy number, and its membership function $\mu_A: U \rightarrow [0,1]$ is given as:

$$\mu_A(x) = \begin{cases} \frac{(x-l)}{(m-l)}, & l < x < m \\ 1 \quad, & x = m \\ \frac{(u-x)}{(u-m)}, & m < x < u \\ 0 \quad, & otherwise \end{cases} \qquad (1)$$

In the Fuzzy Analytic Hierarchy Process (FAHP), fuzzy pairwise comparison matrices are created, as are the matrices with pairwise comparisons of criteria and/or alternatives in the Analytic Hierarchy Process (AHP) proposed by Saaty [9]. Then, method-specific FAHP calculations are made using these matrices and hierarchical structure. Naturally, in FAHP methods, comparisons in these matrices are expressed as fuzzy numbers (usually triangular) as in (2), and a typical workflow for FAHP methods is also illustrated in Fig 1.

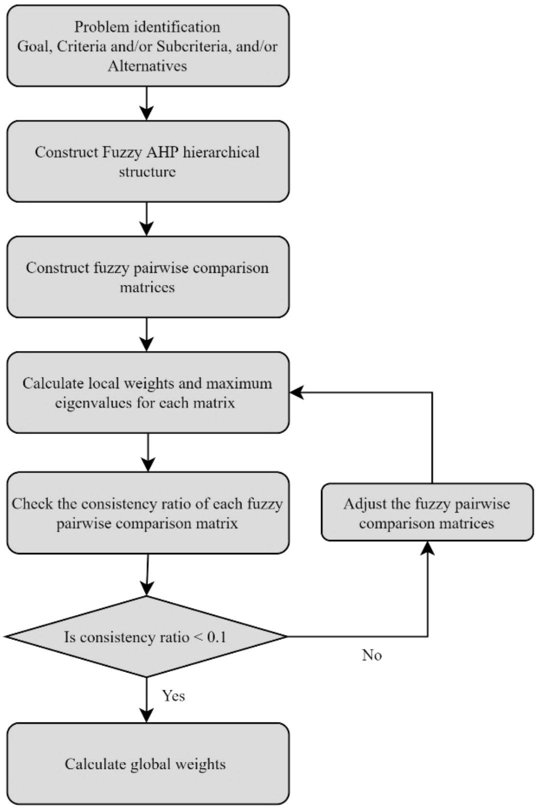

Problem identification
Goal, Criteria and/or Subcriteria, and/or Alternatives
Construct Fuzzy AHP hierarchical structure
Construct fuzzy pairwise comparison matrices
Calculate local weights and maximum eigenvalues for each matrix
Check the consistency ratio of each fuzzy pairwise comparison matrix
Adjust the fuzzy pairwise comparison matrices
Is consistency ratio < 0.1
No
Yes
Calculate global weights

Fig. 1. A typical FAHP workflow.

$$\boldsymbol{A} = (a_{ij})_{n\times n} = \begin{bmatrix} (1,1,1) & (l_{12}, m_{12}, u_{12}) & \dots & (l_{1n}, m_{1n}, u_{1n}) \\ (l_{21}, m_{21}, u_{21}) & (1,1,1) & \dots & (l_{2n}, m_{2n}, u_{2n}) \\ \vdots & \vdots & \vdots & \vdots \\ (l_{n1}, m_{n1}, u_{n1}) & (l_{n2}, m_{n2}, u_{n2}) & \dots & (1,1,1) \end{bmatrix} \qquad (2)$$

In fuzzy pairwise comparison matrices, as in classical AHP, if $a_{ij} = (l_{ij}, m_{ij}, u_{ij})$ then $a_{ji} = a_{ij}^{-1} = (1/u_{ij}, 1/m_{ij}, 1/l_{ij})$, for $i, j = 1, \dots, n$, $i \neq j$.

In order to determine the weights of the COSD challenges, the first of the three FAHP methods used in this study is the algorithm of the GM method, which is known to obtain consistent results. Secondly, the algorithm of the FEA method, which is frequently used in studies but unfortunately causes wrong results, is mentioned. Finally, the algorithm of the MFAHP method, which can produce results close to the GM method in a shorter time, is briefly explained.

A. *Geometric mean method (GM)*

The calculation procedure of the method proposed by Buckley [13] in 1985 is as follows. (For each step, $i = 1, \dots, n$)

Step 1. Calculate the geometric mean of each criterion or alternative from each fuzzy pairwise comparison matrix expressed as in (2).

$$z_i = \left(\prod_{j=1}^{n} a_{ij}\right)^{1/n} \qquad (3)$$

Step 2. Obtain fuzzy weight values r_i of each criterion or each alternative.

$$r_i = z_i \otimes [z_1 \oplus z_2 \oplus \dots \oplus z_n]^{-1} \qquad (4)$$

Step 3. Defuzzify the r_i values by using Center of Area (COA) method.

$$S_i = \frac{l_i + m_i + u_i}{3} \qquad (5)$$

Step 4. Normalize the defuzzified S_i weight values.

$$w_i = \frac{S_i}{\sum_{i=1}^{n} S_i} \qquad (6)$$

B. *Chang's extent analysis on FAHP (FEA)*

The calculation procedure of the method proposed by Chang [9] in 1996 is as follows. (For each step, $i = 1, \dots, n$)

Step 1. Obtain the row sums for each fuzzy pairwise comparison matrix.

$$RS_i = \sum_{j=1}^{n} a_{ij} = \left(\sum_{j=1}^{n} l_{ij}, \sum_{j=1}^{n} m_{ij}, \sum_{j=1}^{n} u_{ij}\right) \qquad (7)$$

Step 2. Calculate the S_i values.

$$S_i = \frac{RS_i}{\sum_{j=1}^{n} RS_j} = \left(\frac{\sum_{j=1}^{n} l_{ij}}{\sum_{k=1}^{n}\sum_{j=1}^{n} u_{kj}}, \frac{\sum_{j=1}^{n} m_{ij}}{\sum_{k=1}^{n}\sum_{j=1}^{n} m_{kj}}, \frac{\sum_{j=1}^{n} u_{ij}}{\sum_{k=1}^{n}\sum_{j=1}^{n} l_{kj}}\right) \qquad (8)$$

Step 3. Calculate the degree of possibility of $S_i \geq S_j$ values.

$$V(S_i \geq S_j) = \begin{cases} 1 & , m_i \geq m_j \\ \frac{(u_i - l_j)}{(u_i - m_i) + (m_j - l_j)}, & l_j \leq u_i, i, j = 1, \dots, n;\ j \neq i \\ 0 & , otherwise \end{cases} \qquad (9)$$

where $S_i = (l_i, m_i, u_i)$ and $S_j = (l_j, m_j, u_j)$ and the visual representation of $V(S_i \geq S_j)$ is shown in Fig. 2.

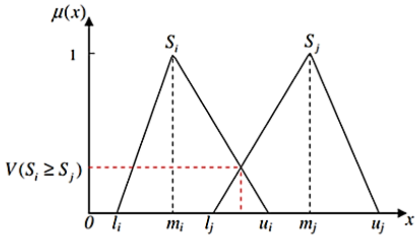


Fig. 2. Visual representation of $V(S_i \geq S_j)$.

Step 4. Calculate the degree of possibility of each S_i over all other $(n-1)$ fuzzy numbers.

$$V(S_i \geq S_j \mid j = 1, \dots, n; j \neq i) = \min_{j \in \{1,\dots,n\}, j \neq i} V(S_i \geq S_j) \quad (10)$$

Step 5. Normalize the calculated values in Step 4.

$$w_i = \frac{V(S_i \geq S_j \mid j=1,\dots,n; j \neq i)}{\sum_{k=1}^{n} V(S_k \geq S_j \mid j=1,\dots,n; j \neq k)} \quad (11)$$

where the weight values are crisp values.

C. Magnitude based fuzzy analytic hierarchy process (MFAHP)

Ranking of alternatives is the knowledge sought to be achieved in FAHP methods. Therefore, the MFAHP method [10] has emerged with the thought that integrating this ranking step into the method for the correct calculation of local and global weights will provide a significant improvement in the final decision. For this reason, many of the proposed [14]–[18] fuzzy number ranking approaches have been examined. Among these methods, the magnitude information of the numbers, where effective results can be obtained due to the nature of fuzzy comparisons used in pairwise comparison matrices, has been added to the steps of the FAHP method. The calculation procedure of the method is as follows. (For each step, $i = 1, \dots, n$)

Step 1. Obtain the row sums for each fuzzy pairwise comparison matrix by using (7).

Step 2. Apply the normalization process as stated in [19].

$$S_i = \frac{RS_i}{\sum_{j=1}^{n} RS_j} = \left(\frac{\sum_{j=1}^{n} l_{ij}}{\sum_{j=1}^{n} l_{ij} + \sum_{k=1, k \neq i}^{n} \sum_{j=1}^{n} u_{kj}}, \frac{\sum_{j=1}^{n} m_{ij}}{\sum_{k=1}^{n} \sum_{j=1}^{n} m_{kj}}, \frac{\sum_{j=1}^{n} u_{ij}}{\sum_{j=1}^{n} u_{ij} + \sum_{k=1, k \neq i}^{n} \sum_{j=1}^{n} l_{kj}} \right) \quad (12)$$

Step 3. Calculate the magnitude value of each S_i value.

$$Mag(S_i) = \frac{l_i + 10 m_i + u_i}{12} \quad (13)$$

Step 4. Normalize the magnitude value of each S_i value.

$$w_i = \frac{Mag(S_i)}{\sum_{j=1}^{n} Mag(S_j)} \quad (14)$$

where the weight values are crisp values.

III. Methodology and Application

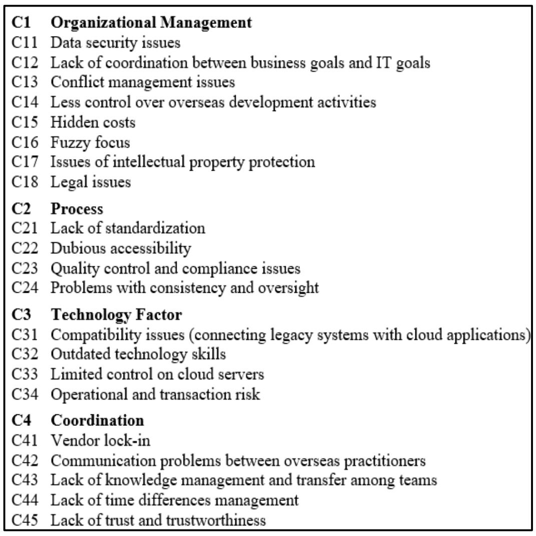
In this section, the decision making problem discussed in [5],that is determining the importance weights (priority degress) of challenges in COSD projects, is solved using 3 different FAHP methods. There are 4 main categories, and under those, a total of 21 subcategories of challenges in the decision making problem.

The authors of [5] conducted an extensive literature review to eventually designate 78 primary studies. They extracted the data from those selected studies and created a list of COSD challenges categories. The list were then validated via a pilot study assessment using Kendall's non-parametric coefficient of concordance test[20] performed involving 5 experts. Following that, a questionnaire survey was conducted with 119 participants who worked in the international software development environments. The participants of the survey were asked to grade challenges regarding their importance on a five-scale Likert scale. It is worth noting that the participants were also provided with an open-ended section in the questionnaire to elicit more challenge categories, but no additional challenges were reported. The determined challenge categories are given in Table I, and their hierarchical structure is displayed in Fig. 3.

The pairwise comparisons of the challenges were obtained via a secondary questionnaire survey with a sub-group of experts who also participated in the first survey. The sample of the questionnaire, the bibliographic information of participants, a sample of FAHP questionnaire, and fuzzy pairwise comparison matrices of the mentioned surveys are shared in the appendices of [5].

In [5], the FEA method was used to obtain the weights of the set of COSD challenges from the fuzzy pairwise comparison matrices. However, FEA has some flaws, which were investigated and laid out in [21] and confirmed by many following studies [19], [22]–[25].The results of the applications given in this paper also support that FEA is a faulty method for the particular case of weighting COSD challenges. Among a variety of FAHP methods, the GM and MFAHP are shown to be the ones to yield the highest quality results [10]. Between these two methods, MFAHP has an additional advantage when it comes to a real-world application like weighting a COSD project's challenges. MFAHP is quite simple to actually use as it abstracts the fuzzy computations from the user by providing a single concise formula, i.e. (13), to directly calculate crisp weight values. It is also shown that a less computation intensive than any other FAHP method [10]. Thus, we suggest and adopt MFAHP as the preferred method for criteria weighting in COSD projects.

TABLE I. List of main COSD challenges and their subcategories



C1	**Organizational Management**
C11	Data security issues
C12	Lack of coordination between business goals and IT goals
C13	Conflict management issues
C14	Less control over overseas development activities
C15	Hidden costs
C16	Fuzzy focus
C17	Issues of intellectual property protection
C18	Legal issues
C2	**Process**
C21	Lack of standardization
C22	Dubious accessibility
C23	Quality control and compliance issues
C24	Problems with consistency and oversight
C3	**Technology Factor**
C31	Compatibility issues (connecting legacy systems with cloud applications)
C32	Outdated technology skills
C33	Limited control on cloud servers
C34	Operational and transaction risk
C4	**Coordination**
C41	Vendor lock-in
C42	Communication problems between overseas practitioners
C43	Lack of knowledge management and transfer among teams
C44	Lack of time differences management
C45	Lack of trust and trustworthiness

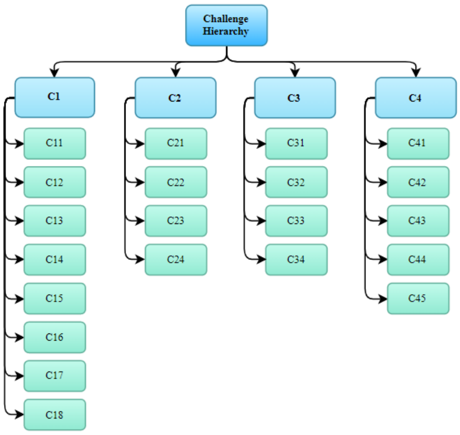


Fig. 3. Hierarchical structure of the problem.

All three FAHP methods presented in the previous section are applied to the problem using fuzzy pairwise comparison matrices of COSD challenges. The triangular linguistic terms used in fuzzy pairwise comparison matrices are given in Table II. The entire fuzzy pairwise comparison matrices involving main categories and subcategories –as expressed in (2)– were originally shared in [5]. The matrices relevant to this study, which were checked for consistency (but not shared in the text for the sake of conciseness) are given in Tables III-VII. However, it is worth noting that our matrices are slightly modified because the source material involved a few small errors.

TABLE II. TRIANGULAR LINGUISTIC TERMS

Linguistic term	Triangular fuzzy scale	Triangular fuzzy reciprocal scale
Just equal	(1, 1, 1)	(1, 1, 1)
Equally important	(1/2, 1, 3/2)	(2/3, 1, 2)
Weakly important	(1, 3/2, 2)	(1/2, 2/3, 1)
Strongly more important	(3/2, 2, 5/2)	(2/5, 1/2, 2/3)
Very strongly more important	(2, 5/2, 3)	(1/3, 2/5, 1/2)
Absolutely more important	(5/2, 3, 7/2)	(2/7, 1/3, 2/5)

TABLE III. FUZZY PAIRWISE COMPARISON MATRIX OF MAIN COSD CHALLENGE CATEGORIES

	C1	C2	C3	C4
C1	(1, 1, 1)	(2, 5/2, 3)	(1, 3/2, 2)	(3/2, 2, 5/2)
C2	(1/3, 2/5, 1/2)	(1, 1, 1)	(2/5, 1/2, 2/3)	(1/2, 2/3, 1)
C3	(1/2, 2/3, 1)	(3/2, 2, 5/2)	(1, 1, 1)	(1, 3/2, 2)
C4	(2/5, 1/2, 2/3)	(1, 3/2, 2)	(1/2, 2/3, 1)	(1, 1, 1)

TABLE IV. FUZZY PAIRWISE COMPARISON MATRIX OF SUBCATEGORIES FOR C1

	C11	C12	C13	C14
C11	(1, 1, 1)	(1, 3/2, 2)	(5/2, 3, 7/2)	(2/3, 1, 2)
C12	(1/2, 2/3, 1)	(1, 1, 1)	(1/2, 2/3, 1)	(1, 3/2, 2)
C13	(2/7, 1/3, 2/5)	(1, 3/2, 2)	(1, 1, 1)	(1/2, 1, 3/2)
C14	(1/2, 1, 3/2)	(1/2, 2/3, 1)	(2/3, 1, 2)	(1, 1, 1)
C15	(2/5, 1/2, 2/3)	(3/2, 2, 5/2)	(1, 3/2, 2)	(5/2, 3, 7/2)
C16	(1/2, 2/3, 1)	(1/2, 2/3, 1)	(1, 3/2, 2)	(1/3, 2/5, 1/2)
C17	(1, 3/2, 2)	(1/3, 2/5, 1/2)	(1/2, 2/3, 1)	(2/3, 1, 2)
C18	(2, 5/2, 3)	(1/2, 2/3, 1)	(1, 3/2, 2)	(1/3, 2/5, 1/2)

	C15	C16	C17	C18
C11	(3/2, 2, 5/2)	(1, 3/2, 2)	(1/2, 2/3, 1)	(1/3, 2/5, 1/2)
C12	(2/5, 1/2, 2/3)	(1, 3/2, 2)	(2, 5/2, 3)	(1, 3/2, 2)
C13	(1/2, 2/3, 1)	(1/2, 2/3, 1)	(1, 3/2, 2)	(1/2, 2/3, 1)
C14	(2/7, 1/3, 2/5)	(2, 5/2, 3)	(1/2, 1, 3/2)	(2, 5/2, 3)
C15	(1, 1, 1)	(2/5, 1/2, 2/3)	(2/7, 1/3, 2/5)	(1, 3/2, 2)
C16	(3/2, 2, 5/2)	(1, 1, 1)	(2/5, 1/2, 2/3)	(2, 5/2, 3)
C17	(5/2, 3, 7/2)	(3/2, 2, 5/2)	(1, 1, 1)	(2/5, 1/2, 2/3)
C18	(1/2, 2/3, 1)	(1/3, 2/5, 1/2)	(3/2, 2, 5/2)	(1, 1, 1)

TABLE V. FUZZY PAIRWISE COMPARISON MATRIX OF SUBCATEGORIES FOR C2

	C21	C22	C23	C24
C21	(1, 1, 1)	(1/3, 2/5, 1/2)	(1, 3/2, 2)	(1/2, 2/3, 1)
C22	(2, 5/2, 3)	(1, 1, 1)	(1/2, 2/3, 1)	(3/2, 2, 5/2)
C23	(2/5, 1/2, 2/3)	(1, 3/2, 2)	(1, 1, 1)	(1, 3/2, 2)
C24	(1, 3/2, 2)	(2/5, 1/2, 2/3)	(1/2, 2/3, 1)	(1, 1, 1)

TABLE VI. FUZZY PAIRWISE COMPARISON MATRIX OF SUBCATEGORIES FOR C3

	C31	C32	C33	C34
C31	(1, 1, 1)	(3/2, 2, 5/2)	(3/2, 2, 5/2)	(3/2, 2, 5/2)
C32	(2/5, 1/2, 2/3)	(1, 1, 1)	(1/2, 2/3, 1)	(1/2, 2/3, 1)
C33	(2/5, 1/2, 2/3)	(1, 3/2, 2)	(1, 1, 1)	(1, 3/2, 2)
C34	(2/5, 1/2, 2/3)	(1, 3/2, 2)	(1/2, 2/3, 1)	(1, 1, 1)

TABLE VII. FUZZY PAIRWISE COMPARISON MATRIX OF SUBCATEGORIES FOR C4

	C41	C42	C43	C44	C45
C41	(1, 1, 1)	(1/3, 2/5, 1/2)	(3/2, 2, 5/2)	(2/5, 1/2, 2/3)	(2/5, 1/2, 2/3)
C42	(2, 5/2, 3)	(1, 1, 1)	(2, 5/2, 3)	(1/2, 1, 3/2)	(1, 3/2, 2)
C43	(2/5, 1/2, 2/3)	(1/3, 2/5, 1/2)	(1, 1, 1)	(2, 5/2, 3)	(5/2, 3, 7/2)
C44	(3/2, 2, 5/2)	(2/3, 1, 2)	(1/3, 2/5, 1/2)	(1, 1, 1)	(1/2, 2/3, 1)
C45	(3/2, 2, 5/2)	(1/2, 2/3, 1)	(2/7, 1/3, 2/5)	(1, 3/2, 2)	(1, 1, 1)

The results of the applications are given in Table VIII-X. Table VIII contains local weights for the main categories of COSD challenges. It is seen that the weight values obtained by MFAHP are very similar to the results of the GM while FEA results are notably different, as expected. The fact that some values seen in the results of FEA method are actually zero (like the result of the C2 main COSD challenge) implies that the criterion is totally irrelevant, which is false. Also note that the numerical results we obtained using FEA are slightly different than the ones given in [5], due to the small numerical errors as previously mentioned.

TABLE VIII. LOCAL WEIGHTS OF MAIN COSD CHALLENGES

	MFAHP	GM	FEA	Rank
C1	0.3783	0.3779	*0.5070*	1
C2	0.1410	0.1423	***0.0000***	4
C3	0.2805	0.2800	*0.3403*	2
C4	0.2002	0.1998	*0.1527*	3

TABLE IX. LOCAL WEIGHTS OF SUBCATEGORIES

	MFAHP	GM	FEA
C11	0.1438	0.1472	0.1446
C12	0.1275	0.1345	0.1287
C13	0.0956	0.1047	0.0910
C14	0.1301	0.1324	0.1317
C15	0.1336	0.1224	0.1348
C16	0.1198	0.1171	0.1193
C17	0.1311	0.1283	0.1322
C18	0.1185	0.1134	0.1176
C21	0.2000	0.1989	0.1588
C22	0.3435	0.3332	0.4029
C23	0.2508	0.2557	0.2641
C24	0.2056	0.2122	0.1742
C31	0.3865	0.3870	0.5275
C32	0.1594	0.1677	0.0297
C33	0.2495	0.2433	0.2779
C34	0.2046	0.2020	0.1649
C41	0.1428	0.1387	0.0834
C42	0.2737	0.3006	0.3186
C43	0.2391	0.2037	0.2665
C44	0.1661	0.1790	0.1618
C45	0.1782	0.1780	0.1696

Table IX contains local weights for the subcategories. When the local weights of the main COSD challenges (Table VIII) and the local weights of the subcategories of COSD challenges (Table IX) were evaluated together, global weight values were obtained for all subcategories for COSD challenges as in Table X. The comparative results in Table X are also visualized in Fig. 4. It is seen that the results of MFAHP and GM are similar. However, FEA results often produce larger or smaller weight values than MFAHP and GM results. Also, FEA produced zero weights for some criteria (C21-C24), which is an anomaly anyway.

TABLE X. GLOBAL WEIGHTS OF COSD CHALLENGES

		MFAHP	GM	FEA
C1	C11	0.0544	0.0556	0.0733
	C12	0.0482	0.0508	0.0652
	C13	0.0362	0.0396	0.0462
	C14	0.0492	0.0500	0.0668
	C15	0.0505	0.0462	0.0683
	C16	0.0453	0.0442	0.0605
	C17	0.0496	0.0485	0.0670
	C18	0.0448	0.0428	0.0596
C2	C21	0.0282	0.0283	0.0000
	C22	0.0484	0.0474	0.0000
	C23	0.0354	0.0364	0.0000
	C24	0.0290	0.0302	0.0000
C3	C31	0.1084	0.1084	0.1795
	C32	0.0447	0.0470	0.0101
	C33	0.0700	0.0681	0.0946
	C34	0.0574	0.0566	0.0561
C4	C41	0.0286	0.0277	0.0127
	C42	0.0548	0.0601	0.0487
	C43	0.0479	0.0407	0.0407
	C44	0.0332	0.0358	0.0247
	C45	0.0357	0.0356	0.0259

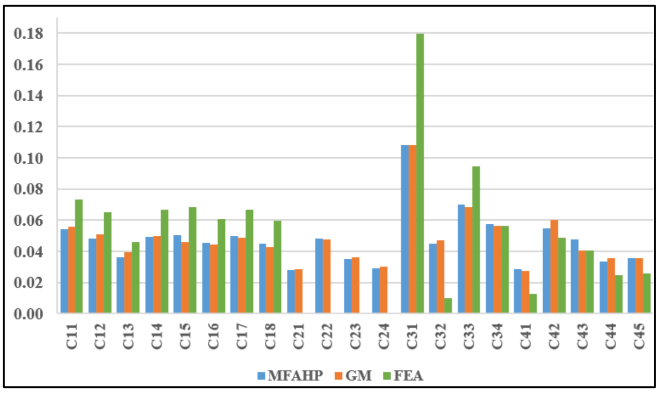


Fig. 4. Graphical representation of global weights of COSD challenges.

Regardless of the FAHP method used, the main challenge categories are prioritized in the same order. Organizational management (C1) challenges are of the highest priority, followed by technology factor (C3) and coordination (C4) challenges. The development process (C2) category is the lowest ranked of main challenge priorities. Though the priority rankings are essentially indifferent, FEA generated considerably different and inconsistent weights compared to other two methods. The weights spread considerably wider in the case of FEA, such that C3 is more than twice as important than C4 and total weights of C3 and C4 are smaller than C1 alone. C2 is assigned a zero-weight by FEA, which would translate to that challenges regarding the development process are entirely effectless. Of course, this points to a flaw in the method, rather than the interpretetion. MFAHP and GM assigned very similar weight values with one another. However, though the differences between weight values are considerably small, the weight ranking is of subcategories are different even for MFAHP and GM. For example, the third highest weight value is assigned to operational and transactional risk (C34) using MFAHP, while GM gives the third plac to communication problems between overseas practitioners (C42). Considering the results of MFAHP and GM, compatibility issues (C31) are expected to pose the greatest challange in a COSD project among 21 subcategories, by quite a margin. Fig. 5 summarizes the collective outcome of the MFAHP application.

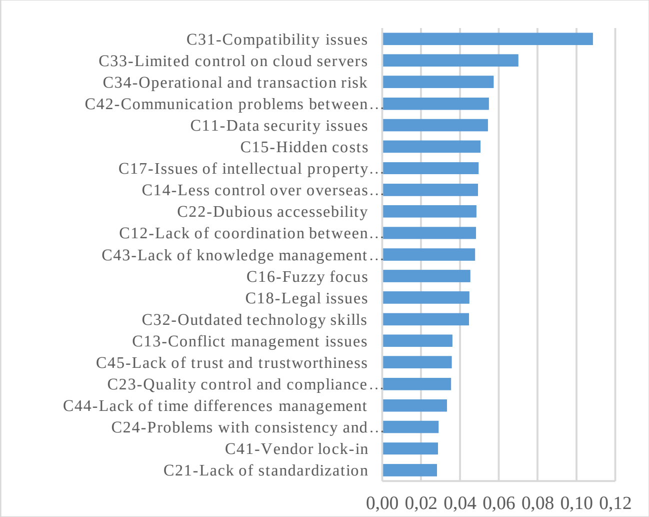


Fig. 5. Sorted global weights of COSD challenges obtained by the MFAHP.

TABLE XI. RUNNING TIMES OF EACH METHOD

	Running times (in ms.)
MFAHP	0.0287
GM	0.1923
FEA	0.1677

Finally, when the methods are examined in terms of the running times given in Table XI, it is known that GM is slow due to the calculation procedure and FEA is the fastest. For this example, MFAHP performed faster than FEA. Of course, it is not correct to generalize for a single example. However, in [10] where the methods are compared in detail, it has already been shown that MFAHP works as fast as FEA. All of the mentioned applications were programmed in C# language and run on a personal computer with a 10th Generation Intel Core i7 CPU clocked at 1.8 GHz and 16 GB of RAM.

IV. CONCLUSION

Managing a COSD process has a variety of challenges, some of them being unique to the process. This study provides a methodology tailored for a COSD process assessment in the form of weighted criteria. The methodology combines the challenge categorization presented in [5] and the MFAHP method proposed in[10].The results of the example application indicate that choice of the FAHP method is superior in one way or another to its rivals, especially to FEA used in [5] in terms of outcome accuracy. Therefore, we believe that the study could benefit to decision makers of a COSD process in a rather simple way.

In the future, evaluating a set of real COSD projects based on the results achieved in this study is of primary importance. Also, the problem of selecting cloud service providers could be integrated into the methodology presented in this study to further assist the decision makers towards the success of their projects.

REFERENCES

[1] S. Schneider and A. Sunyaev, “Determinant factors of cloud-sourcing decisions: Reflecting on the IT outsourcing literature in the era of cloud computing,” Journal of Information Technology, vol. 31, no. 1, 2016, doi: 10.1057/jit.2014.25.

[2] S. Dhar, “From outsourcing to Cloud computing: Evolution of IT services,” Management Research Review, vol. 35, no. 8, 2012, doi: 10.1108/01409171211247677.

[3] S. Leimeister, M. Böhm, C. Riedl, and H. Krcmar, “The business perspective of cloud computing: Actors, roles, and value networks,” 2010.

[4] M. Böhm, S. Leimeister, C. Riedl, and H. Krcmar, “Cloud Computing – Outsourcing 2.0 or a new Business Model for IT Provisioning?”, in Application Management, 2011. doi: 10.1007/978-3-8349-6492-2_2.

[5] M. A. Akbar, M. Shameem, S. Mahmood, A. Alsanad, and A. Gumaei, “Prioritization based Taxonomy of Cloud-based Outsource Software Development Challenges: Fuzzy AHP analysis,” Applied Soft Computing Journal, vol. 95, no. 106557, 2020, doi: 10.1016/j.asoc.2020.106557.

[6] S. U. Khan, M. Niazi, and R. Ahmad, “Factors influencing clients in the selection of offshore software outsourcing vendors: An exploratory study using a systematic literature review,” in Journal of Systems and Software, 2011, vol. 84, no. 4. doi: 10.1016/j.jss.2010.12.010.

[7] A. B. Steven, Y. Dong, and T. Corsi, “Global sourcing and quality recalls: An empirical study of outsourcing-supplier concentration-product recalls linkages,” Journal of Operations Management, vol. 32, no. 5, 2014, doi: 10.1016/j.jom.2014.04.003.

[8] R. Jabangwe, D. Šmite, and E. Hessbo, “Distributed software development in an offshore outsourcing project: A case study of source code evolution and quality,” Information and Software Technology, vol. 72, 2016, doi: 10.1016/j.infsof.2015.12.005.

[9] D. Y. Chang, “Applications of the extent analysis method on fuzzy AHP,” European Journal of Operational Research, vol. 95, no. 3, pp. 649–655, 1996, doi: 10.1016/0377-2217(95)00300-2.

[10] A. O. Kinay and B. T. Tezel, “Modification of the fuzzy analytic hierarchy process via different ranking methods,” International Journal of Intelligent Systems, pp. 1–29, 2021, doi: 10.1002/int.22628.

[11] L. Zadeh, “Fuzzy Sets,” Information and Control, vol. 8, no. 3, pp. 338–353, 1965.

[12] T. L. Saaty, The analytic hierarchy process: planning, priority setting, resource allocation, 1st edition. New York: McGraw-Hill International Book Co., 1980.

[13] J. J. Buckley, “Fuzzy hierarchical analysis,” Fuzzy Sets and Systems, vol. 17, no. 3, pp. 233–247, 1985, doi: 10.1016/0165-0114(85)90090-9.

[14] S. Abbasbandy and T. Hajjari, “A new approach for ranking of trapezoidal fuzzy numbers,” Computers and Mathematics with Applications, vol. 57, no. 3, pp. 413–419, 2009, doi: 10.1016/j.camwa.2008.10.090.

[15] X. Wang and E. E. Kerre, “Reasonable properties for the ordering of fuzzy quantities (I),” Fuzzy Sets and Systems, vol. 118, no. 3, pp. 375–385, 2001, doi: 10.1016/S0165-0114(99)00062-7.

[16] X. Wang and E. E. Kerre, “Reasonable properties for the ordering of fuzzy quantities (II),” Fuzzy Sets and Systems, vol. 118, no. 3, pp. 387–405, 2001, doi: 10.1016/S0165-0114(99)00063-9.

[17] G. Bortolan and R. Degani, “A review of some methods for ranking fuzzy subsets,” Fuzzy Sets and Systems, vol. 15, no. 1, pp. 1–19, 1985, doi: 10.1016/0165-0114(85)90012-0.

[18] N. van Hop, “Ranking fuzzy numbers based on relative positions and shape characteristics,” Expert Systems with Applications, vol. 191, 2022, doi: 10.1016/j.eswa.2021.116312.

[19] Y. M. Wang and T. M. S. Elhag, “On the normalization of interval and fuzzy weights,” Fuzzy Sets and Systems, vol. 157, no. 18, pp. 2456–2471, 2006, doi: 10.1016/j.fss.2006.06.008.

[20] M. G. Kendall, Rank Correlation Methods. 1948.

[21] Y. M. Wang, Y. Luo, and Z. Hua, “On the extent analysis method for fuzzy AHP and its applications,” European Journal of Operational Research, vol. 186, no. 2, pp. 735–747, 2008, doi: 10.1016/j.ejor.2007.01.050.

[22] K. Zhü, “Fuzzy analytic hierarchy process: Fallacy of the popular methods,” European Journal of Operational Research, vol. 236, no. 1, 2014, doi: 10.1016/j.ejor.2013.10.034.

[23] S. Kubler, J. Robert, W. Derigent, A. Voisin, and Y. le Traon, “A state-of the-art survey & testbed of fuzzy AHP (FAHP) applications,” Expert Systems with Applications, vol. 65, pp. 398–422, 2016, doi: 10.1016/j.eswa.2016.08.064.

[24] F. R. Lima-Junior and L. C. R. Carpinetti, “Dealing with the problem of null weights and scores in Fuzzy Analytic Hierarchy Process,” Soft Computing, vol. 24, no. 13, pp. 9557–9573, 2020, doi: 10.1007/s00500-019-04464-8.

[25] F. Ahmed and K. Kilic, “Fuzzy Analytic Hierarchy Process: A performance analysis of various algorithms,” Fuzzy Sets and Systems, vol. 362, pp. 110–128, 2019, doi: 10.1016/j.fss.2018.08.009.

Journal of Emerging Computer Technologies
Research Article
Received: *2022-04-08* | ***Reviewing:*** *2022-04-08 & 2022-05-20*| ***Accepted:*** *2022-05-21* | ***Online:*** *2022-05-22* | ***Issue Date:*** *2022-06-30*

Ensemble Regression-Based Gold Price (XAU/USD) Prediction

Zeynep Hilal Kilimci
Department of Information Systems Engineering
Kocaeli University
Kocaeli, 41001, Turkey
zeynep.kilimci@kocaeli.edu.tr
0000-0003-1497-305X

***Abstract*—The prediction of any commodities such as cryptocurrency, stocks, silver, gold is a challenging task for the investors, researchers, and analysts. In this work, we propose a model that forecasts the value of 1 ounce of gold in dollars by using regression ensemble-based approaches. To our knowledge, this is the very first study in terms of combining regression models for the prediction of XAU/USD index although there are plenty of methods employed in the literature to forecast the price of gold. The contributions of this study are fivefold. First, the dataset is gathered between July 2019 and July 2020 from global financial websites in the world, and cleaned for modeling. Then, feature space is extended with technical and statistical indicators in addition to opening, closing, highest, lowest prices of gold index. Next, different regression and ensemble-based regression models are carried out. These are linear regression, polynomial regression, decision tree regression, random forest regression, support vector regression, voting regressor, stacking regressor. Experiment results demonstrate that the usage of stacking regression combination model exhibits considerable results with 2.2036 of MAPE for forecasting the price of XAU/USD index.**

Keywords—Gold price prediction, XAU/USD index forecast, ensemble regression, stacking regressor

I. INTRODUCTION

Gold, a precious metal, has maintained its popularity among societies for thousands of dec as a barter, reserve unit, and jewelry. Considering the durability of the gold mine due to its structure, the convenience it ensures in terms of its workability and other benefits, it is significant for the business in production and for the financial markets as a commodity. For this reason, the price of gold is widely followed in the world financial markets. If the gold index is the price of an ounce of gold traded in US dollars (XAU/USD), that is, it refers to how many US dollars it takes to buy an ounce of gold. Gold, as a commodity, is considered one of the most important investment instruments not only for companies that are in close contact with the outside world, but also for any country. Countries and multinational companies use the exchange rate, which is one of the most important economic variables, as well as gold reserves as variables to ensure their connection with the outside world. This makes the gold index and the gold market one of the largest and most important financial markets in the world. For this reason, the gold index can be quickly affected in a positive or negative way by many developments that may occur in the markets, the economy and political policies. Taking into account external factors, it becomes almost impossible to control the future level of the gold index and its market. This makes gold index forecasting a more attractive and active research area for researchers, and investors. Within the scope of this study, it is proposed to construct a model that predicts the price of the gold index based on the regression ensemble-based approach.

Regression analysis is known as a collection of statistical procedures for predicting the relations between a dependent argument and one or more independent arguments. Regression analysis is especially employed for two conceptively different objectives. Firstly, regression analysis is evaluated to conclude causal relations between the independent and dependent arguments. Second, regression analysis is commonly utilized for forecasting, where its use has drastically coincided with the area of machine learning. The second usage of regression analysis is the main focus of this work. The most used kinds of regression analysis are linear regression, logistic regression ridge regression, etc. In this study, linear regressor, decision tree regressor, random forest regressor, and support vector regressor are evaluated as base regressors on the other hand voting regressor, and stacking regressor are assessed as ensemble regression models to predict the gold price.

In this study, it is proposed to forecast the price of gold index (XAU/USD). Movements in XAU/USD are analyzed between July 2019 and July 2020 by gathering data from global financial websites. In order to compose feature set is, opening, closing, the highest, and the lowest gold prices are included to the dataset. The same variables of the dollar index that have an effect on gold have been added to the dataset. In order to extend feature space, technical indicators are also included namely, simple moving average, relative strength index, and Bollinger band. Then, five different regression models are constructed namely, linear regression, polynomial regression, decision tree regression, random forest regression, support vector regression. Finally, voting regressor and stacking regressor models are employed by consolidating previous four regression models to get more robust prediction of gold price. There are plenty of methods employed by academic circumferences to evaluate and forecast the price of gold, such models are based on linear regression (MLR), support vector machine (SVM), artificial neural network (ANN), etc. To our knowledge, this is the very first attempt in terms of combining regression models for the prediction of XAU/USD index. Experiment results demonstrate that the combining of regression models is an effective method to acquire more robust results for forecasting the gold price instead of employing individual estimates.

The rest of the article is organized as follows: Section 2 presents a summary of studies that analyze predict direction on financial investment instruments. Section 3 contains the architecture of the proposed model and the methods used for the construction of the system. The results of the experiment and the conclusions are presented in Section 4 and Section 5.

II. RELATED WORK

This part provides a summary of the literature studies on regression analysis to estimate the price or direction of

Cite (APA): Kilimci, Z.H. (2022). Ensemble Regression-Based Gold Price (XAU/USD) Prediction. *Journal of Emerging Computer Technologies*, 2(1), 7-12.

different instruments such as digital currencies, stocks, mineral commodities such as gold, silver, bonds, funds and such products.

In a study [1], authors propose to forecast gold prices using multiple regression method (MLR). Various parameters which have an impact on the gold prices are employed to construct feature set such as Commodity Research Bureau future index (CRB), USD/Euro Foreign Exchange Rate (EUROUSD), Inflation rate (INF), Money Supply (M1), New York Stock Exchange (NYSE), Standard and Poor 500 (SPX), Treasury Bill (T-BILL), and US Dollar index (USDX). Authors report that the success of proposed model is competitive for forecasting the price of gold with 85.2% of sample variations in monthly described by the model. In another study [2], authors aim to predict price of gold employing multiple linear regression with principal component analysis (PCA). In order to eliminate the problem of presence of multicollinearity of the explanatory variables, PCA is employed in the experiments. The usage of PCA contributes to the performance of the proposed system by improving prediction accuracy from 0.572 to 0.625. In other study [3], a new consolidated approach (ICA- GRUNN) based on independent component analysis (ICA) and gate recurrent unit neural network (GRUNN) is presented on the estimation of gold price. Authors conclude the paper that ICA-GRUNN outperforms the traditional techniques such as autoregressive integrated moving average (ARIMA), radial basis function neural network (RBFNN), long short-term memory neural network (LSTM), GRUNN, and ICA-LSTM in terms of accuracy.

In a study [4], authors investigate the impact of Chicago Board Options Exchange (CBOE) gold and silver implied volatility on gold futures volatility in China using heterogeneous autoregressive (HAR) and Ridge regression methods. To demonstrate the effectiveness of the proposed models, data is gathered between March 18, 2011 and June 29, 2018. Authors report that HAR and Ridge regression-based models perform better predictive performance compared to the benchmark models. Pierdzioch et al. present a real-time approach based on quantile-regression in order to assess forecast out-of-sample gold returns with the help of macroeconomic and financial parameters [5]. With the usage of real-time quantile-regression technique, model instability, uncertainty, and the possibility that a forecaster has an asymmetric loss function is provided. Authors inform that the forecasts calculated employing the real-time quantile-regression model performs better than an autoregressive model. In a study [6], Suranart et al. focus on the various models namely, neural network, radial basis function network and support vector regression (SVR) to forecast the price of gold. The dataset is collected between June 2008 and April 2013 and evaluated as monthly and weekly. Experiment results indicate that SVR exhibits superior performance on both weekly and monthly data by performing 1.140 of MAPE and 0.908 of MAPE, respectively. In a study [7], authors investigate the effect of decision tree and support vector regression models on the prediction of gold price. The decision tree technique is utilized for the feature selection task while the regression is performed for the purpose of gold index estimation. Experiment results show that the consolidation of decision tree and support vector regression models boost the prediction performance compared to the techniques namely, linear regression and neural network. Sadorsky proposes tree-based classifiers to forecast the direction of gold and silver price [8]. For this purpose, bagging, stochastic gradient boosting, and random forests are employed. The prediction performance of tree-based methods that ranges from 85% and 90% excels when compared to the logit models for 20-day and 10-day estimates. In another study [9], Mithu et. al propose to estimate the gold price employing regression methods for stock market inconsistency and settling economic. For this purpose, authors evaluate support vector regression (SVR), random forest regressor (RFR), decision tree, gradient boosting, and XGBoost methods separately to estimate the daily gold price. The paper is concluded the superior performance of RFR algorithm with 99% of accuracy score. In [10], authors focus on estimating artificial neural networks of gold prices using multiple linear regression models. Experiment results demonstrate that MRL analysis in an effective way to compose the network pattern and get more accurate estimation score with 0.004264% of MSE.

III. Methods

In this part, regression and ensemble regression-based models are introduced used in this work.

A. Linear Regression (LR)

In statistics, linear regression is a linear model for constructing the relation between a scalar answer (dependent) and one or more expository parameters (independent). The state of one expository parameter is named as simple linear regression while the situation is named as multiple linear regression for more than one variable [11]. However, this is rather different from multivariate linear regression, where multiple related dependent parameters are estimated, rather than a one scalar parameter [12]. The correlations are composed of employing linear estimator functions whose unknown model variables are predicted from the data in linear regression. These are called as linear models [13]. Linear regression is the first and broadly employed type of regression analysis in various applications [14]. These are trend estimation in time series data, epidemiology, finance, economics, environmental science, machine learning, etc. Because linear regression is extensively applied in social, behavioral, biological and sciences to define possible relations between parameters, it is as one of the most significant tools employed in these fields.

B. Polynomial Regression (PR)

Polynomial regression is a form of Linear regression employed as a specific version of multiple linear regression which forecasts the relation between independent and dependent variables as an nth degree polynomial [15]. Even though polynomial regression complies a nonlinear model to the data, as a statistical forecast problem it is linear, that is the regression method is linear in the unknown variables that are forecasted from the data. That is why polynomial regression is accepted to be a specific situation of multiple linear regression.

C. Decision Tree Regression (DTR)

This approach is also known as decision tree learning or induction of decision trees which is one of the predictive modelling techniques used in machine learning, data mining, and statistics. It utilizes a decision tree as a predictive model to move from samples about an item demonstrated in the branches to outcomes about the item's goal value indicated in the leaves. In tree models, if the goal parameter will take a discrete set of values are named as classification trees while goal variable will get continuous values are entitled as regression trees [16]. Decision trees are assessed as intelligibly and simple among the other widely-used machine learning techniques [17].

D. Random Forest Regression (RFR)

Random decision forests or random forests are supervised learning techniques that are utilized for both classification and regression. The term forest stands for lots of decision trees. The model builds an ensemble of decision trees at training phase and blends their decisions by making inferences about the goal category (classification) that is the highly voted by the community of decision trees or average estimation (regression) of the base trees. Random forest approach ensures superior predictive accuracy, and do not enable overfitting if there are adequate trees in the forest [18], is proposed by Breiman [19]. In this work, random forest algorithm is employed with 100 estimators.

E. Support Vector Regression (SVR)

Support vector machine (SVM) is initially presented by Vapnik [20]. Support vector classification (SVC) and support vector regression (SVR) are two major categories for SVMs. SVM is a learning framework employing a high dimensional feature space. It provides functions of prediction that are enlarged on a sub set of support vectors. A variant of a SVM for regression model is asserted in 1997 by Vapnik et al. [21]. This model is named as support vector regression (SVR). The method generated by SVC just bounds up with a sub set of the training data, for the cost function for constructing the method does not attach importance to training data that reach out further the margin. Likewise, the model generated by SVR solely is contingent on a subset of the training data, inasmuch as the cost function for establishing the system disregards any training data that is near to the model estimation. SVR is the widely applied version of SVMs. One of the basic properties of SVR is that in place of reducing the error of observed training, SVR proposes to decrease the generalized error bound so as to accomplish generalized success. The generalization error is composed of the consolidation of a regularization term that checks the complexity of the hypothesis space and the error of training.

F. Voting Regressor (VR)

A voting ensemble or majority voting is an ensemble machine learning technique that consolidates the estimations from a lot other method. This generally is employed to develop the success of the system by integrating various models instead of using any single learning technique. In other words, majority voting is utilized for both classification or regression, is performed by consolidating the decisions from multiple methods. In the event of regression [22], it is calculated by taking average of the estimations of all models, while the predictions for each class are aggregated and the final decision on class with the majority vote is estimated in classification [23]. In summary, a voting regressor used in this work is an ensemble meta-predictor that conforms distinct individual regressors, each on the entire dataset. Then, the average of individual forecasts is ensured to decide a final estimation.

G. Stacking Regressor (SR)

Stacking is an ensemble technique that employs a meta-learning method to learn how to best consolidate the estimations from two or more individual machine learning techniques [24]. The advantage of stacking approach is that it can benefit the abilities of a set of well-performed methods on a regression or classification task and make estimations that have better success than any individual method in the ensemble. The meta-model is frequently basic, maintaining a smooth explication of the estimations performed by the individual models. For this reason, linear methods are constantly utilized as the meta-model. For instance, linear regression is employed for regression tasks while logistic regression used for classification tasks for the purpose of forecasting a label of class.

H. Proposed Framework

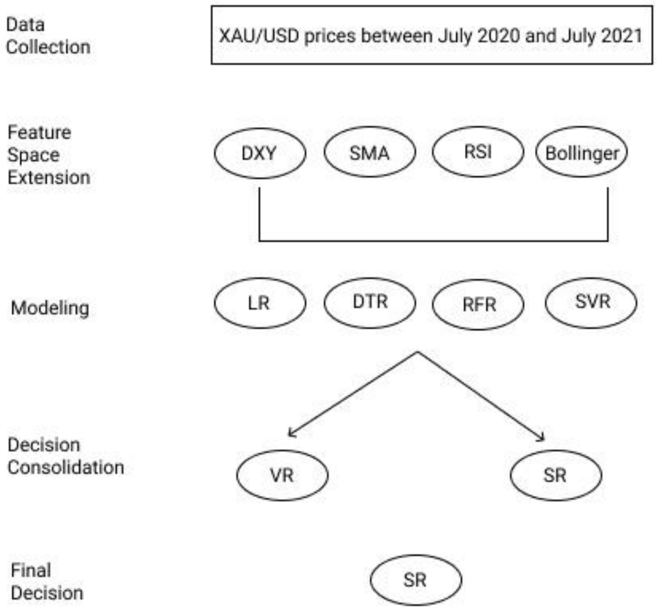


Fig. 1. The architecture of proposed model.

In this work, the prediction of gold price index (XAU/USD) is proposed. First, data collection process is performed between July 2019 and July 2020. Due to weekends and public holidays, 252 days of data were collected from the financial website. Missing 113 days of data is completed with a quadratic decal calculation method. Prices of opening, closing, highest, and lowest XAU/USD index are collected via the investing.com website with the help of Selenium library. The feature space of the data set is extended by collecting different features that may affect the direction and price of the gold index in the same date range. These are simple moving average (SMA) for 20, 50, and 100 days, opening, closing, highest and lowest dollar index (DXY) prices, 14-day relative strength index (RSI), the upper, middle and lower values of the Bollinger band (BB). SMA-20, SMA-50, SMA-100 mean 20-day, 50-day, 100-day simple moving averages of 1 ounce of gold, respectively. RSI-14 represents the 14-day relative strength index of 1 ounce of gold. BB-upper, BB-middle, BB-lower value symbolize the upper, middle, and lower band values of the Bollinger band of 1 ounce of gold. Historical indicator values are calculated using the technical analysis library named TA-lib. After extending feature space, modelling stage is implemented by utilizing various regression models namely, linear regressor (LR), decision tree regressor (DTR), random forest regressor (RFR), support vector regressor (SVR). To consolidate decisions of each regression model, voting (VR) and stacking regressors (SR) employed for the purpose of acquiring more robust decision on predicting XAU/USD index. Figure 1 demonstrates major stages of tasks carried out along this work.

IV. Experiment Results

In this section, we propose to forecast the value of 1 ounce of gold in dollars with the help of regression ensemble-based approaches. Experiment results of base regression models and ensemble regression methods are presented to demonstrate the impact of ensemble-based regression models on estimating XAU/USD index. Voting regressor (VR) and stacking Regressor (SR) are utilized as decision integration approach while linear regression (LR), polynomial regression (PR), decision tree regressor (DTR), random forest regressor (RFR), and support vector regressor (SVR) are evaluated as individual learners. The dataset is randomly splitted into 80% training and 20% test sets. To assess the performance of the proposed model, various evaluation metrics are employed namely, mean absolute percentage error (MAPE), mean absolute error (MAE), mean squared error (MSE), and R-squared (R^2) as below:

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_a - y_f}{y_a}\right| \times 100 \tag{1}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_a - y_f\right| \tag{2}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(y_a - y_f\right)^2 \tag{3}$$

$$R^2 = 1 - \frac{\sum\left(y_a - y_f\right)^2}{\sum\left(y_a - y_m\right)^2} \tag{4}$$

where N is the size of test set, y_a and y_f are the actual and forecasted observations, respectively. In R-squared calculation, numerator corresponds to sum squared regression that is the sum of residuals squared while denominator corresponds to total sum of squares which the sum of the distance the data is away from the mean all squared. The statistical characteristics of the data are given followingly: The minimum value of data is 1404.00, the maximum value is 1810.78, the mean is 1574.508, and median of the data is 1547.00.

In Table I, experiment results of individual and ensemble-based regression techniques are presented in terms of different evaluation metrics. It is obviously observed that stacking regressor with 2.2036 of MAPE outperforms both the other ensemble-based regression model and base regression approaches. It is followed by SVR, PR, and LR with 2.2745, 2.3042, and 2.3816 of MAPE results. In this case, the usage of SVR, PR, or LR as an individual learner is more meaningful rather than the utilization of voting ensemble model in terms of both predict performance and time. On the other hand, voting regressor as an ensemble-based approach exhibits poor MAPE result compared to LR, PR, and SVR techniques. Moreover, it is clearly observed that the utilization of both DTR with 5.1923 of MAPE and RFR with 4.2544 of MAPE is not convenient to forecast the price of XAU/USD index. SVR as an individual method with 2.2745 of MAPE is competitive for forecasting of XAUSD index when compared to the proposed decision consolidation method, namely SR. Also, the utilization of decision tree-based models namely, decision tree and random forest regressors are not convenient to predict the price of XAUUSD when considering both system success and time.

TABLE I. Experiment Results of Individual and Ensemble-based Regression Models

	Evaluation Metrics			
Models	**MAPE**	**MAE**	**MSE**	**R^2**
LR	2.3816	0.0198	0.0022	0.9975
PR	2.3042	0.0182	0.0021	0.9977
DTR	5.1923	0.1396	0.1275	0.9861
RFR	4.2544	0.0213	0.1065	0.9905
SVR	2.2745	0.0116	0.0018	0.9980
VR	3.1004	0.0174	0.0043	0.9969
SR	**2.2036**	0.0109	0.0011	**0.9986**
Avg.	**3.1016**	0.0341	0.0350	0.9950

The inclusion of stacking regressor model ensures roughly 3% improvement while voting regressor provides nearly 2% enhancement in terms of mean absolute percentage error compared to the poorest technique, namely DTR. As a result of Table I, the performance order of models can be summarized as: SR> SVR> PR> LR> VR> RFR> DTR. The success order of all models is similar compared to the other evaluation metrics. In Figure 2, the forecasts of seven different regression models for XAUUSD price are presented. Figure 2 demonstrates actual and predicted prices of XAUUSD employing different regression models. In each model, scatter plot is employed to observe the performance of different models for shuffled train and test data sets. It is obviously seen that the points estimated by the SR model for the test data almost completely coincide with the actual price.

V. Discussion and Conclusion

In this work, regression ensemble-based model is proposed to estimate the value of 1 ounce of gold in dollars (XAU/USD index). To demonstrate the effect of proposed regression ensemble-based model, seven different regression models namely, linear regression, polynomial regression, decision tree regression, random forest regression, support vector regression, voting regressor, stacking regressor are evaluated. For this purpose, the dataset is collected between July 2019 and July 2020 from financial websites, and enhanced with various indicators such as simple moving average (SMA) for 20, 50, and 100 days, opening, closing, highest and lowest dollar index (DXY) prices, 14-day relative strength index (RSI), the upper, middle and lower values of the Bollinger band (BB), and prepared for model construction. To the best of our knowledge, this is the very first attempt considering the consolidation of regression models for the estimation of XAU/USD index. Experiment results indicate that the utilization of stacking regression combination model demonstrates remarkable score with 2.2036 of MAPE for predicting the price of XAU/USD index. In future research, we plan to design a framework which blends both the results of time series analysis and regression models by adding various technical and statistical indicators into dataset.

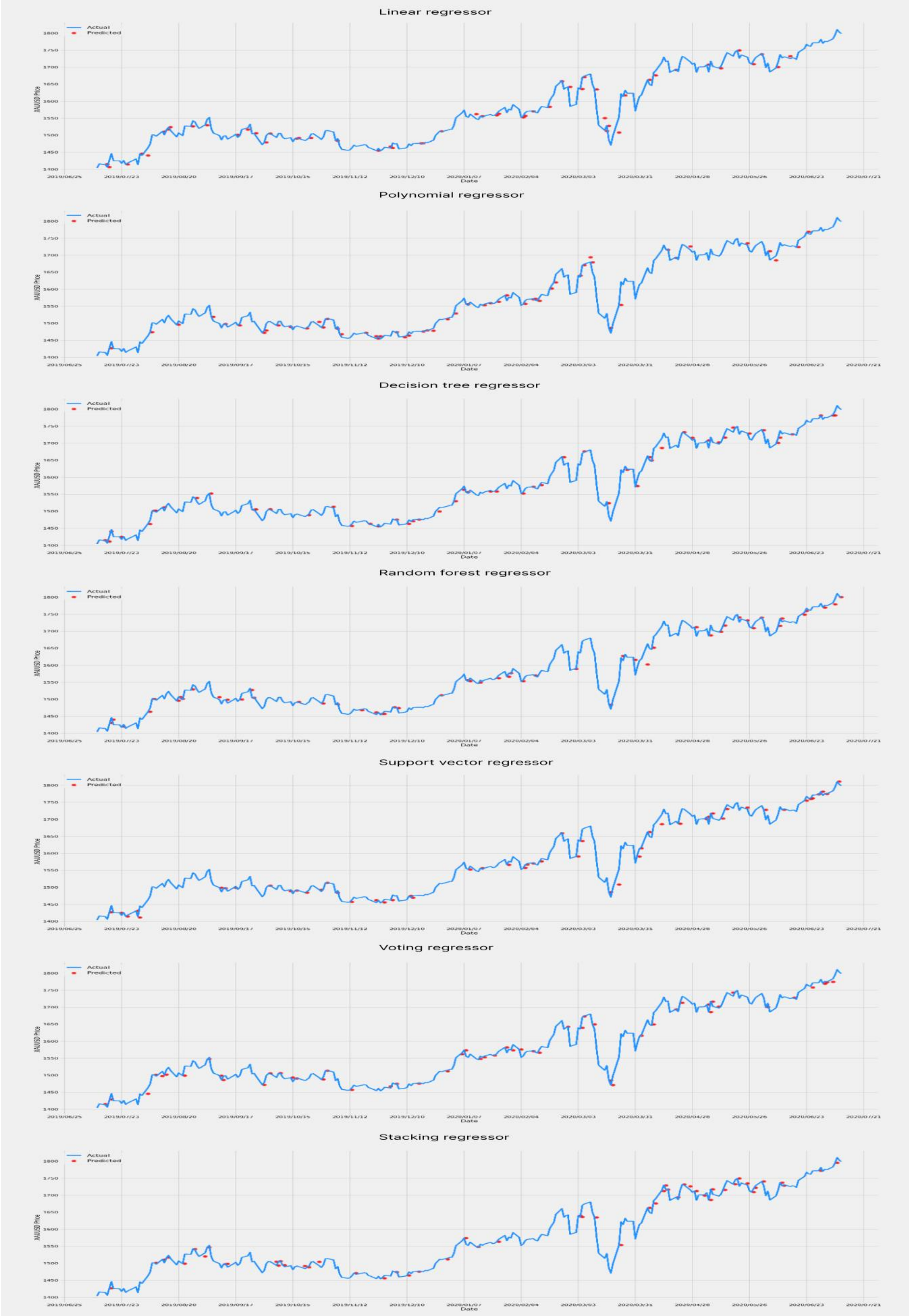


FIG. 2. THE FORECASTS OF SEVEN DIFFERENT REGRESSION MODELS FOR XAUUSD PRICE

REFERENCES

[1]. Ismail, Z., Yahya, A., & Shabri, A. (2009). Forecasting gold prices using multiple linear regression method. American Journal of Applied Sciences, 6(8), 1509.

[2]. Manoj, J., & Suresh, K. K. (2019). Forecast Model for Price of Gold: Multiple Linear Regression with Principal Component Analysis. Thailand Statistician, 17(1), 125-131.

[3]. Jianwei, E., Ye, J., & Jin, H. (2019). A novel hybrid model on the prediction of time series and its application for the gold price analysis and forecasting. Physica A: Statistical Mechanics and Its Applications, 527, 121454.

[4]. Wei, Y., Liang, C., Li, Y., Zhang, X., & Wei, G. (2020). Can CBOE gold and silver implied volatility help to forecast gold futures volatility in China? Evidence based on HAR and Ridge regression models. Finance Research Letters, 35, 101287.

[5]. Pierdzioch, C., Risse, M., & Rohloff, S. (2015). A real-time quantile-regression approach to forecasting gold returns under asymmetric loss. Resources Policy, 45, 299-306.

[6]. Suranart, K., Kiattisin, S., & Leelasantitham, A. (2014, March). Analysis of comparisons for forecasting gold price using neural network, radial basis function network and support vector regression. In The 4th Joint International Conference on Information and Communication Technology, Electronic and Electrical Engineering (JICTEE) (pp. 1-5). IEEE.

[7]. Ongsritrakul, P., & Soonthornphisaj, N. (2003, July). Apply decision tree and support vector regression to predict the gold price. In Proceedings of the International Joint Conference on Neural Networks, 2003. (Vol. 4, pp. 2488-2492). IEEE.

[8]. Sadorsky, P. (2021). Predicting gold and silver price direction using tree-based classifiers. Journal of Risk and Financial Management, 14(5), 198.

[9]. M. A. Mithu, K. M. Rahman, R. A. Razu, M. Riajuliislam, S. I. Momo and A. Sattar. (2021, July). Gold price forecasting using regression techniques for settling economic and stock market inconsistency. In 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), pp. 1-4, IEEE.

[10]. Yanto, M., Sanjaya, S., Yulasmi, Y., Guswandi, D., & Arlis, S. (2021). Implementation multiple linear regresion in neural network predict gold price. Indonesian Journal of Electrical Engineering and Computer Science, 22(3), 1635-1642.

[11]. Freedman, D. A. (2009). Statistical models: theory and practice. cambridge university press.

[12]. Rencher, A. C., & Christensen, W. F. (2012). Chapter 10, Multivariate regression–Section 10.1, Introduction. Methods of multivariate analysis, Wiley Series in Probability and Statistics, 709, 19.

[13]. Hilary, L. (1967). Seal. Studies in the history of probability and statistics. XV: The historical development of the Gauss linear model. Biometrika, 1-24.

[14]. Yan, X., & Su, X. (2009). Linear regression analysis: theory and computing. World Scientific.

[15]. Ostertagová, E. (2012). Modelling using polynomial regression. Procedia Engineering, 48, 500-506.

[16]. Quinlan, J. R. (1990). Decision trees and decision-making. IEEE Transactions on Systems, Man, and Cybernetics, 20(2), 339-346.

[17]. Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Angus, L. B., Yu, P. S., Zhou, Z., Steinberg, D. (2008). Top 10 algorithms in data mining. Knowledge and information systems, 14(1), 1-37.

[18]. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[19]. Korting, T. S. (2006). C4. 5 algorithm and multivariate decision trees. Image Processing Division, National Institute for Space Research–INPE Sao Jose dos Campos–SP, Brazil, 22.

[20]. Vapnik, V. N. (1999). An overview of statistical learning theory. IEEE transactions on neural networks, 10(5), 988-999.

[21]. Vapnik, V., Golowich, S., & Smola, A. (1996). Support vector method for function approximation, regression estimation and signal processing. Advances in neural information processing systems, 9.

[22]. An, K., & Meng, J. (2010, August). Voting-averaged combination method for regressor ensemble. In International Conference on Intelligent Computing (pp. 540-546). Springer, Berlin, Heidelberg.

[23]. Ruta, D., & Gabrys, B. (2005). Classifier selection for majority voting. Information fusion, 6(1), 63-81.

[24]. Džeroski, S., & Ženko, B. (2004). Is combining classifiers with stacking better than selecting the best one?. Machine learning, 54(3), 255-273.

Journal of Emerging Computer Technologies
Research Article
Received: *2022-04-14* | ***Reviewing:*** *2022-04-14 & 2022-05-21* | ***Accepted:*** *2022-05-21* | ***Online:*** *2022-05-22* | ***Issue Date:*** *2022-06-30*

Social Media Content Review of MMORPG Games: Reddit Comment Scraping and Sentiment Analysis

Kaan ARIK
Beykoz University
Department of Digital Game Design
İstanbul, Türkiye
kaanarik@beykoz.edu.tr
0000-0002-0930-8955

Abstract **— Social media is a system that provides access from one-way information sharing to two-way and simultaneous information sharing, introducing Web 2.0 as a service to users. Social media is a set of dialogues and exchanges that people have with each other on the Internet. Reddit is which can call an important audience in these environments. Popular topics and content are available, such as science, sports, gaming, music, food and drink, and photography. After the release of MOBA games, there has been a serious decrease in playing time of the MMORPG genre. The research aims to sentiment analysis of content created on the MMORPG subreddit channel on Reddit. In my study, I focused on sentiment analysis of MMORPG games, which have been very popular for years. Possible reasons for that were tried to be evaluated relative to players' opinions. Sentiment analysis was performed based on posts from the 'MMORPG' subreddit on Reddit. Negative, positive, and neural structures are explained. Frequency analysis of often used words is also included.**

Keywords — MMORPG games, Reddit, sentiment analysis, social media

I. Introduction

The concept of the game has taken its place in the literature with different definitions of scientists working in fields such as philosophy, history, and sociology. It has done a lot of work to understand and explain the content of games and how they can relate to each other. It is difficult to come up with a precise definition of a process or structure in constant change. Because of its game structure, it is an activity that appeals to different audiences in different periods [1]–[11]. The aim here is to find the common elements of the definitions, distinguish the problems, and then present the basic definition of the game, considering the previous definitions.

Different scholars from psychology, physiology, philosophy, history, sociology, and cultural studies have tried to define the change and structure of the concept of plain game definition. Johan Huizinga, a philosopher and history professor who contributed to the studies of games, mentions that "human society" has been familiar with the game since ancient times and has been based on extensive primitive activities [12]. It also emphasizes that the game contains some elements, that it is based on integrity and voluntariness, and that it must occur within the framework of certain rules. A game is an activity that happens in a certain place, within limits of time and will, under an order, in line with an individual's consent and away from the reality of contemporary life, and is an activity that directs social and cultural activities [6], [12] [13].

A. Digital Game Trend

Devices such as computers, tablets, smartphones, and PDAs offer a different innovations to our lives every day. We often need and actively use these devices while shopping, having fun, communicating, and in our daily life. These devices affect our lives. Along with developing technology, more time is allocated to the use of technology compared with previous generations. Also, the definition of individuals born and raised as "Digital Natives" are at a higher level to have technology competence compared to their predecessors [14]. Games and the development of technology have become an indispensable part of our lives and are playable on different platforms. Undoubtedly, in this dimension, mostly devoted to technology is spent on digital games, considering the number of players and time played [15].

Digital games, which are in constant change and development, have become a domain of study that gathers together different disciplines. Digital games are an interdisciplinary study with science fields such as computer science, film, art, animation, new media, business and management sciences, semiotics, and psychology [16]. According to these digital games A new type of expression such as drama, opera, or film is described as a new form of collective behavior according to the social science approach, and as an invention according to computer scientists, engineers, and industrial designers [17]. Newman who carried out research on game science and computer games at Bath Spa University in the United Kingdom stated that academic studies on game theory were insufficient and ignored [18]. Likewise, Wolf and Perron [16] stated that a clear working group has not emerged because of the sterility of academic studies in digital games. Along with the continuation of academic studies, technology has also developed and taken an important place in our lives. Digital games, which include video and computer games with their new definition, appear as a cultural form, but continue to boom as a media technology and a global industry. Digital games are text-based or image-based entertainment software that uses an electronic platform, such as personal computers or consoles, and plays one or more players in a physical or networked environment [19]–[24]. According to [24], a digital game is any game that can be played on digital devices and an interactive program for one or more players, developed to increase the fun as much as possible at the simplest level. It is an activity similar to a normal game with intense interest and competition, but performed through a device with its own unique rules [25].

Cite (APA): Arık, K. (2022). Social Media Content Review of MMORPG Games: Reddit Comment Scraping and Sentiment Analysis. *Journal of Emerging Computer Technologies*, 2(1), 13-21.

A digital game is an electronic game played with images on a video screen, often involving fast-paced action. These are software systems that interact with a user interface to create visual feedback on a computer or video device, using many elements such as entertainment, games, win/lose, and competition [21]. It is a particular type of digital entertainment in which the player interacts with a digital interface and encounters different types of difficulties, depending on the game [26]. The digital game is an activity that arises from the recreational activities of a certain community in their spare time and has become one of the major structures of the culture of modern societies [27]. In its most basic form, digital game is interactive digital entertainment that can be played via computer, console, smartphone, or tablet [28]. Again, with a different definition, digital game is the general name of software code designed to entertain or educate the individual [29].

B. Evolution of Digital Games from the ’90s

In 1990, Microsoft introduced the classic card game, Solitare, into the operating system with the Windows 3.0 version as a package. Based on the period, Solitare [30], which used an easy-to-play and ordinary game model in its period, was among the most popular games for players. In 1991, SEGA company took a big step in the game industry and released a game console called GENESIS, known as MEGA DRIVE in Japan, which was produced for playing at home with 16-bit memory. He developed a game called SONIC for this console. In 1992, Dune II, which was developed by Westwood Studios, inspired by the movie Dune, which met with the audience in 1984, and Frank Herbert’s sci-fi novel with the same name, has become another popular game played by actors [31]. In the same year, “Mortal Kombat” was released as a fighting game by Midway Games. About 3 years after the game’s release, the movie was shot in 1995 and the game became known by everyone.

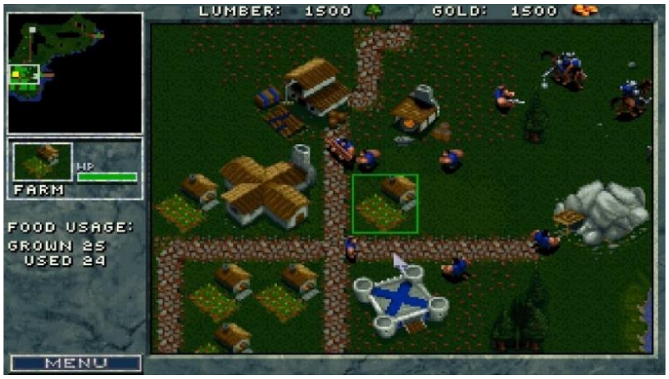


FIG. 1. SCENE OF WARCRAFT: ORCS & HUMANS

In 1993, Doom, which is very common in the game’s history, appears as an FPS computer game by the company “id Software”. Doom was presented to the players with different versions in the following years and has become one of the popular games still played by some players today. In 1994, the game WarCraft Orcs & Humans [32] by Blizzard Entertainment, which is a game still influencing today and is a developer company of many popular games, made a lot of noise. Although WoW is not the first game of its kind to be released, it has become popular much faster because players love its mechanics and dynamics.

C. Digital Game Platforms

Digital games have undergone a substantial change and development since the release of their predecessors, both in terms of content and gameplay. The diversity of digital gaming platforms is increasing every year. While these platforms offer unique experiences to players, they engage players in a fictional world virtually. Interaction between players in digital games can be with personal computers or mobile phones, as well as on coin-operated machines or consoles specially developed for certain games. Digital games can range from relatively simple and text-based games to graphically rich virtual environments developed by a large team. While some games are played in small groups as a single player, others can be played in multiplayer by masses that are not geographically close to each other [33].

The first game in the electronic games category was developed in 1947 on a cathode ray tube amusement device. In the digital game industry, various consoles have been offered to players since the 1960s. While games were played on analog system electronic devices in the 1960s, they were now divided into different categories as computers, consoles, and mobile devices with the development of technology. While computers and consoles were used until the 1990s, today mobile devices are included in this genre. While expressing the type of digital games, the “platform” category refers to the hardware and systems in which the game is played. Personal computers, consoles, and smart mobile devices can also be added to the group [34]. On the personal computer platform, Windows and Linux are grouped as PlayStation, XBOX, Nintendo Switch, and Wii in the console category, and Android and IOS on the mobile platform. A detailed visual diagram of the platforms is given in Figure 2.

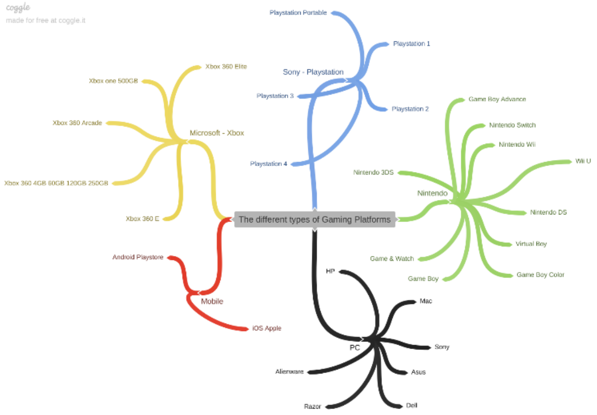


FIG. 2. DIGITAL GAMING PLATFORMS AND SUB-BRANCHES

Categorizing game genres is an important aspect of understanding games in a particular order and providing easy access to them. Video game genres are evaluated under different categories. Categorization is based on the playing styles and camera angles of the games [35]. When categorizing a video game, the basic nature of the interaction within the game itself is considered, rather than asking whether the game is played in a single-player, multiplayer, or over a network. A game is based on the platform on which it is played (PC, mobile, or console), the style of play it provides (multiplayer, networked, or single player), the game world (first person or third person), the rules and objectives that make up the style of play (racing game or action-adventure, etc.) or representational aspects (science fiction, fantasy or action). All of these classifications are closely related to each other [36]. These species are; action/adventure, board, combat, platforming, simulation, racing, role-playing, and strategy. Below are detailed definitions according to the gameplay (gameplay) and camera angle (perspective and viewpoint) factor. Definitions and categorization of game genres below were developed by [34] and [16] and included in literature by [37] in the future. These game types are given below, and sample games have been added to each genre by the researcher.

D. Game Genre

FIG. 3. SCENE OF ALBION: ONLINE

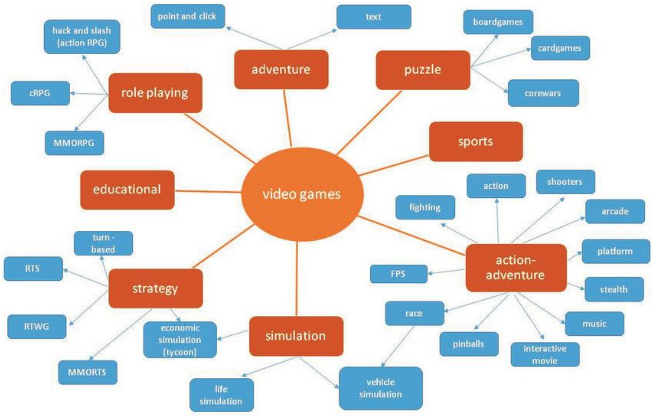


FIG. 4. TYPES OF DIGITAL GAMES BY GAMEPLAY AND CAMERA ANGLES

- **RPG (Role Playing Game):** It is a type of digital game in which players assume the role of characters in a fictional environment and learn a story from beginning to end and accomplish it [38]. Actors take on the responsibility of portraying these roles within a narrative, along with a structured decision-making process regarding actual roles or character development. Games such as The Witcher series, Undertale and Dark Souls are in this category.

- **MMORPG (Massively multiplayer online role-playing game):** It is the name given to the online role-playing game genre played by gamers. Also, type of game in which players from different geographical locations actively get on a character and direct with an internet connection. Blade and Soul, WoW, Albion Online, and EVE online are examples of MMORPGs.

- **MOBA (Multiplayer Online Battle Arena):** Online multiplayer battle arena, or MOBA, is a genre of real-time strategy war game. League of Legends, DOTA II, and Heroes of the Storm Fall could include in this category.

II. SOCIAL MEDIA AND REDDIT

Social media is a system that enables access from one-way information sharing to two-way and simultaneous information sharing, putting Web 2.0 at the service of users [39]. Social media is actively used by many individuals and institutions. In this way, quick access is facilitated, users can view content, articles, news, thoughts, daily events, and photos through social media and reflect their views on this social network. If we are to define the social media that is widely used by people today, the first term that comes to mind is "sharing"[40].

It is a human communication that focuses on sharing and discussion, without time and place constraints (mobile-based). On social media platforms, you meet people and communicate with them. You also help people, get help, answer their questions, and ask questions. In this regard, social media is also one of the ways of informal education. Technology has a structure in which telecommunication and social communication are mediated through words, images, and sound files. People also have a framework in which they share their stories and experiences in this context [41].

Social media has some features of “user-sourced media” based on many-to-many paradigms in terms of communication, rather than broadcasting to many people from a single point, as in traditional media. Social media has also revealed concepts of “Content Produced by Users” and “Media Produced by Customers”, and with this structure, it has gained meaning in the commercial plan.

Data scraping is done through different social media platforms. After data is obtained, sentiment analysis is performed with different natural language processing techniques. Access to data is provided by using different topics and keywords. It's frequently used in the entertainment industry such as finance, news, banking, automotive, movies, and games. In this study, the studies carried out in the field of gaming are highlighted. By scraping Youtube comments, trend model definition and sentiment analysis of game channels were carried out [42]. On the other hand, by accessing the data via Twitter, another popular social media platform, research was conducted according to the game preferences of the players during the COVID 19 period [43]. Another study included the toxicity analysis of players in online communities [44]. Finally, there is a study on the analysis of sexist approaches to games on gender, masculinity, and video game keywords with the data scraped on Reddit [45].

A. Reddit

Reddit is an American social news discussion site. Users registered to site can share posts that contain links, texts, photos, surveys, and videos, and other users can vote on these posts so that the post goes up or down on the page. Users create posts on Reddit; it’s shared in sub-groups called “subreddit”, which cover different topics such as sports, games, music, food, photography, and news [46]. If a post shared on a subreddit gets enough votes, that post will appear on Reddit's home page. Every user registered on the site can create their subreddit. Users create posts on Reddit; it's shared in sub-groups called "subreddit" that deal with different topics such as sports, games, music, food, photography. Users can comment on posts and upvote (upvote) or downvote the post. Users can also vote on comments and reply to them as well. Users get the so-called "karma" score based on the votes they receive. This score appears on their profile and can drop to a negative number.

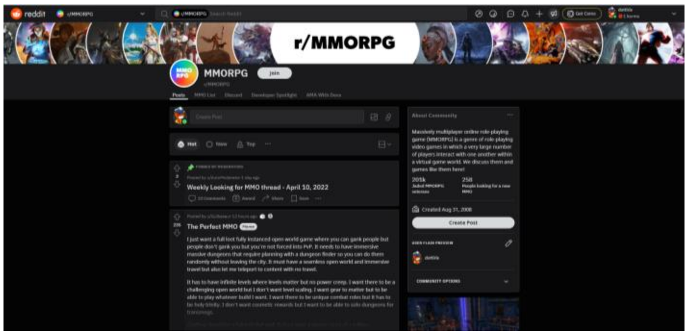


FIG. 5. REDDIT WEB PAGE

B. PRAW and Reddit Data

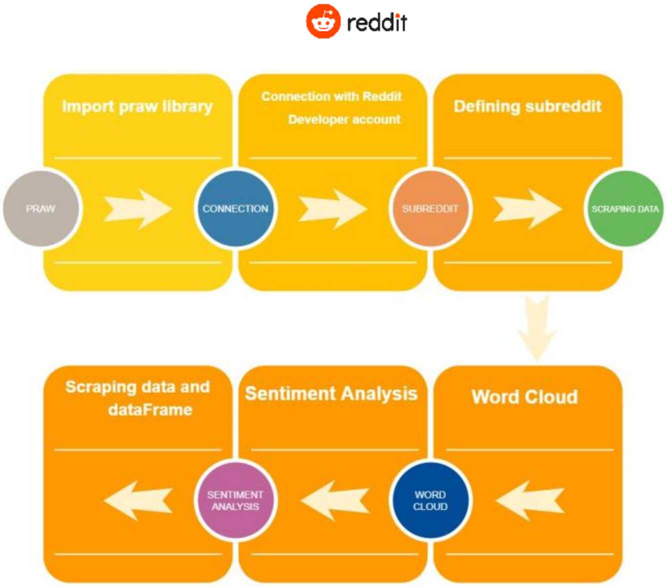
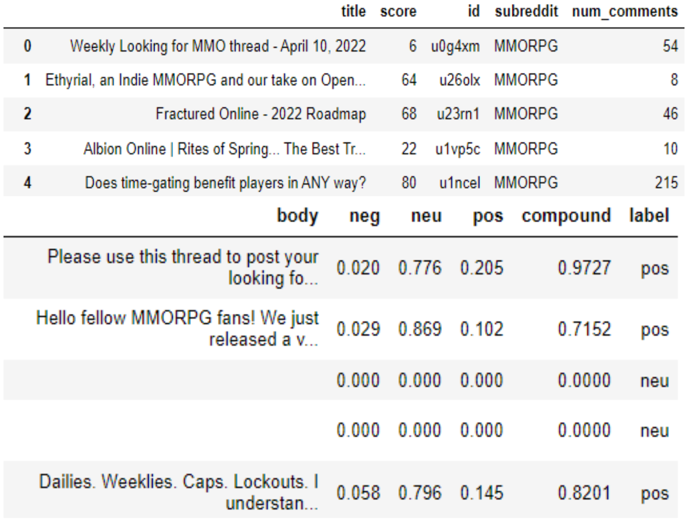


FIG. 6. PRAW AND DATA COLLECTION SCHEMA

PRAW) an acronym for "Python Reddit API Wrapper", is a Python package that allows for simple access to Reddit's API. PRAW aims to be easy to use and internally follows all of Reddit's API rules [47].

Steps of PRAW library:

- Import PRAW library,
- Connection with Reddit Developer account,
- Defining subreddit,
- Scraping data and data frame,
- Sentiment Analysis,
- Word Clouds



```
                                              title  score      id  subreddit  num_comments
0     Weekly Looking for MMO thread - April 10, 2022      6  u0g4xm     MMORPG            54
1  Ethyrial, an Indie MMORPG and our take on Open...     64  u26olx     MMORPG             8
2                     Fractured Online - 2022 Roadmap     68  u23rn1     MMORPG            46
3       Albion Online | Rites of Spring... The Best Tr...  22  u1vp5c     MMORPG            10
4    Does time-gating benefit players in ANY way?     80  u1ncel     MMORPG           215

                              body    neg    neu    pos  compound  label
   Please use this thread to post your
                         looking fo...  0.020  0.776  0.205    0.9727    pos
  Hello fellow MMORPG fans! We just
                         released a v...  0.029  0.869  0.102    0.7152    pos
                                        0.000  0.000  0.000    0.0000    neu
                                        0.000  0.000  0.000    0.0000    neu
  Dailies. Weeklies. Caps. Lockouts. I
                         understan...  0.058  0.796  0.145    0.8201    pos
```

- FIG. 7. EXAMPLE OF REDDIT SCRAPING DATA

As seen in Figure 7 title, score, id, subreddit, number of comments, body, negative, neutral, positive, compound, and label attributes are included in data retrieved from Reddit, and those are written to a .csv file. Data consists of 869 rows (comments) and 11 columns.

TABLE 1: BREAKDOWN OF DATASET

Data Collection Summary		
First Post	Last Log	Number of Post
14.06.2007	30.03.2022	5158 post

TABLE 2: DATASET INFORMATION AND ATTRIBUTES

Data Information	
Title	Title of topic
Score	Score of topics
ID	Identifier number
Subreddit	Name of subreddit
num_comments	Number of comments
Body	Topic text and comments
Neg	Negative ratio
Nue	Neutral ratio
Pos	Positive ratio
Compound	The compound ratio of text
label	Grouping positive, negative, and neutral

III. NATURAL LANGUAGE PROCESSING (NLP)

Natural language processing (NLP) is "computers can understand, process, interpret, and even produce sentences in spoken language." It is also a discipline in which computer science (especially artificial intelligence and machine learning) and linguistics are used together. Technologies such as the chatbot we use today on a bank's website, the commands we give to the assistant on our phone, the translations we make in Google/Microsoft translate, and the prediction of the next word by our phone when writing a message are the result of natural language processing. Text mining, which has become very popular recently, is also part of natural language processing. Thanks to text mining, we can process the thoughts that pile up on the Internet and make sense of them. For example, using the Twitter API, we can find out what percentage of those who receive and write tweets about climate change think climate change is dangerous and have information about the public. In addition, natural language processing is used in speech recognition. Technologies such as speech recognition and automatic lip-reading are used to assist the hearing impaired as well as in monitoring [48].

Natural language processing processes differ depending on the language. The computer first examines the transformation of words with suffixes at the root. The first process is called lexical. Then it tries to understand what the words mean based on their order in the sentence. The second process is called syntactic. Then it examines what the sentence is trying to explain. The third process is called semantic. Finally, it examines what the sentences are trying to express by putting them together, which is pragmatic.

In summary, the computer learns the context of speech by examining the root of word separately, ordering of words separately, meaning of sentence and speech, and extracts a meaning from this speech [49].

Examples of 10 different projects can be given to the application areas [50].

- Text Classification and Categorization
- Named Entity Recognition (NER)
- Part-of-Speech Tagging
- Semantic Parsing and Question Answering
- Finding Interpretation (Paraphrase Detection)
- Language Generation and Multi-document Summarization
- Language Translation (Machine Translation)
- Speech Recognition
- Character Recognition
- Spell Checking

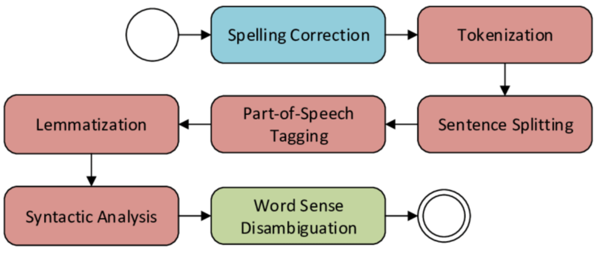


FIG. 9. NLP PIPELINE [51]

A. VADER (Valence Aware Dictionary for Sentiment Reasoning)

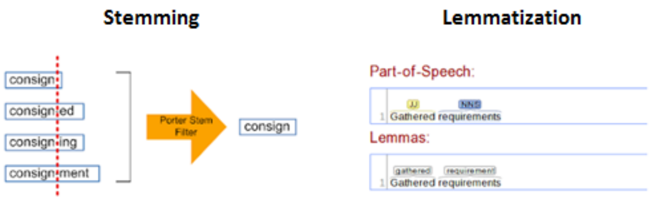


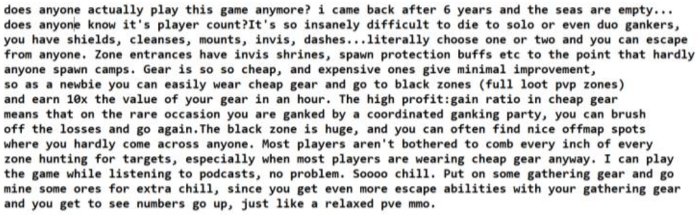
FIG. 10. STEMMING AND LEMMATIZATION WORKFLOW

VADER is a dictionary and rule-based sentiment analysis tool specifically designed for predicting emotions in social media texts. They create a generalizable and valence-based gold standard sentiment dictionary for social microblogging platforms by using a combination of some qualitative and quantitative methods in 2014. They then combine these lexical features with five general rules that people use when expressing or emphasizing emotional intensity. In this approach, each of the words in the dictionary has valence values that indicate both positive and negative emotional polarity and emotional intensity on a scale of -4 to +4. VADER analyzes a text by checking whether any of the words in the text are in the dictionary. It also checks some of the rules mentioned above. It then uses the word ratings to create four emotion metrics: positive, neutral, negative, and composite ratings. The first three, positive, neutral, and negative, indicate the proportion of the text that falls into these categories. The last score, the composite score, is calculated by adding the valence scores of each word in the dictionary, adjusting them according to the rules, and then normalizing them from -1 (most negative) to +1 (most positive) [52].

B. Text Cleaning and Processing

Let's look at a few terms before we delve into text cleanup.

- **Stem:** Stem allows us to get to root of word. To give an example with three words, if words eat, eat, enough are found in same text, stemming process takes root of all three.
- **Root:** Root is a word root, as in English, and is expected to have a meaning.
- **Lemma:** Lemma is morphological rooting of the word. For example, the lemma of the word "touchy" is "to take", we can define it as dictionary equivalent of plain form of words.
- **Stemming:** Stemming is name given to taking root of word. Stemming NLP varies according to applied language.



```
does anyone actually play this game anymore? i came back after 6 years and the seas are empty...
does anyone know it's player count?It's so insanely difficult to die to solo or even duo gankers,
you have shields, cleanses, mounts, invis, dashes...literally choose one or two and you can escape
from anyone. Zone entrances have invis shrines, spawn protection buffs etc to the point that hardly
anyone spawn camps. Gear is so so cheap, and expensive ones give minimal improvement,
so as a newbie you can easily wear cheap gear and go to black zones (full loot pvp zones)
and earn 10x the value of your gear in an hour. The high profit:gain ratio in cheap gear
means that on the rare occasion you are ganked by a coordinated ganking party, you can brush
off the losses and go again.The black zone is huge, and you can often find nice offmap spots
where you hardly come across anyone. Most players aren't bothered to comb every inch of every
zone hunting for targets, especially when most players are wearing cheap gear anyway. I can play
the game while listening to podcasts, no problem. Soooo chill. Put on some gathering gear and go
mine some ores for extra chill, since you get even more escape abilities with your gathering gear
and you get to see numbers go up, just like a relaxed pve mmo.
```

FIG. 11. EXAMPLE OF COMMENT DATA AFTER TEXT-CLEANSING

There are three types of stemming algorithms: **Snowball Stemmer, Porter Stemmer and Lancaster Stemmer** [53] [54] [55]. All are available in Python's NLTK library. Porter Stemmer is the oldest of them all, and to put it simply, he tries to find a common root by removing the common endings of the words he finds. Snowball Stemmer, an improved and more aggressive version of Porter Stemmer – also called Porter2 – runs faster, so it is more used. Lancaster Stemmer is the most aggressive algorithm among them, so that sometimes it can find roots that don't really mean anything, but the good thing is that it's more tamperable.

- **Lemmatization:** Lemmatization examines words morphologically. As an example: "They are going" consists of third person plural form of verb to go in present tense. Here, initial unconjugated form of word is called a lemma. In this example, going is a lemma. Lemmatization algorithms need a dictionary to work. Likewise, if we give an example in English, "Feeds" is augmented form of third person singular of verb feed.

- **Tokenizing:** Tokenizing can be defined as breaking a sentence into smaller meaningful units. Tokens are meaningful small units, symbols, words, phrases can be given as examples of tokens. The parsing changes depending on the tokenizer you use.

IV. FINDINGS

VADER is sentiment analysis method in research method. After stemming, lemmatization and tokenizing stages, text seen in Figure 10.

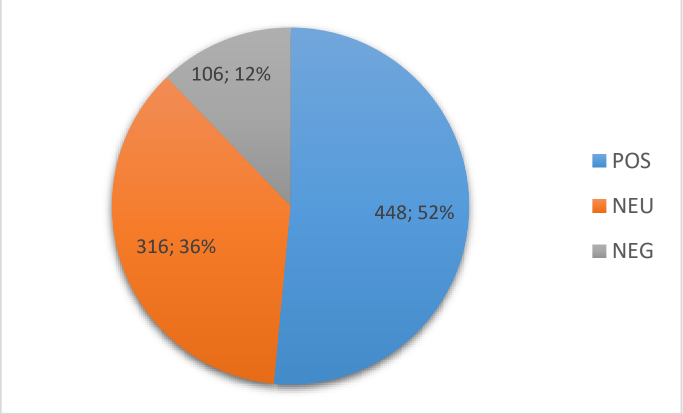


FIG. 12. PIE-CHART FOR DISTRIBUTION OF SENTIMENT FREQUENCY

As seen in Figure 12, sentiment analysis distributed positive (n=448, %52), neutral (n=316, %36), and negative (n=106, %12). Comments are mostly positive and neutral. Number of negative comments is very few.

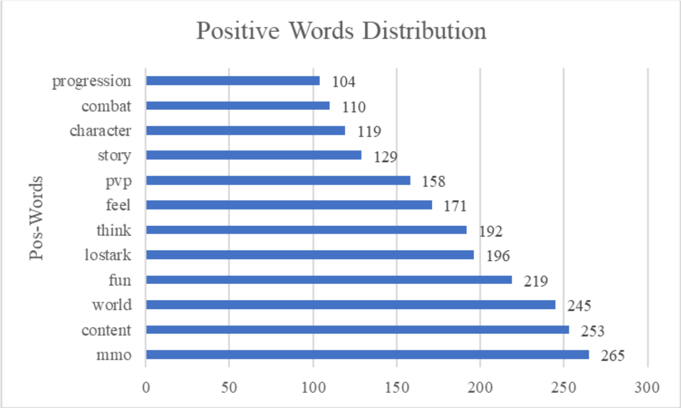


FIG. 13. POSITIVE WORDS DISTRIBUTION

According to Figure 13, can interpret MMO (Massively Multiplayer Online), which is in positive group, focuses on content, world, entertainment, and feeling. Content and fun players are among most important structures make MMORPG games unique.

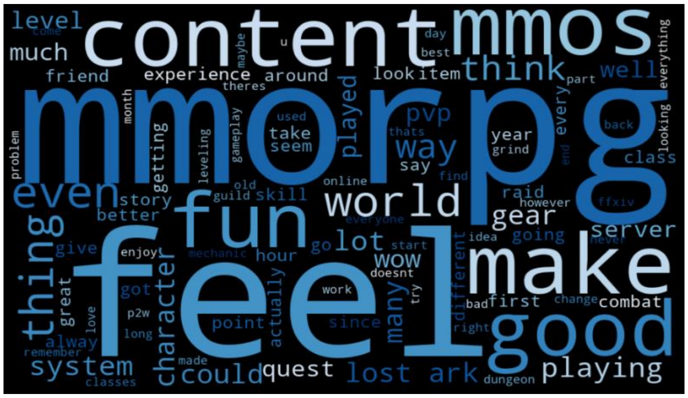


FIG. 14. WORD CLOUD POSITIVE WORDS

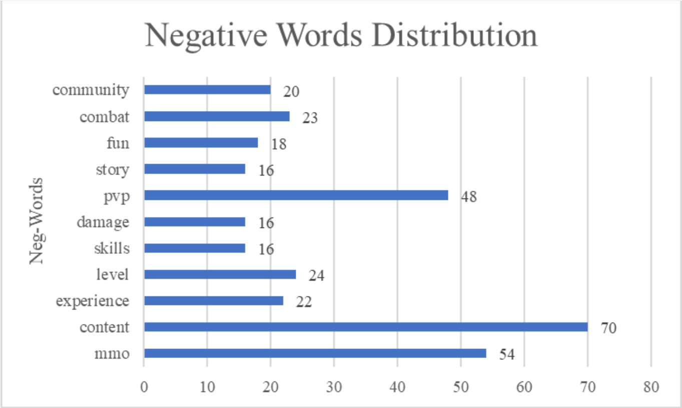


FIG. 15. NEGATIVE WORDS DISTRIBUTION

According to Figure 15, can interpret which is in negative group, focuses on content, pvp, story, combat and community. Content and fun players are among most important structures that make MMORPG games unique.

According to Figure 17, can interpret which is in neutral group, focuses on game, players, items, people, combat and content. Combat and items are among most important structures after players and game that make MMORPG games unique.

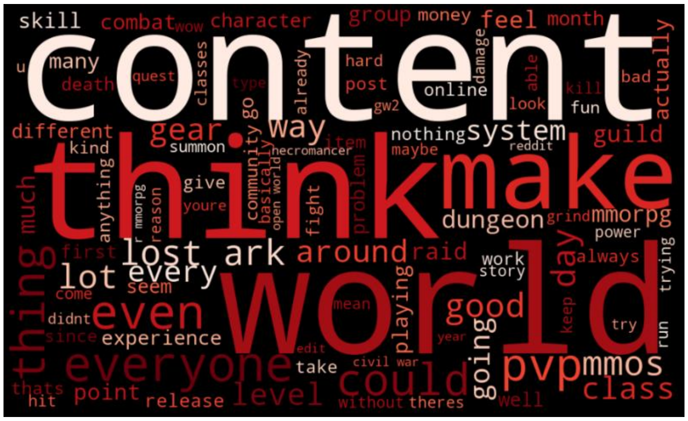


FIG. 16. WORD CLOUD NEGATIVE WORDS

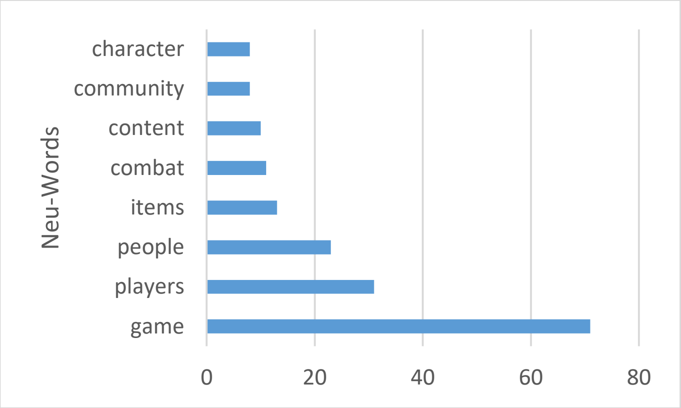


FIG. 17. WORD CLOUD NEUTRAL WORDS

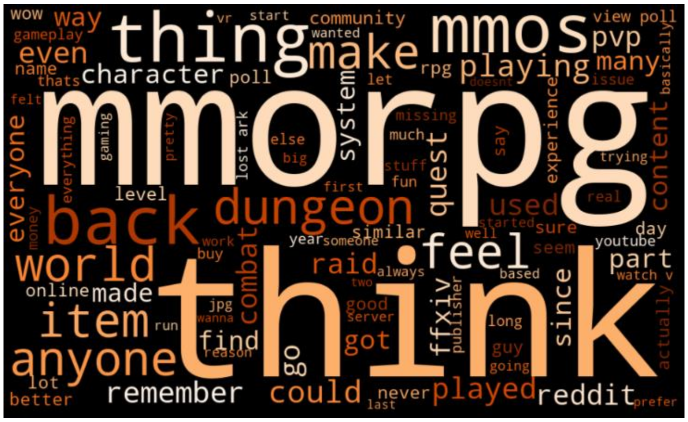


FIG. 18. WORD CLOUD NEUTRAL WORDS

V. CONCLUSION

People who are attracted to MMORPG are people who like to Role-play online with other real people, hence the acronym MMORPG (massively multiplayer online roleplaying game). Not surprisingly, quite a lot of MMORPG gamers rarely play any other type of game (or at the least, don't even come close to putting the amount of time into other game genres).

It's the social interaction and the character development that's the real draw. While other genres like single-player RPGs, open-world, 1st, and 3rd-person action, and etc., have bits and pieces of the same features, MMORPG generally covers all the bases for giving a player the feeling of being part of a living, breathing virtual world. Features like player housing, player-driven economy, factions, the high capacity of players, and the immense size of the world, are what typically set it apart from other video game genres [56].

Positive, neutral and negative sentiment distributions are given in Figure 12. Information in the comments, gameplay accessibility, mechanics, arena battles, and art style why it's so booming in a major base game is reflected in the attractive feature results. Another factor is the story. A strong story of WoW and similar games are among the top preferences of the players [57]. Some concepts are in both groups, this is because some players like factors such as content, combat, community, etc., while others do not, for example.

As a result, there are common words in positive, neutral, and negative group. These words are positive for some players and negative for others. However, after an element in the game industry is mentioned negatively by the players, it remains in the minds for a long time. In this respect, although the content and multiplayer structure are in the positive group, they represent the negative emotion more.

However, one of main reasons why MMORPG games are still played by players is the community, which is one of the important elements. That is, players play the game to belong to a community and to be in constant communication with that group [58]. In this respect, it is important that games that have been or will be released in this genre pay attention to the community factor.

REFERENCES

[1] C. S. Ang and P. Zaphiris, “Computer Games and Language Learning:,” in *Handbook of Research on Instructional Systems and Technology*, T. T. Kidd and H. Song, Eds. IGI Global, 2008, pp. 449–462. doi: 10.4018/978-1-59904-865-9.ch032.

[2] E. M. Avedon and B. Sutton-Smith, *The study of games*. New York, N.Y. [u.a]: Ishi Press, 2015.

[3] G. Costkyan, “I have no words & I must design: toward a critical vocabulary for games,” in *Proceedings of the computer games and digital cultures conference, Finland*, 2002, pp. 9–33.

[4] R. E. Cardona-Rivera, J. P. Zagal, and M. S. Debus, “GFI: A Formal Approach to Narrative Design and Game Research,” in *Interactive Storytelling*, vol. 12497, A.-G. Bosser, D. E. Millard, and C. Hargood, Eds. Cham: Springer International Publishing, 2020, pp. 133–148. doi: 10.1007/978-3-030-62516-0_13.

[5] J. Juul, “Games telling stories? A brief note on games and narratives,” *Game Stud.*, vol. 1, no. 1, pp. 1–12, 2001.

[6] Kevin Maroney, “My Entire Waking Life,” *http://www.thegamesjournal.com/*, 2011. http://www.thegamesjournal.com/articles/MyEntireWakingLife.shtml (accessed Apr. 13, 2022).

[7] R. A. Myers and G. Mertz, “The Limits of Exploitation: A Precautionary Approach,” *Ecol. Appl.*, vol. 8, no. 1, p. S165, Feb. 1998, doi: 10.2307/2641375.

[8] R. J. Paddick, “*The Grasshopper: Games, Life and Utopia.* By Bernard Suits. Toronto, University of Toronto Press 1978,” *J. Philos. Sport*, vol. 6, no. 1, pp. 73–78, Jan. 1979, doi: 10.1080/00948705.1979.10654153.

[9] G. Tavinor, “Definition of videogames,” *Contemp. Aesthet. J. Arch.*, vol. 6, no. 1, p. 16, 2008.

[10] N. Whitton, *Digital Games and Learning: Research and Theory*, 0 ed. Routledge, 2014. doi: 10.4324/9780203095935.

[11] K. S. Tekinbaş and E. Zimmerman, *Rules of play: game design fundamentals*. Cambridge, Mass: MIT Press, 2003.

[12] J. Huizinga, *Homo ludens: a study of the play-element in culture*. Kettering, OH: Angelico Press, 2016.

[13] C. Crawford, *Chris Crawford on game design*. Indianapolis, Ind: New Riders, 2003.

[14] M. Prensky, "Digital natives, digital immigrants part 2: Do they really think differently?," *Horiz.*, 2001.

[15] GAMINGINTURKEY, "Türkiye Oyun Sektörü 2021 Raporu ve Detayları," Dec. 11, 2021. https://www.gaminginturkey.com/tr/turkiye-oyun-sektoru-2021-raporu-ve-detaylari/ (accessed Apr. 13, 2022).

[16] M. J. P. Wolf and B. Perron, Eds., *The video game theory reader*. New York ; London: Routledge, 2003.

[17] J. Murray, I. Bogost, M. Mateas, and M. Nitsche, "Game Design Education: Integrating Computation and Culture," *Computer*, vol. 39, no. 6, pp. 43–51, Jun. 2006, doi: 10.1109/MC.2006.195.

[18] J. Newman, *Videogames*, 2nd ed. London ; New York: Routledge, 2013.

[19] G. Frasca, "Videogames of the oppressed: Videogames as a means for critical thinking and debate," Master's Thesis, School of Literature, communication, and culture, Georgia Institute of …, 2001.

[20] N. B. Sardone and R. Devlin-Scherer, "Teacher Candidates' Views of Digital Games as Learning Devices," vol. 18, no. 2, p. 21, 2009.

[21] S. S. Shabanah, J. X. Chen, H. Wechsler, D. Carr, and E. Wegman, "Designing Computer Games to Teach Algorithms," in *2010 Seventh International Conference on Information Technology: New Generations*, Las Vegas, NV, USA, 2010, pp. 1119–1126. doi: 10.1109/ITNG.2010.78.

[22] T. Govender and J. Arnedo-Moreno, "An analysis of game design elements used in digital game-based language learning," *Sustainability*, vol. 13, no. 12, p. 6679, 2021.

[23] J. Arjoranta, "How to Define Games and Why We Need to," *Comput. Games J.*, vol. 8, no. 3–4, pp. 109–120, Dec. 2019, doi: 10.1007/s40869-019-00080-6.

[24] Beth E. Kolko, "Digital Games Course Definitions," *Technical Communication 498 Digital Games*. http://faculty.washington.edu/bkolko/games/definitions.shtml (accessed Apr. 13, 2022).

[25] A. K. Przybylski and N. Weinstein, "How we see electronic games," *PeerJ*, vol. 4, p. e1931, Apr. 2016, doi: 10.7717/peerj.1931.

[26] P. Zackariasson and T. L. Wilson, Eds., *The video game industry: formation, present state, and future*, 1. published. New York, NY; London: Routledge, 2012.

[27] J. Breuer, J. Vogelgesang, T. Quandt, and R. Wendt, "Violent Video Games and Physical Aggression: Evidence for a Selection Effect Among Adolescents," *Psychol. Pop. Media Cult.*, vol. 4, pp. 305–328, Feb. 2015, doi: 10.1037/ppm0000035.

[28] Phil Owen, "What Is A Video Game? A Short Explainer." https://www.thewrap.com/what-is-a-video-game-a-short-explainer/ (accessed Apr. 13, 2022).

[29] Computer Hope, "What is a Game?" https://www.computerhope.com/jargon/g/game.htm (accessed Apr. 13, 2022).

[30] R. D. Halliburton and J. Pearson, "Ronald D. Halliburton Inventions, Patents and Patent Applications - Justia Patents Search." https://patents.justia.com/inventor/ronald-d-halliburton (accessed Apr. 13, 2022).

[31] V. Huard Pelletier, A. Lessard, F. Piché, C. Tétreau, and M. Descarreaux, "Video games and their associations with physical health: a scoping review," *BMJ Open Sport Exerc. Med.*, vol. 6, no. 1, p. e000832, Oct. 2020, doi: 10.1136/bmjsem-2020-000832.

[32] Blizzard, "Warcraft: Orcs and Humans on GOG.com." https://www.gog.com/game/warcraft_orcs_and_humans (accessed Apr. 13, 2022).

[33] M. A. Winget, "Videogame preservation and massively multiplayer online role-playing games: A review of the literature," *J. Am. Soc. Inf. Sci. Technol.*, vol. 62, no. 10, pp. 1869–1883, Oct. 2011, doi: 10.1002/asi.21530.

[34] T. H. Apperley, "Genre and game studies: Toward a critical approach to video game genres," *Simul. Gaming*, vol. 37, no. 1, pp. 6–23, Mar. 2006, doi: 10.1177/1046878105282278.

[35] J. H. Lee, N. Karlova, R. I. Clarke, K. Thornton, and A. Perti, "Facet Analysis of Video Game Genres," presented at the iConference 2014 Proceedings: Breaking Down Walls. Culture - Context - Computing, Mar. 2014. doi: 10.9776/14057.

[36] D. Carr, D. Buckingham, A. Burn, and G. Schott, *Computer Games: Text, Narrative and Play*. Polity, 2006.

[37] T. Krzywinska and D. Brown, "Online Games and Genre," in *The International Encyclopedia of Digital Communication and Society*, 1st ed., P. H. Ang and R. Mansell, Eds. Wiley, 2015, pp. 1–4. doi: 10.1002/9781118767771.wbiedcs043.

[38] A. Tychsen, M. Hitchens, T. Brolund, and M. Kavakli, "Live Action Role-Playing Games: Control, Communication, Storytelling, and MMORPG Similarities," *Games Cult.*, vol. 1, no. 3, pp. 252–275, Jul. 2006, doi: 10.1177/1555412006290445.

[39] T. Aichner, M. Grünfelder, O. Maurer, and D. Jegeni, "Twenty-five years of social media: a review of social media applications and definitions from 1994 to 2019," *Cyberpsychology Behav. Soc. Netw.*, vol. 24, no. 4, pp. 215–222, 2021.

[40] C. T. Carr and R. A. Hayes, “Social media: Defining, developing, and divining,” *Atl. J. Commun.*, vol. 23, no. 1, pp. 46–65, 2015.

[41] danah m. boyd and N. B. Ellison, “Social Network Sites: Definition, History, and Scholarship,” *J. Comput.-Mediat. Commun.*, vol. 13, no. 1, pp. 210–230, Oct. 2007, doi: 10.1111/j.1083-6101.2007.00393.x.

[42] G. M. H. C. Gajanayake and T. C. Sandanayake, “Trending Pattern Identification of YouTube Gaming Channels Using Sentiment Analysis,” in *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*, Colombo, Sri Lanka, Nov. 2020, pp. 149–154. doi: 10.1109/ICTer51097.2020.9325476.

[43] C. Krittanawong *et al.*, “Association of Social Gaming with Well-Being (Escape COVID-19): A Sentiment Analysis,” *Am. J. Med.*, vol. 135, no. 2, pp. 254–257, Feb. 2022, doi: 10.1016/j.amjmed.2021.10.010.

[44] A. Ghosh, “Analyzing Toxicity in Online Gaming Communities,” *Turk. J. Comput. Math. Educ. TURCOMAT*, vol. 12, no. 10, Art. no. 10, Apr. 2021, doi: 10.17762/turcomat.v12i10.5182.

[45] M. Maloney, S. Roberts, and T. Graham, *Gender, masculinity and video gaming: analysing Reddit’s r/gaming community*. Cham, Switzerland: Palgrave Macmillan, 2019.

[46] “Reddit,” *Vikipedi*. Apr. 03, 2022. Accessed: Apr. 13, 2022. [Online]. Available: https://tr.wikipedia.org/w/index.php?title=Reddit&oldid=27500499

[47] *PRAW: The Python Reddit API Wrapper*. Python Reddit API Wrapper Development, 2022. Accessed: Apr. 13, 2022. [Online]. Available: https://github.com/praw-dev/praw

[48] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python*, 1st ed. Beijing ; Cambridge [Mass.]: O’Reilly, 2009.

[49] T. Hoobyar, T. Dotz, and S. Sanders, *NLP: the essential guide to neuro-linguistic programming*, 1st ed. New York: William Morrow, 2013.

[50] B. Alshemali and J. Kalita, “Improving the Reliability of Deep Neural Networks in NLP: A Review,” *Knowl.-Based Syst.*, vol. 191, p. 105210, Mar. 2020, doi: 10.1016/j.knosys.2019.105210.

[51] K. Schouten, F. Frasincar, and F. de Jong, “Ontology-Enhanced Aspect-Based Sentiment Analysis,” in *Web Engineering*, vol. 10360, J. Cabot, R. De Virgilio, and R. Torlone, Eds. Cham: Springer International Publishing, 2017, pp. 302–320. doi: 10.1007/978-3-319-60131-1_17.

[52] C. J. Hutto and E. Gilbert, “VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text,” p. 10.

[53] P. Willett, “The Porter stemming algorithm: then and now,” *Program*, vol. 40, no. 3, pp. 219–223, Jul. 2006, doi: 10.1108/00330330610681295.

[54] “Snowball: A language for stemming algorithms.” http://snowball.tartarus.org/texts/introduction.html (accessed Apr. 13, 2022).

[55] “NLTK :: nltk.stem.lancaster.” https://www.nltk.org/_modules/nltk/stem/lancaster.html (accessed Apr. 13, 2022).

[56] Joe Seabreeze, “Why do people play MMORPGs?,” *Quora*. https://www.quora.com/Why-do-people-play-MMORPGs (accessed Apr. 13, 2022).

[57] dottiris, “Attraction for WoW!,” *r/wow*, Apr. 09, 2022. www.reddit.com/r/wow/comments/u02j1k/attraction_for_wow/ (accessed Apr. 13, 2022).

[58] C. S. Ang and P. Zaphiris, “SOCIAL ROLES OF PLAYERS IN MMORPG GUILDS: A social network analytic perspective,” *Inf. Commun. Soc.*, vol. 13, no. 4, pp. 592–614, Jun. 2010

Journal of Emerging Computer Technologies
Research Article
Received: *2022-06-01* | ***Reviewing:*** *2022-06-02* & *2022-07-18* | ***Accepted:*** *2022-07-18* | ***Online:*** *2022-07-19* | ***Issue Date:*** *2022-06-30*

Symptom Based COVID-19 Prediction Using Machine Learning and Deep Learning Algorithms

Nesibe Yalçın
Department of Computer Engineering
Erciyes University
Kayseri, Türkiye
nesibeyalcin@erciyes.edu.tr
0000-0003-0324-9111

Sibel Ünaldı
Department of Electrical Electronics Engineering
Bilecik Şeyh Edebali University
Bilecik, Türkiye
sibel.unaldi@bilecik.edu.tr
0000-0001-9948-4284

***Abstract*— Research studies are carried out in many areas of science to cope with the impacts of the COVID-19 crisis in the world. Machine learning can be used for purposes such as understanding, addressing, fighting, and preventing - controlling COVID-19. In this research, the presence of COVID-19 has been predicted using K Nearest Neighbor, Support Vector Machines, Logistic Regression, and Multilayer Perceptual Neural Networks machine learning and Gated Recurrent Unit (GRU) and Long Short-Term Memory deep learning algorithms. A publicly available dataset that includes various features (i.e. wearing masks, abroad travel, contact with the COVID patient) and symptoms (i.e. breathing problems, fever, and dry cough) is used for the COVID-19 diagnosis prediction. The learning algorithms have been compared according to the evaluation metrics. The experimental results have been shown that GRU deep learning algorithm is more reliable with a prediction accuracy of 98.65% and a loss/mean squared error of 0.0126.**

Keywords— COVID-19, deep learning, symptom, machine learning, prediction

I. Introduction

COrona **VI**rus **D**isease 2019 (COVID-19) epidemic caused by SARS-CoV-2 still brings many problems globally. This epidemic, which emerged in Wuhan, Hubei Province of China in 2019 December, has spread worldwide rapidly. A novel coronavirus, whose symptoms may include dry cough, fever, and anosmia, was identified on 7 January 2020 [1, 2]. The World Health Organization (WHO) announced a pandemic on 11 March 2020 [3]. In October 2020, the total number of patients exceeded 39,500,000 [4]. According to WHO [5], as of 7 January 2022, a total of 298,915,721 COVID-19 cases and 5,469,303 COVID-19 deaths were announced globally, while the number of COVID-19 cases in the last 7 days was 13,307,762 and the number of COVID-19 deaths was 40,868. This pandemic emphasizes the ability of viral spread from animals to cause significant disease in humans [6].

Machine learning, a subfield of artificial intelligence, is a method that enables machines to produce new solutions based on previous solutions [7]. Machine learning can play an important role in research and predictions of COVID-19 or other diseases. It can be used to analyze, evaluate and triage COVID-19 cases by integrating into health provider programs and strategies [8]. Machine learning based applications/platforms show a huge potential for accelerating COVID-19 diagnosis and treatment [8-10]. It can be interpreted that the machine learning methods will be useful in improving the diagnostic accuracy by using it together with Polymerase Chain Reaction (PCR) test or other tests [11]. Prediction of diagnosis according to symptoms in pandemics is of great importance in terms of both initiating treatments with early diagnosis and creating highly accurate alternatives that can alleviate the workload of healthcare professionals.

In the literature, many studies are aiming to obtain faster and more accurate results for COVID-19 diagnosis, including machine learning and deep learning based on artificial intelligence principles [12-14]. In addition, Computed Tomography and X-ray medical images are used to accurately segment infected parts with artificial intelligence to increase the efficiency of COVID-19 diagnosis [15, 16]. In [8], it is aimed to figure out the role of machine learning algorithms in different studies dealing with COVID-19. Supervised learning algorithms have presented better results with 92.9% test accuracy compared to unsupervised learning algorithms. Reference [17] have studied on COVID-19 dataset and developed a model utilizing the Support Vector Machine (SVM) to estimate patients as COVID or not. An accuracy of 87% has been achieved in estimating 3 cases: not infected, mildly infected, and severely infected. In [18], Prophet, Random Forest, AutoRegressive Integrated Moving Average (ARIMA), Polynomial Regression, and Linear Regression models have been built up to detect COVID-19 confirmed cases in the USA using machine learning algorithms and Polynomial Regression has outweighed the other algorithms by giving the best estimates. Reference [19] has created a model for the future COVID-19 forecast of 7 countries (including Turkey) considering the number of cases. In addition to classical forecasting methods, machine learning methods have been implemented to a COVID-19 dataset and Facebook's Prophet method has given the lowest forecasting error for all countries. Various supervised machine learning algorithms have been applied to estimate COVID-19 in [20] worked on a COVID-19 dataset. The algorithms' performance has been evaluated using 10-fold cross validation, and after comparing all experiments, the highest accuracy rate has been obtained by SVM with 98.81%. Reference [21] have applied Naïve Bayes, Logistic Regression (LR), SVM, Decision Tree, and K Nearest Neighbors (KNN) machine learning algorithms for the determination of COVID-19 patients and have studied on a worldwide accessible dataset. In this prediction study based on their symptoms, Naïve Bayes and Decision Tree having an accuracy of 93.70% present the best performance. Reference [22] has proposed forecast models including Long Short - Term Memory (LSTM), bidirectional LSTM, ARIMA, and support vector regression for COVID-19 prediction. Bidirectional LSTM outperforms better for pandemic

Cite (APA): Yalçın, N., Ünaldı, S. (2022). Symptom Based COVID-19 Prediction Using Machine Learning and Deep Learning Algorithms. *Journal of Emerging Computer Technologies*, 2(1), 22-29.

prediction in planning and management in the public health system. Recent studies on the detection of COVID-19 are summarized in Table I.

In the study on human infection caused by 2019 novel coronavirus (2019-nCoV) [27], treatment and clinical features, and epidemiological, radiological, and laboratory characteristics of the 2019-nCoV infected patients have been reported. To clarify the clinical and epidemiological characteristics of 2019-nCoV, [28] also analyzed demographic, epidemiological, radiological, and clinical features and laboratory data of coronavirus patients. A patient with no history of diabetes, hepatitis, or tuberculosis is studied in [6]. This patient was admitted to the hospital 6 days after the onset of coronavirus disease and reported fever, dizziness, cough, and serious respiratory syndrome at presentation. The dataset used in [4] includes 8 basic features: demographic information (gender and age 60+), clinical symptoms (cough, fever, sore throat, shortness of breath, headache), and known contact with a confirmed COVID patient.

TABLE I. SUMMARY OF LITERATURE REVIEW ON COVID-19 DETECTION

Reference	Month - Year	Dataset	Description	Methods	Results
[4]	January 2021	Records from tested individuals	Prediction of COVID-19 diagnosis based on symptoms	Gradient Boosting	0.90area under a receiver operating characteristic (AUROC), 0.66 are under precision-recall curve (AUPRC)
[13]	February 2022	Laboratory blood tests	A novel Deep Neural Network (DNN) modelfor early COVID-19 diagnosis	LR, KNN, Decision Tree, Extremely Randomized Trees, SVM, Naïve Bayes, Random Forest, XGBoost, LSTM, DNN, Recurrent Neural Network (RNN), and Convolutional Neural network (CNN)	The proposed DNN model achieved an accuracy of 93.33%.
[14]	January 2022	RT-PCRvirology test results	A deep learning model to improve COVID-19 diagnostic performance	LSTM	The model exceeded a sensitivity of 90%
[15]	October 2021	X-ray and CT-scan medical images	Deep learning methods applied to medical images for COVID-19 detection	CNN models, VGG16, DenseNet121, ResNet50, ResNet152, and Fast.AI ResNet	High accuracy of 99%
[17]	May 2021	Extracted critical symptoms	Detection of COVID-19 from the symptoms	SVM	An accuracy of 87%
[18]	January 2021	Confirmed cases	Prediction of COVID-19 cases	Random Forest, Polynomial Regression, Linear Regression, Prophet, and ARIMA	A mean absolute error (MAE) of 1.86%
[19]	May 2020	Confirmed cases from 7 countries	Prediction of possible confirmed cases and mortality numbers	SVM, Holt-Winters, Prophet, and LSTM	The prophet model presented the lowest RMSE for all countries.
[20]	May 2021	Possible factors	Detection of COVID-19 presence	SVM, KNN, J48 Decision Tree, Random Forest, and Naïve Bayes	An accuracy of 98.81%
[21]	May 2021	Patient recordscontaining symptoms and actual results	Determination of COVID-19 patients among various age groups	SVM, KNN, Decision Tree, LR, and Naïve Bayes	An accuracy of 93.70%
[22]	August 2020	Confirmed and recovered cases	Prediction for COVID-19	LSTM, GRU, and Bidirectional-LSTM	MAE value of 0.007 and R^2 value of 0.9997
[23]	July 2020	Laboratory findings	Clinical predictive modelsto estimate COVID-19 infection	Artificial Neural Network (ANN), CNN, RNN, LSTM, SNNLLSTM, and CNNRNN	Deep learning models have an accuracy of over 84%.
[24]	January 2020	Clinical features	Prediction of COVID-19 mortality	MLPNN, KNN, J48 Decision Tree, Random Forest, Naïve Bayes, LR, and XGBoost	An accuracy of 95.03%
[25]	October 2020	Multiple features of patients	Prediction of COVID-19 risk	LR	An accuracy of 92%.
[26]	December 2021	Clinical symptoms	Symptom based prediction model for diagnosis of COVID-19 in children	Random forest, LR, MLP, SVM, Boosted Trees	Area under ROC of 0.65
This study	-	Symptoms and various features	Prediction of COVID-19 presence	KNN, SVM, LR, MLPNN, GRU and LSTM	Prediction accuracy of 98.65%, AUROC of 0.989 and AUPRC of 0.998 with 95% confidence interval (CI)

This research aims to analyze and estimate the COVID-19 presence based on the symptoms & features. For this purpose, KNN, LR, SVM, Multilayer Perceptron Neural Network (MLPNN) machine learning algorithms, and LSTM and Gated Recurrent Unit (GRU) deep learning algorithms have been used. GRU has been announced to be an appropriate algorithm for COVID-19 diagnosis prediction due to its better accuracy. The following section details the COVID-19 dataset used in the research. Section III provides the information about the preparation of the study and briefly explains the used algorithms and evaluation metrics. In Section IV, the experimental results are given and evaluated. The final section concludes and provides suggestions for future study.

II. COVID-19 DATASET

In this study, it is used a publicly available dataset entitled "Symptoms and COVID Presence" from Kaggle [29]. The dataset is updated on 2020-08-18. It covers data between 2020-04-17 and 2020-08-29. It contains 20 features that indicate the presence of various symptoms and 1 class feature (the person has COVID or not). The total number of examples in the dataset is 5434, the number of COVID patients is 4383 (80.7%), and the number of healthy people is 1051 (19.3%), as shown in Fig. I. The presence of COVID in potential patients is indicated as "Yes" or "No".

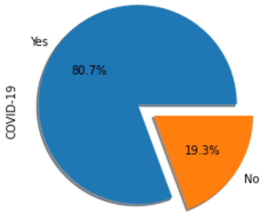


FIG. I. THE PRESENCE OF THE COVID-19

Different people are affected by COVID-19 in various ways. Infected patients develop various levels of symptoms. In addition to several symptoms of COVID-19 infection such as shortness of breath, fever, and dry cough, some infected people have experienced fatigue, anosmia (loss of taste or smell), and muscle aches [30]. The correlation between each feature/symptom and coronavirus disease has been computed and the obtained correlation coefficients have been listed in Table II. In addition, the number and rates of COVID-19 positive and negative cases are given according to the status of each feature/symptom.

TABLE II. THE BASIC STATISTICS OF THE COVID-19 DATASET

Feature & Symptom	Correlation Coefficient	Status	Total (n=5,434)		COVID-19 Positive (n=4,383)		COVID-19 Negative (n=1,051)	
			n	*%*	*n*	*%*	*n*	*%*
Breathing problem	0.444	Yes	3,620	66.6	3,369	76.9	251	23.9
		No	1,814	33.4	1,014	23.1	800	76.1
Fever	0.353	Yes	4,273	78.6	3,757	85.7	516	49.1
		No	1,161	21.4	626	14.3	535	50.9
Dry cough	0.464	Yes	4,307	79.3	3,878	88.5	429	40.8
		No	1,127	20.7	505	11.5	622	59.2
Sore throat	0.503	Yes	3,953	72.7	3,669	83.7	284	27
		No	1,481	27.3	714	16.3	767	73
Running nose	-0.006	Yes	2,952	54.3	2,375	54.2	577	54.9
		No	2,482	45.7	2,008	45.8	474	45.1
Asthma	0.09	Yes	2,514	46.3	2,124	48.5	390	37.1
		No	2,920	53.7	2,259	51,5	661	62.9
Chronic lung disease	-0.057	Yes	2,565	47.2	2,008	45.8	557	53
		No	2,869	52.8	2,375	54.1	494	47
Headache	-0.028	Yes	2,736	50.3	2,177	49.7	559	53.2
		No	2,698	49.7	2,206	50.3	492	46.8
Heart disease	0.027	Yes	2,523	46.4	2,064	47.1	459	43.7
		No	2,911	53.6	2,319	52.9	592	56.3
Diabetes	0.041	Yes	2,588	47.6	2,131	48.6	457	43.5
		No	2,846	52.4	2,252	51.4	594	56.5
Hypertension	0.103	Yes	2,663	49	2,258	51.5	405	38.5
		No	2,771	51	2,125	48.5	646	61.5
Fatigue	-0.044	Yes	2,821	51.9	2,228	50.8	593	56.4
		No	2,613	48.1	2,155	49.2	458	43.6
Gastrointestinal	-0.003	Yes	2,551	46.9	2,054	46.9	497	47.3
		No	2,883	53.1	2,329	53.1	554	52.7
Abroad travel	0.444	Yes	2,451	45.1	2,451	55.9	0	0
		No	2,983	54.9	1,932	44.1	1,051	100
Contact with COVID patient	0.357	Yes	2,726	50.2	2,582	58.9	144	13.7
		No	2,708	49.8	1,801	41.1	907	86.3
Attended large gathering	0.39	Yes	2,510	46.2	2,442	55.7	68	6.5
		No	2,924	53.8	1,941	44.3	983	93.5
Visited public exposed places	0.12	Yes	2,820	51.9	2,403	54.8	417	39.7
		No	2,614	48.1	1,980	45.2	634	60.3
Family working in public exposed places	0.16	Yes	2,262	41.6	1,994	45.5	268	25.5
		No	3,172	58.4	2,389	54.5	783	74.5
Wearing masks	-	Yes	0	0	0	0	0	0
		No	5,434	100	4,383	100	1,051	100
Sanitization from market	-	Yes	0	0	0	0	0	0
		No	5,434	100	4,383	100	1,051	100

According to Table II, there is a strong positive relationship between COVID-19 presence and sore throat, dry cough, and breathing problems. A sore throat was observed in 83.7% of those infected with the coronavirus. 88.5% and 85.7% of patients with COVID-19 have dry cough and fever complaints, respectively. The percentage of infections in people who have had close contact with COVID-19 patients is 94.71%. People who have recently traveled abroad and about 80.7% of people not wearing masks are infected.

III. METHODS

This research presents machine learning / deep learning based models to detect COVID-19 diagnosis. The deep learning and machine learning algorithms involved in the research have been built with Python programming language and the Google Colab platform (a free online cloud-based product from Google Research) has been selected to execute the algorithms. Popular Python libraries such as NumPy, Pandas, Keras, Scikit-Learn, and Matplotlib have been used to develop the models. KNN, LR, SVM, and MLPNN supervised machine learning algorithms, and LSTM and GRU deep learning algorithms have been selected for this study to construct the prediction model. The models have been trained on the training dataset using the learning algorithms and then the trained models have been tested with the testing dataset. The architecture of the proposed prediction model is illustrated in Fig. II.

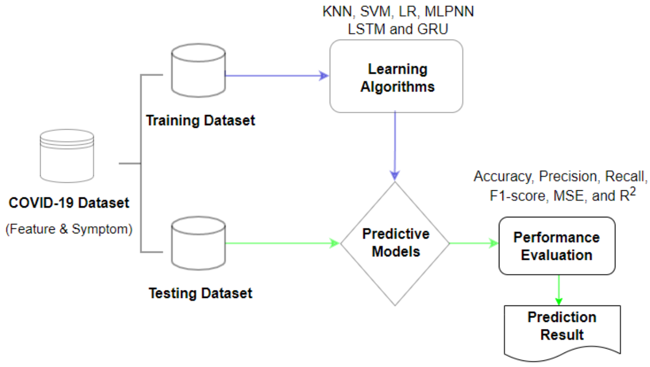


FIG. II. GRAPHICAL REPRESENTATION OF THE COVID-19 PREDICTION MODEL

KNN is the oldest classification algorithm and has some advantages simplicity in terms of complexity and quick calculation time [31]. The number of neighbors, k is a hyperparameter for building the prediction model in KNN. SVM is one of the widely preferred machine learning algorithms for classification or prediction problems. It uses the maximum margin concept and converts low-dimensional input space to higher dimensional space to create separable classes, depending on kernel functions [17]. MLPNN is a feed-forward neural network model using a backpropagation algorithm for training [32]. It consists of input, hidden (at least), and output layers and aims to minimize the difference between the target (desired output) and the output of the network [33]. LR, despite its name, is a linear model for predicting classes rather than regression. It is also a commonly studied simple machine learning algorithm for binary classification. It describes the relationship between at least one independent variable and a categorical dependent variable [25]. Recurrent Neural Networks (RNNs) are an extension of feedforward neural networks. LSTM and GRU networks are popular RNN architectures. LSTM is structurally composed of 3 main gates, namely forget gate, input gate, and output gate [22]. It takes into account crucial lessons acquired from previous experiences [19], unlike conventional neural networks. GRU is a gating mechanism in RNN and is less complex than LSTM. GRU has reset gate and update gate (combination of input gate and forget gate) [34].

Various metrics have been used to describe and evaluate the performance of each algorithm. Confusion matrix is often used to describe and illustrate the performance of the prediction/classification methods. As shown in Table III, it presents a summary table about the number of incorrect and correct predictions. True Negative (TN) and True Positive (TP) are the total numbers of correctly predicted negative and positive examples, respectively. False Negative (FN) and False Positive (FP) are incorrect predictions. False Negative (FN) and False Positive (FP) are the total numbers of incorrectly predicted positive and negative examples, respectively.

TABLE III. THE STRUCTURE OF THE CONFUSION MATRIX

Class		**Predicted**	
		Negative	***Positive***
Actual	***Negative***	TN	FP
	Positive	FN	TP

Accuracy, a common evaluation metric, is the ratio of accurate predictions (TP and TN) over all (correctly and incorrectly) predictions. As can be seen in (1), it is computed depending on the confusion matrix.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \qquad (1)$$

Precision, recall, and F1-score performance metrics have been also used to determine the algorithm(s) making the most accurate predictions and can be defined by (2), (3), and (4), respectively. Precision is the ratio of correct positive predictions to all positive predictions. Recall is the ratio of positive predictions to the total positive examples. F1-score measures the harmony and the balance of precision and recall metrics.

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \qquad (3)$$

$$\text{F1} - \text{score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \qquad (4)$$

Mean Square Error (MSE) has been used as a loss function for computing the loss between the real values and predictions. It is defined by (5).

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (t_i - y_i)^2 \qquad (5)$$

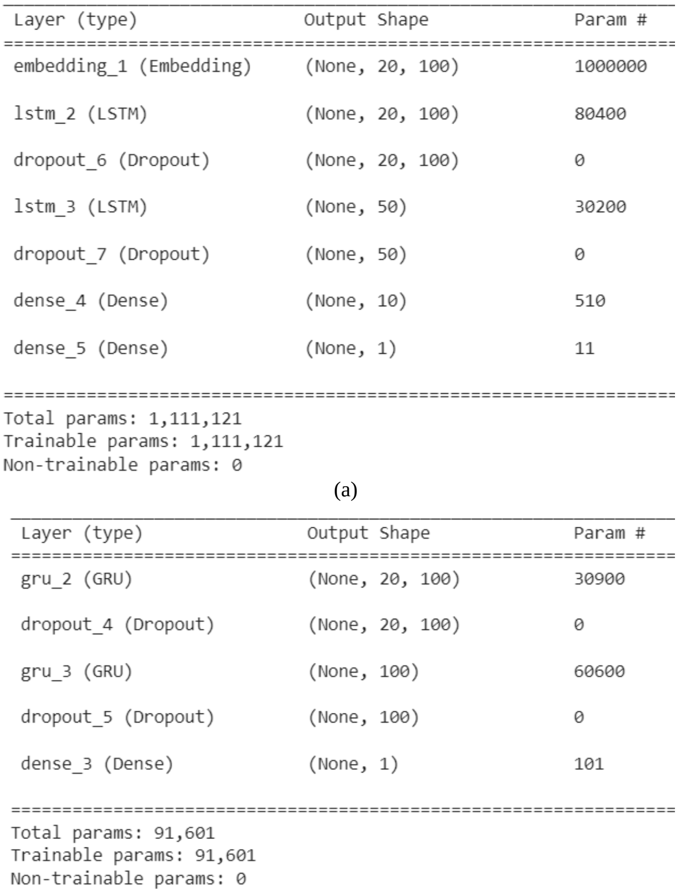
where$n$ is the total number of examples in the dataset and i is the index (i = 1, 2, 3, …, n). $t_i$is the target (actual, desired) output value, and y_i is the predicted output using the learning algorithm for i. example. Explanatory coefficient R^2 is computed by (6) and t_{ort} is the average of the target output values.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(t_i - y_i)^2}{\sum_{i=1}^{n}(t_i - t_{ort})^2} \qquad (6)$$

The dataset has been divided into two subsets (85% - 15%): a training dataset and a testing dataset. There are 5,434 examples in the dataset, the training dataset includes examples of 4,618 (3,716 COVID and 902 healthy) and the testing dataset examples of 816 (667 COVID and 149 healthy). Then the dataset has been passed on to the learning algorithms.

IV. Results and Discussion

The development of all the models has been performed under the Google Colab environment. In order to decide the optimal value of k in KNN, error rates have been calculated for all k neighbor numbers between 1 and 40, and the k value corresponding to the lowest error rate has been determined as 3. Euclidean distance is used as the distance function in KNN. As a result of experiments, the polynomial kernel is selected as the kernel function in SVM. Different neural network models (i.e., a maximum number of hidden layers/neurons and epochs, activation function) have been designed to determine the best network structure. After extensive experiments, the number of hidden layers is 2 and the numbers of hidden neurons are 64 and 32, respectively. The rectified linear unit (relu) activation function has been utilized for the hidden layers in MLPNN. The learning rate for weight updates between layers is a constant of 0.001. The summaries of the developed LSTM and GRU model architectures are depicted in Fig. III. The output shape and number of parameters in each layer can be seen clearly.



```
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, 20, 100)           1000000

lstm_2 (LSTM)                (None, 20, 100)           80400

dropout_6 (Dropout)          (None, 20, 100)           0

lstm_3 (LSTM)                (None, 50)                30200

dropout_7 (Dropout)          (None, 50)                0

dense_4 (Dense)              (None, 10)                510

dense_5 (Dense)              (None, 1)                 11

=================================================================
Total params: 1,111,121
Trainable params: 1,111,121
Non-trainable params: 0
```

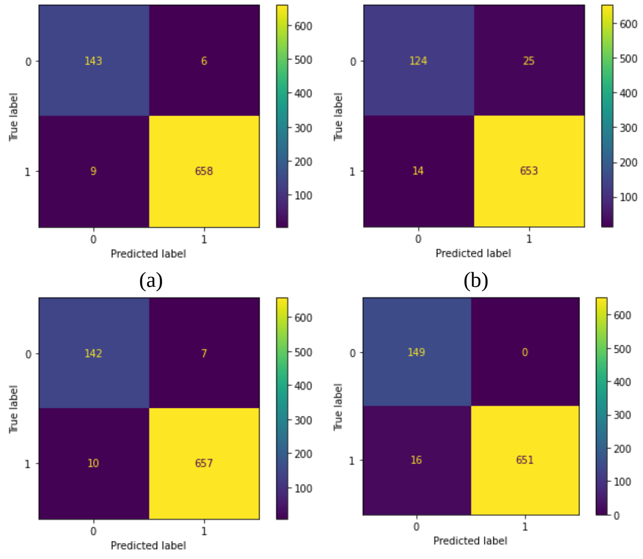
(a)

```
Layer (type)                 Output Shape              Param #
=================================================================
gru_2 (GRU)                  (None, 20, 100)           30900

dropout_4 (Dropout)          (None, 20, 100)           0

gru_3 (GRU)                  (None, 100)               60600

dropout_5 (Dropout)          (None, 100)               0

dense_3 (Dense)              (None, 1)                 101

=================================================================
Total params: 91,601
Trainable params: 91,601
Non-trainable params: 0
```

(b)

Fig. III. The developed (a) LSTM and (b) GRU models

Fig. IV provides a complete inside into the test results obtained after applying machine learning algorithms in the diagnosis prediction. KNN and MLPNN are more satisfactory in terms of the TP and TN values, respectively. It has been observed that KNN classifies the infected people less incorrectly and LR predicts a higher number of healthy people as ill in comparison to other algorithms.

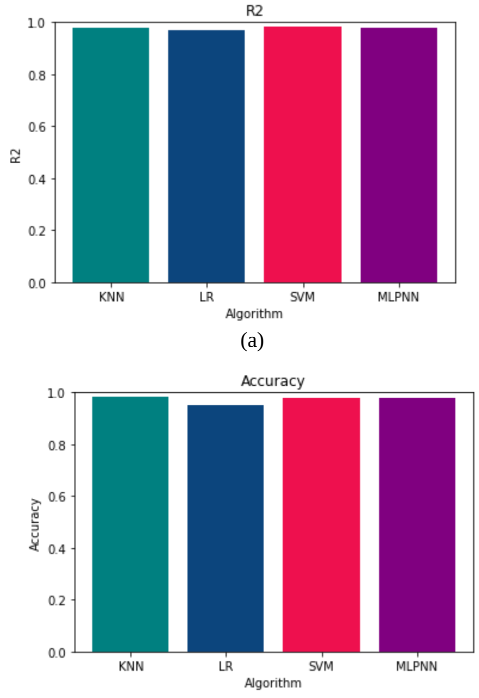


(a) (b)

(c) (d)

Fig. IV. The confusion matrices obtained after training on the dataset using (a) KNN, (b) LR, (c) SVM, and (d) MLPNN algorithms

The direct comparison of the studied machine learning algorithms' performance for the COVID-19 prediction is presented in Fig. V. The R^2 obtained for the training dataset and testing accuracy values show a similar distribution.

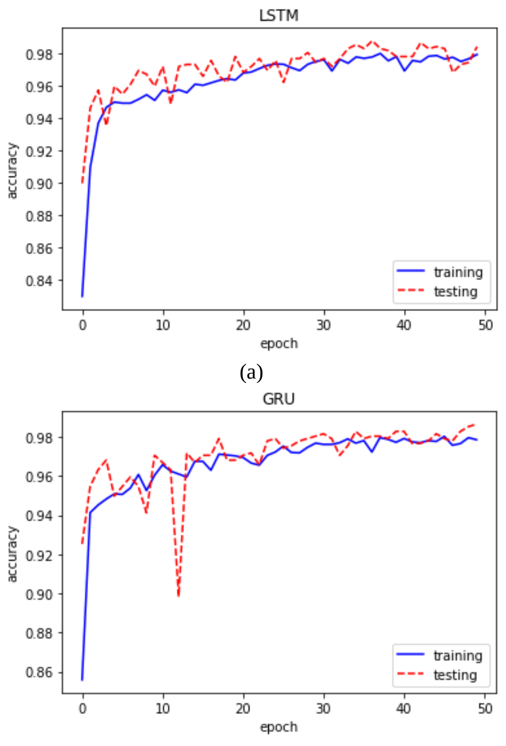


(a)

(b)

Fig. V. The (a) R^2 and (b) accuracy graphics of the machine learning algorithms

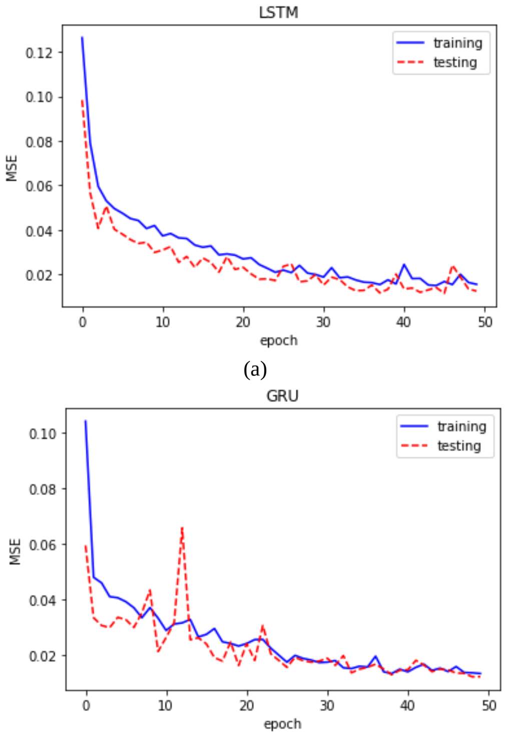
Fig. VI and Fig. VII provide comparisons between LSTM and GRU deep learning algorithms in terms of the most important performance metrics “accuracy” and “MSE” for training and testing datasets. From the overall comparison, it can be observed that both LSTM and GRU have prediction accuracy of more than 98% and also offer acceptable performance with a loss/MSE of about 1% in both training and testing.



(a)

(b)

FIG. VI. THE OBTAINED ACCURACY RESULTS USING (a) LSTM AND (b) GRU ALGORITHMS



(a)

(b)

FIG. VII. THE OBTAINED LOSS GRAPHICS USING (a) LSTM AND (b) GRU ALGORITHMS

The obtained test results using the different performance metrics are presented comparatively in Table IV to give an idea about the prediction success of the symptom-based COVID-19 diagnosis for each algorithm. All of the used algorithms produce promising results with an accuracy of above 95%. The results show that GRU having a 98.65% accuracy is the best learning algorithm. That is to say, it has the highest level of accuracy when compared with KNN, LR, SVM, MLPNN, and LSTM.

TABLE IV THE SUCCESS OF THE ALGORITHMS

Algorithm	**Accuracy (%)**	**Precision**	**Recall**	**F1-score**	**Loss (MSE)**
KNN	98.16	0.991	0.987	0.989	0.0184
LR	95.22	0.963	0.979	0.971	0.0478
SVM	97.92	0.989	0.985	0.987	0.0208
MLPNN	98.04	1.000	0.976	0.988	0.0196
LSTM	98.41	1.000	0.98	0.99	0.0125
GRU	**98.65**	0.997	0.986	0.992	0.0126

TABLE V. COMPARISON WITH DIFFERENT LEARNING ALGORITHMS ON THE SAME DATASET

Reference	**Algorithms**	**Best Algorithm**	**Highest accuracy (%)**
[20]	J48 Decision Tree, Naïve Bayes, SVM, KNN, Random Forest	SVM	98.81
This study	KNN, LR, SVM, MLPNN	KNN	98.16
This study	LSTM, GRU	GRU	98.65

Villavicencio et al. [20] has also used the same dataset [29] for predicting the COVID-19 infected patients. Table V summarizes the performance of the studies on the same dataset with different algorithms in terms of accuracy rates. It can be observed that LSTM and GRU are provided 98.65% and 98.41% accuracy success for COVID-19 prediction, respectively. The SVM algorithm has an accuracy of 98.81% and is the best method reported in [20] for the detection of the potential presence of COVID-19. GRU prediction model have run 50 epochs and 30 times. According to the results of the best run, the model has predicted with the AUROC of 0.97-0.989 and AUPRC of 0.993 and 0.998 with 95% CI: 97.9% - 98.7% accuracy, 98.0% - 98.6% recall (sensitivity), and 95.5% - 99.4% specificity.

The accuracy, precision, and recall values computed for common machine learning algorithms (SVM and KNN) used in this study and [20] are compared comprehensively in Table VI. The results show that SVM in [20] and KNN in our study have better performance than the other algorithms used in each study. The better values are produced in our study when the same parameters are used for SVM and KNN algorithms. This study is achieved a higher prediction accuracy than the study of [20] when the polynomial kernel function is used for SVM and the number of nearest neighbors, k is selected as 3, and cross-validation is not performed for KNN.

TABLE VI. A COMPREHENSIVE COMPARISON OF THE STUDIES USED THE SAME DATASETS

Reference	Algorithm	Accuracy (%)	Precision	Recall
[20]	KNN (k = 1,10-fold cross-validation)	**98.69**	0.987	0.987
[20]	KNN (k = 3, no cross validation)	97.57	-	-
This study	KNN (k = 3, no cross validation)	**98.16**	0.987	**0.989**
[20]	SVM (Pearson VII universal kernel, 10-fold cross-validation)	**98.81**	0.988	**0.988**
[20]	SVM (Polynomial kernel, 10-fold cross-validation)	95.48	-	-
This study	SVM (Polynomial kernel no cross-validation)	95.48	-	-
This study	SVM (Polynomial kernel, no cross-validation)	**97.92**	**0.989**	0.985

V. CONCLUSION

This research aims to analyze and estimate the diagnosis of COVID-19 based on COVID-19 symptoms using several machine learning and deep learning algorithms. KNN, LR, SVM, MLPNN, GRU, and LSTM algorithms have been used for the COVID-19 diagnostic estimation. This research has been carried out on a worldwide available COVID-19 database for the diagnosis of this viral disease and has shown promising results with high accuracy, precision, recall, F1-score, and MSE. The best performance has been obtained by GRU (98.65% accuracy) and the lowest accuracy by LR (95.22%). The second-best results in terms of prediction success and error rate have been presented by LSTM.

The results demonstrate the capability of machine learning and deep learning algorithms in predicting COVID-19. The study shows that using the learning algorithms together with other tests such as PCR can be a good alternative in terms of increasing the diagnostic accuracy. In the future, this study can be extended to address COVID-19 variants in the COVID-19 health care applications.

AUTHORS' CONTRIBUTIONS

Nesibe YALÇIN: Conceptualization, Methodology, Software, Validation, Investigation, Resources, Performing the algorithms, Evaluating the results with performance metrics, and Writing the manuscript.

Sibel ÜNALDI: Conceptualization, Validation, Investigation, Resources, Analyzing the results, and Writing the manuscript

CONFLICT OF INTEREST

There is no conflict of interest in this study.

REFERENCES

[1] N.Alballa and I. Al-Turaiki, "Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: A review", Inform. Med. Unlocked, vol. 24, pp. 100564 (1-17), 2021.

[2] C. I.Paules,H. D. Marston, and A. S.Fauci, "Coronavirus infections - more than just the common cold", JAMA: J. Am. Med. Assoc., vol. 323, pp. 707-708, 2020.

[3] WHO, "Virtual press conference on COVID-19 - 11 March 2020", 25 January 2022, Available online: https://www.who.int/docs/default-source/coronaviruse/transcripts/who -audio-emergencies-coronavirus-press-conference-full-and-final-11mar2020.pdf, 2020.

[4] Y.Zoabi, S.Deri-Rozov, andN.Shomron, "Machine learning-based prediction of COVID-19 diagnosis based on symptoms", NPJ Digit. Med., vol. 4, pp. 3 (1-5), 2021.

[5] WHO, "WHO coronavirus disease (COVID-19) dashboard", 13 January 2022, Available online: https://covid19.who.int/, 2022.

[6] F. Wu et al., "A new coronavirus associated with human respiratory disease in China", Nature, vol. 579 (7798), pp. 265-269, 2020.

[7] O. Sevli andV. G.Başer, "COVID-19 salgınına yönelik zaman serisi verileri ile Prophet model kullanarak makine öğrenmesi temelli vaka tahminlemesi", European Journal of Science and Technology, vol. 19, pp. 827-835, 2020.

[8] A. S. Kwekha-Rashid, H. N. Abduljabbar, andB. Alhayani, "Coronavirus disease (COVID-19) cases analysis using machine-learning applications", Appl. Nanosci., pp. 1-13, 2021.

[9] M.Naseem, R.Akhund, H.Arshad, andM. T. Ibrahim, "Exploring the potential of artificial intelligence and machine learning to combat COVID-19 and existing opportunities for LMIC: a Scoping review", J. Prim. Care Community Health, vol. 11, pp. 1-11, 2020.

[10] Jamshidi M. et al., "Artificial intelligence and COVID-19: deep learning approaches for diagnosis and treatment", IEEE Access, vol. 8, pp. 109581–109595, 2020.

[11] E.Dinçmen, "Makine öğrenmesi ve Covid-19", 20 January 2022, Available online: https://www.isikun.edu.tr/web/1695-15661-1-1/isik_universitesi/hakkinda/yonetim__ idari_birimler__kurumsal_iletisim_daire_baskanligi__basinda_isik_universitesi__isik_yazilari/makine_ogrenmesi_ve_covid-19 (2022)

[12] Z. A. A. Alyasseri et al., "Review on COVID-19 diagnosis models based on machine learning and deep learning approaches", Expert Syst., pp. e12759 (1-32), 2021.

[13] S. B.Rikan, A. S.Azar, A.Ghafari, J. B.Mohasefi, and H.Pirnejad, "COVID-19 diagnosis from routine blood tests using artificial intelligence techniques", Biomed. Signal Process. Control, vol. 72,pp. 103263 (1-16), 2022.

[14] Y. Lee et al., "The application of a deep learning system developed to reduce the time for RT-PCR in COVID-19 detection", Sci. Rep., vol. 12, pp. 1234 (1-10), 2022.

[15] D.Yang, C.Martinez, L.Visuña, H.Khandhar, C.Bhatt, and J.Carretero, "Detection and analysis of COVID-19 in medical images using deep learning techniques", Sci. Rep., vol. 11, pp. 19638 (1-13), 2021.

[16] F.Zhang, "Application of machine learning in CT images and X-rays of COVID-19 pneumonia", Medicine, vol. 100 (36), pp. e26855 (1-13), 2021.

[17] S.Guhathakurata, S.Kundu, A.Chakraborty, and J. S.Banerjee, "A novel approach to predict COVID-19 using support vector machine", Data Science for COVID-19, pp. 351-364, 2021.

[18] N. S.Özen, S.Saraç, and M.Koyuncu, "COVID-19 vakalarının makine öğrenmesi algoritmaları ile tahmini: Amerika Birleşik Devletleri örneği", European Journal of Science and Technology, vol. 22,pp. 134-139, 2021.

[19] R. Ünlü and E.Namlı, "Machine learning and classical forecasting methods based decision support systems for COVID-19", Comput., Mater. Contin., vol. 64(3), pp. 1383-1399, 2020.

[20] C. N.Villavicencio, J. J. E.Macrohon, X. AInbaraj., J. -H.Jeng, and J. -G.Hsieh, "COVID-19 prediction applying supervised machine learning algorithms with comparative analysis using WEKA", Algorithms, vol. 14(7), pp. 201 (1-22), 2021.

[21] M. Malik et al., "Determination of COVID-19 patients using machine learning algorithms", Intell. Autom. Soft Comput., vol. 31(1), pp. 207-222, 2022.

[22] F.Shahid, A.Zameer, and M.Muneeb, "Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM", Chaos Solit. Fractals, vol. 140, pp. 110212 (1-9), 2020.

[23] T. B.Alakus and I.Turkoglu, "Comparison of deep learning approaches to predict COVID-19 infection", Chaos Solit. Fractals, vol. 140, pp. 110120 (1-7), 2020.

[24] K.Moulaei, M.Shanbehzadeh, Z.Mohammadi-Taghiabad, and H.Kazemi-Arpanahi, "Comparing machine learning algorithms for predicting COVID-19 mortality", BMC Medical Inform. Decis. Mak., vol. 22, pp. 2(1-12), 2022.

[25] A. B.Majumder, S.Gupta, D.Singh, and S.Majumder, "An intelligent system for prediction of COVID-19 case using machine learning framework-logistic regression", J. Phys. Conf. Ser., vol. 1797 (1),pp. 012011(1-9), 2021.

[26] J. M. Antoñanzas et al. "Symptom-Based Predictive Model of COVID-19 Disease in Children", Viruses, vol. 14, pp. 63, 2022.

[27] C. Huang et al., "Clinical features of patients infected with 2019 novel Coronavirus in Wuhan, China", The Lancet, vol. 395(10223), pp. 497-506, 2020.

[28] N. Chen et al., "Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study", The Lancet, vol. 395(10223),pp. 507-513, 2020.

[29] Kaggle, "Symptoms and COVID Presence", 8 January 2022. Available online: https://www.kaggle.com/hemanthhari/symptoms-and-covid-presence

[30] A. Tsatsakis et al., "SARS-CoV-2 pathophysiology and its clinical implications: An integrative overview of the pharmacotherapeutic management of COVID-19", Food Chem. Toxicol., vol. 146,pp. 111769, 2020.

[31] E. Karaahmetoğlu, S.Ersöz, A. K.Türker, V.Ateş, and A. F.İnal, "Evaluation of profession predictions for today and the future with machine learning methods: emperical evidence from Turkey", Journal of Polytechnic, (in press)

[32] I.Balikci Cicek and Z.Kucukakcali, "Classification of prostate cancer and determination of related factors with different artificial neural network", Middle Black Sea Journal of Health Science, vol. 6(3), pp. 325-332, 2020.

[33] N.Yalcin, G.Tezel, and C.Karakuzu, "Epilepsy diagnosis using artificial neural network learned by PSO", Turk. J. Electr. Eng. Comput. Sci., vol. 23(2), pp. 421-432, 2015.

[34] S.Cakir, S.Toklu, and N.Yalcin, "RPL attack detection and prevention in the Internet of Things networks using a GRU based deep learning", IEEE Access, vol. 8, pp. 183678-183689, 2020.

Journal of Emerging Computer Technologies
Research Article
Received: *2022-06-24* | ***Reviewing:*** *2022-06-24 & 2022-07-18*| ***Accepted:*** *2022-07-19* | ***Online:*** *2022-07-19* | ***Issue Date:*** *2022-06-30*

Data Visualization Tools - Preview and Comparison

Fehmi Skender
PhD Candidate, International Vision University.
Faculty of Engineering and Architecture
Gostivar. North Macedonia.
fehmi.skender@vizyon.edu.mk
0000-0001-6046-5772

Violeta Manevska
Full Professor, University St Kliment Ohridski
The Faculty of Information and Communication Technologies
Bitola. North Macedonia.
violeta.manevska@uklo.edu.mk
0000-0003-4401-1230

Abstract— **Data visualization enables the display of information and data in a visual context. Data visualization tools and visual data analysis through visual elemQents make it possible to see trends and changes in data.**

Data visualization involves converting data into a visual context with the sole purpose of facilitating the human brain to understand and inspect the data, that is, to understand the results of their processing. The visual canal is the fastest human cognitive channel. Due to this, more people want to see data rather than hear or read it. In a time of digital data overproduction, it is usual for fast data processing, analysis, and display to become a priority increasingly in everyday life. On the one hand, developers and the other hand, managers and users are in constant communication, which leads to continuous updating and updating of processing tools and visual data analysis.

In this context, the paper will review the current visualization and visual data analysis tools and compare their performance.

Keywords: Big Data, Data Visualization, Visual Data Analysis, Data Visualization Tools.

I. Introduction

Visual data analysis (VDA) is a technique for preparing and displaying data to clarify the data for analysts and managers, as well as all those who deal with the management and use of big data in their management of activities. Big data analysts are particularly interested in data visualization tools, which will enable the interpretation of as much useful information as possible and support them in the decision-making process [1].

The rapid development of technologies is primarily based on the use of IT technologies but also on data analysis with the growing use of visualization, i.e., more accessible information in the visual form [2]. The visualization itself is based on the rapid perception of visual forms by man because the brain of an average person quickly memorizes visual representations and data that it will receive from the environment through the visual sense. Science is increasingly concerned with finding ways and procedures to facilitate daily human life. The competition between the tools available in the market for data visualization is based on the rule that they are "closer" to users, ie. User friendly.

At a time of increasing dynamism in all spheres of life, the process of making the right decision is becoming more and more relevant. If it is known that the wrong decision is better than indecision, then it is very clear why the tools for visual display of data are increasingly a factor for success in decision making. Because of all this, different data visualization and visual data analysis tools are increasingly present in the software markets today. While developers try to find simple, content-friendly, and easy-to-use tools, users have a more excellent choice with more user-friendly tools.

II.Related Works

The capabilities of the highest-ranking tools for data visualization and visual data analysis are continuously applied and, simultaneously, upgraded with new features by many researchers and analysts. To summarize some of their results, we will review the most commonly used techniques, methods, and tools used by big data analysts and researchers. Development in this area has been enhanced by the advent of powerful tools for interactive visual data analysis that is available through the many user-friendly tools available on the market. Minatogawa et al. used a design methodology to develop an artifact to assist in seeking business model innovation. The customer-driven artifact uses indicators to measure the performance of companies' business models, powered by Big Data analytics to enable business model innovations. The research results show the artifact's successful assistance in the proactive and continuous effort toward creating the business model. Although based on technical concepts, the artifact is accessible to the small business context, helping to democratize business model innovation practices and big data analytics outside large organizations. Big data and its visualization facilitate the results of large statistical operations and help decide and mediate innovations, especially in the business sector [3].

Alharthi et al. visually researched the organization of the Hajj religious rite in Mecca, Saudi Arabia. They emphasize the importance of using graphical visualization with graphs and tables to analyze big data. Their specific research is related to the organization of religious rites and required and offered services. It flows of believers for specific research on the necessary data for Hajj, which the organizers needed in the decision-making process. The paper shows how to effectively use big data visualization based on deeper details in service visualization, thus identifying needs for improvement of certain services. The research clarifies the fact that ordinary visual representations on a map do not provide details that focus on improving services by taking the right action at the right time. Advanced visual data analysis

Cite (APA): Skender, F., Manevska, V. (2022). Data Visualization Tools - Preview and Comparison. *Journal of Emerging Computer Technologies*, 2(1), 30-35.

techniques are needed to detect problems and focus the attention of decision-makers on proper assessment to improve services. [4].

Kennedy et al. researched an excellent approach to studying the factors influencing the general public and the media. Among other things, the research proposes bridging the paradigms for HCI (human-computer interaction) and media and communication studies to develop awareness among people about the effective use of data visualization as a method. The conclusion of the paper mentions the need for greater understanding and interpretation of the visual data of the general public, i.e., for different groups of society [1].

Jagadish et al. explored the technical challenges of visualizing large amounts of data. In addition to defining many concepts of big data and their visualization, it examines the challenges posed by big data visualization but also explains big data as a certainty in the modern way of life, in business, and their use by the wider public [5].

Ventartaman et al. explored the possibilities of using cloud computing in visualization, with predictions about the new data center constellation and the use of the VDA. Here, the big data visualization process analysis is based on providing business intelligence for timely and effective decision-making. The paper recognizes the value of creating a big data infrastructure and delivering higher performance and scalable business intelligence in different organizations. The use of state-of-the-art tools and technologies for big data infrastructure, as well as the NIST (National Institute for Technology Standards) framework, is demonstrated. The advantages of data visualization are illustrated with in-depth scenarios of various production examples. In short, the paper contributes greatly to providing valuable insights into the flow of big data from organizations to enable a scalable infrastructure to make more informed and better critical decisions [6].

Rushton, G. has taken rudimentary steps to establish the basis for spatial analysis for Geographic Information Systems - GIS, for future data processing and more effective data processing and decision-making primarily in the health sector. In short, the paper contains systems that contribute to general public health in the United States that are geographically mapped so that they can be considered the beginning of the visualization of modern data [7].

Kumar and Singh in their paper reach large amounts of structured, unstructured, and semi-structured data, in short, heterogeneous data concerning big data, their processing tools, analysis, and decision-making techniques. Their research is based on the impact of big data in the healthcare sector and various tools such as Hadoop in exploring the conceptual architecture of data analysis. The research is based on the use of the Hadoop and MapReduce tools [8].

Debrus R. presents detailed information and a range of capabilities on Power BI data visualization boards. Explains the reasons why processed data becomes the very final process visualization essentially. The paper presents the advantages of Power BI in research, such as user interface, Consolidation of multiple data sources (Excel, CSV, XML, Text), interactive reports and map mapping, R, Python, and SQL integration, through various examples of the benefits of data analysis in decision making have been demonstrated [9].

Der. G and Everitt B in SAS research papers, justify the justification of emphasizing the skills necessary to perform statistical analysis based on visualized results. The effectiveness and importance of the SAS tool as a basis for processing big data in the study are also explained. The paper also contains research done on divergent diagrams, correlation, simple regression, and operations offered by the SAS tool [10].

In the research of Mani. M and Fei. S data visualization is considered an integral part of big data analysis. The research examines effective ways of visualizing big data, focusing on visualizing interactive processes. During an interactive visualization session, the analyst can issue multiple visualization requests, with each visualization subsequently building on previous visualizations. The research covers integrating distributed data processing systems that can effectively process large data with a visualization system, even effective interactive visualization for smaller amounts of data. Emphasis is also placed on the search for alternatives so that the delay period of the visualization is minimized. All this is taken as a conclusion which is also a possible answer to future demands, as a basis for new experiences for analysts, which would increase their productivity [11].

Many other researchers have worked with tools that enable visual data analysis of big data and have obtained outstanding visual representations. Because of this, we will look at the possibilities of the tools most commonly used today for VDA and interactive visualization.

III. Preview Of Data Visualization Tools

Undoubtedly, data visualization is a rapidly growing scientific field that arouses great interest among data scientists and the scientific community that applies this science in practice to obtain more talkative, faster, and more efficient observations that will lead to data penetration. And will help make more efficient, quicker, and more useful decisions for their organizations. Therefore, presenting the data, which prefers using innovative techniques, primitives such as colors, elements, and dimensions, and analyses that affect the representativeness of the data, is very important. Because data visualization, especially in big data, requires not only knowledge of design and data but also basic statistical knowledge, a variety of visualization tools have been developed and used. We would single out a few of them.

a. Tableau Public

Tableau is one of the Gartner group's most highly rated visualization and visual data analysis tools. It is rated as one of the best visualization tools today with several versions, including the version available to the general public but with limited capabilities, named Tableau Public. The tool stands out with its easy, intuitive user interface and simple use. With Tableau Public, all kinds of visualizations can be performed easily and quickly, without requiring code information. To get the full capabilities of the software, you need to pay for a license and get the full paid version. In the paid version of the Tableau tool, the possibilities are greater, and at the same time, different types of data inherent in Excel or PDF format can be combined. In short, with Tableau Public, data can be visualized simply by creating tables, lists, maps, and many tools that are particularly effective for visualizing big data, but also for interactive data analysis between the data and the visualizations themselves [12].

With Tableau Public, data and tables can be combined or linked together. Data can be edited using group and cluster properties. Tableau Public is used in many sectors, from private companies to public institutions. You can share your own visualized data with other people online. The data can also be shared with mobile devices compatible with the tool [13].

The architecture of the Tableau lives version makes it easy to connect to Hive (Big Data Business Intelligence) resources, enabling online data analysis without needing any data transfer (when data reaches an order of five bytes in size). Tableau live's architecture protects investments in both big data and data warehouses. Tableau can also visualize large data by connecting directly to the operating database in memory. Tableau enables the analysis of hundreds of millions of rows of data with tremendous speed and capacity, and streaming capabilities.

The main features of the Tableau tool are: Adding a comment to the table, changing the view, dragging and dropping, having a Tableau Reader for viewing data, converting data into visualizations, the ability to reduce the size of data at different intervals, creating interactive tables, data exchange, highlighting and filtering data, spreadsheet sharing, data notifications, automatic display of new features and security permissions at any level.

b. Yellowfin Bi

Yellowfin BI is an analytics platform specializing in spreadsheets and visual analysis. As a modern tool for data visualization, it brings with it a rich library of pre-developed spreadsheets. The tool, with its structure, allows for greater performance of the users. The manufacturer offers a wide range of additional tools with the possibility of automated insight, built-in explanations, and contextual analysis. The tool easily connects to various resources and relational databases, such as Hadoop and NoSQL. With the features provided by the tool, faster responses to KPI based requests can also be found [14].

Features of the Yellowfin BI tool are: The answers are based on machine learning, as a result of automated insights, and machine learning algorithms, providing on-demand visualization and quite understandable and easy-to-present visualizations. Getting focused results by selecting or excluding certain data points in a spreadsheet, as well as enabling complete control of data points in their visualization, is another advantage of this tool. The mobile application is based on HTML5 and replaces the source application in version 6.3 of the tool. The tool automatically detects changes in the data and gives the results at the same time [15].

c. SAP ANALYTICS CLOUD

SAP ANALYTICS CLOUD is a tool for data visualization for business research, primarily designed for planning companies and analyzing their oversights and decision-making. Different data can be repeated and used at any time during processing. SAP ANALYTICS CLOUD supports hands-on collaboration based on simply sharing information in PDF format with SAP Analytics Cloud users as well as external collaborators [16].

Features of the SAP Cloud Analytics Story tool are: SAP Cloud Analytics Story software enables the use of a wide range of professional, interactive, important lists and other objects to display your data responsibly. By applying custom filters to widgets and spreadsheets, data can be personalized, and bookmarks can be entered into specific views for future research. The built-in calendar allows assigning, planning, and monitoring of the status, as well as temporary reminders for the same. Data can be shared on the spot with all dynamic elements when using the tool and in PDF format without dynamic functionality. At the same time, it provides security at the level of companies with access permits for owners, and users, individually or in groups.

d. ORACLE ANALYTICS CLOUD

ORACLE ANALYTICS CLOUD is a data visualization tool for joint reporting with analytics for organizations of all data sizes, with the capabilities of the machine learning tool enabling the detection of critical information through intuitive data visualization.

One of the major advantages of ORACLE ANALYTICS CLOUD is the processing of data in the natural language, i.e., there is no need for good mastery of code to discover and analyze data that is of interest to the user. At the same time, ORACLE ANALYTICS CLOUD, with the help of multidimensional data analysis and scenario simulations, provides a much more detailed view of business data [17].

Features of the ORACLE ANALYTICS CLOUD tool are: The ability to automatically create visualizations by identifying algorithms based on machine learning in big data analysis is multiple features of the tool. The intuitive interface of the tool provides identical data analysis in various data visualization techniques and methods.

e. DOMO

DOMO is a cloud-based Business Intelligence (BI) platform with thousands of built-in visualizations, including more than 1,500 list types and approximately 7,000 maps. Machine learning enables automated data detection by alerting and asking questions in natural language. The DOMO tool can also access, filter, sort, and group data deep after its data genesis. The tool lets you customize your layout by adding custom colors, images, and text [18].

Features of the DOMO tool are: Ability to combine individual views, for example, combining sales, return on investment, performance measurement, and all other key performance indicators that can be selected in one interface. With the help of the Pop-Up menus, the individual visualizations from the table can be separated. The tool provides the opportunity to improve the appearance of the visual display by selecting different types of lists, as well as data series and filters to enhance the visualization itself interactively.

f. POWER BI

Power BI is a collaborative tool for software services, applications, and links that work together to enable the transformation of independent data sources into consistent, visual, and interactive insights. The data can be from Excel, a collection of cloud data, or hybrid systems in various formats. Power BI makes it easy to connect to current data sources, and visualize and share any part of the results in the cloud or between groups of users. Power BI consists of several different components that work together, including the three basic elements:

- Windows desktop application or Power BI Desktop,
- Online SaaS or Power BI service,
- Power BI mobile applications for Windows, iOS and Android devices.

Features of Power BI tool are: The biggest advantage of Power BI is its availability and being relatively inexpensive as a data visualization tool. Power BI Desktop version is free. It is easy to download, install and use.

The use of Power BI is with easy custom visualizations in reports and spreadsheets and allows interactivity in visualization. Features include many tools, such as visualizing key performance indicators, maps, charts, graphics, R image scripts, dashboards, and more. An important advantage of using Power BI as a data analysis tool is the ability to transfer data from a wide range of sources and link data to XML and JSON formats. In addition to the many advantages, one of the most important features of Power BI is the establishment of direct access to big data resources. It should also be noted that Excel integration includes the option to load Excel data into Power BI. The option to import unprocessed data into Excel allows it to be available for processing, to cut some of it as data from Power BI, and to easily transfer it back to Excel. Power BI can be used to add various visualizations to the report with the ability to establish interaction [19].

g. Other Tools

Other Gartner-rated data visualization software applications include:

TIBCO SpotFire **-** is a software platform that assists in software integration and business visual analysis. The software is also used as software for data integration and analysis, as well as for rapid detection of concepts for better decision making. TIBCO Spotfire is a software platform that simply"animates" data, in various big data surveys [20].

SAS VDA - SAS (Statistical Analysis System) is a software package that provides statistical analysis that provides access to big data, data management, analysis, and presentation of data. It also offers a range of products that can be used in many areas related to data analysis, data cleansing, or data analysis operations. SAS is mainly used in big data processing and visualization [21].

Google Data Studio is a free online tool that converts data into customizable info reports and spreadsheets. Google Data Studio combines the marketing tools of Google or different data sources in one interface, all with the sole purpose of creating a spreadsheet. It is quite easy to use and has a simple interface [22].

Qlik Sense -is a visualization and data discovery tool that allows you to create flexible, interactive visualizations and make meaningful decisions. The tool is licensed, which answers questions that come after the inspection or visualization. Qlik Sense answers questions constantly with its relational model, allowing continuing insights [23].

Amazon QuickSight - is a licensed tool. Amazon QuickSight lets everyone in organizations understand the data by asking questions in natural language, browsing through interactive spreadsheets, or automated machine learning. QuickSight supports millions of spreadsheet views for customers all the time with the sole purpose of assisting customers in decision-making [24].

IV. Discussion and Results

The research of several papers in the field of data and information visualization shows that the competition moves the software of the tools proportionally to the needs and opportunities. In summary, the current visualization tools in terms of their characteristics can be shown in the following table:

TABLE I. STRENGTHS AND WEAKNESSES OF DATA VISUALIZATION TOOLS.

DATA VISUALIZATION TOOLS	
ADVANTAGES	*WEAKNESSES*
TABLEAU PUBLIC	
• Manages big data and machine learning applications. • Salesforce can integrate with advanced database solutions like Hadoop, SAP, Teradata. • Effective graphics can be created.	o There is no option to refresh reports automatically. o The solution is not so comprehensive, and knowledge of SQL is inevitable.
YELLOWFIN BI	
• Ability to analyze operational and strategic planning of the organization by following the critical metrics through updated visualizations. • Expand spreadsheets by combining widgets, action keys, and codeless functionality with the JavaScript API. Create floating panels, filter controls with HTML, CSS, and JavaScript	o Does not allow change of hierarchies and data types inherited from OLAP. o Does not support financial planning and profit analysis
SAP ANALYTICS CLOUD	
• The tool allows deeper penetration into the data source through the Smart Predict, Smart Insights and Smart Discovery modules, allowing in-depth analysis. • Eliminate foresight in the decision-making process. • Save time, select data to find information when automatically generating relevant insights and descriptions	o Sometimes does not provide solid data validation for AI-based features. o Data modeling is not flexible enough.
ORACLE ANALYTICS CLOUD	
• Ability to automatically create submersible visualizations by identifying insights through machine learning algorithms. • Its intuitive interface provides the same data analysis in multiple visualizations. • Provides specific answers to specific business questions through personalized searches. • Regardless of technical skills, it provides opportunities for research and analysis.	o Does not provide enough graphics. o Not intuitive in terms of reinforcement and recovery. o Does not include visualization like its competitors.
DOMO	
• Ability to ask and answer questions from everyday life, the ability to lower the standards for in-depth analysis. • Working in teams with clients on visualization reduces the need for meetings. • Ability to record specific data points, leave comments and observations so they can be tracked by other clients.	o Does not allow export of more than 5 MB of additional data in the scheduled CSV reports. o Does not allow easy removal of data from the platform. o Not intuitive enough in terms of user interface.
POWER BI	
• The Power BI user receives data from various data sources, such as files, Azure Resources, online services, queries, or gateway resources. Then they work with this data in a development tool like Power BI Desktop. Here, the imported data is cleaned and converted according to the needs of the user. • Reports created on the Power BI Desktop can then be published on two types of platforms: Power BI Service and Power BI Report Server. Power BI Service is a public cloud-based platform, and Power BI Report Server is a firewall-protected platform. • Spreadsheets and reports can be shared using web browsers, tablets, laptops, phones, etc.	o Provides AutoML only in Power BI DataFlows. This means that it is not possible to run AutoML on a Power BI Desktop. o It is also not possible to autom after creating the data model. o There is no intelligence for an advanced position. o Lack of opportunity for graphical analysis. o PowerQuery & m is the most powerful tool in Power BI, while DAX and the data model are an obstacle that weighs equal to Power BI.

V. Conclusion

As a result of the analysis of the mentioned tools for visual data analysis and visualization of big data, we conclude that they are suitable for use in organizations. These tools have been developed for solving dynamic problems such as creating losses or illiquidity. Modern organizations use visualization tools effectively, thus advancing production and services much more confidently and with better quality.

Big data processing and its visualization greatly help all users, consumers, manufacturers, and managers. More and more attention is being paid to using visual facts when deciding and investing. Using tables and graphs is a certainty in today's dynamic time in such dynamic environment.

The number of tools for data visualization is growing rapidly, and they are constantly improving and increasing their capabilities for analysis, especially in the field of Big data. If we know that the human eye is attracted by visual representations, colors, and patterns, i.e., 90% of the information presented in the brain is visual, it can be easily concluded that data visualization is the most desirable technique for business and, in general for everyday needs for visual data analysis. in any area of life. As a result, the trend of increasing the use of visualization tools is increasing, and the use of tools is becoming more than necessary.

Therefore, the paper analyzed research on previously published articles that use big data visualization for various tools, including tool analysis, video data analysis, and interaction. The study focuses on big data using visualization tools where their key performance indicators are analyzed. Finally, it can be rightly noted that Big Data, Visualization, Machine Learning, and Artificial Intelligence are not only a challenge for analysts, managers, managers, and professionals but are increasingly becoming a necessity and inevitability for every human being. The analysis supported by the possibility of interaction in the processing and visualization of big data results in a solid basis for obtaining useful and insightful information as a prerequisite for success in all fields.

References

[1] H. Kennedy, "Engaging with data visualisation," http://journals.uic.edu/ojs/index.php/fm/rt/printerFriendly/6389/5652, 2016.

[2] S. Aybeyan, "Systematic Review Of Big Data, Digital Transformation Areas And Industry 4.0 Trends In 2021," International Scientific Journal Vision, vol. 6, no. 2, pp. 27-41, 2021.

[3] M. Vinicius , M. M. V. Franco, I. S. Rampasso and R. Anholon, "Operationalizing Business Model Innovation through Big Data Analytics for Sustainable Organizations," Business model Innovation Research - Sustainability 12(1):277, no. 12, p. 277, 2022.

[4] A. Alharthi, V. Krotov and M. Bowman, "Addressing barriers to big data," p. https://www.researchgate.net/publication/317279522, 2017.

[5] Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou and J. Patel, "Big Data and Its Technical Challenges," Communications of the ACM, pp. 86-94, 2014.

[6] VENKATRAMAN, "Cloud Computing A Research Roadmap In Coalesence With Software Engineering," An International Journal (SEIJ), vol. 2, no. 2, 2012.

[7] G. Rushton, "Public health, GİS and spatial analytic tools," Annu.Rev. Pub.Health 2003 - Department of geography - Unıversıty of Iowa, USA, no. 24, 2003.

[8] Maninder and S. S. Kumar , "Big Data Analytics for Healthcare Industry," Big Data Mining And Analytics, vol. 2, no. ISSN222096-0654 l l05/06l, pp. 48-57, 2019.

[9] D. Roxane, "Power BI in Clinical Data exploration,," PHUSE Virtual EU Connect 2020, 2020.

[10] G. Der and B. S. Everitt, Statistical Analysis of Medical Data Using SAS, vol. 16, 2006, p. http://ftp.sas.com/samples/A60928.

[11] S. F. Murali Mani, "Effective Big Data Visualization," In Proceedings of IDEAS ’17, Bristol, United Kingdom, July 12-14, 2017, 6 pages., 2017.

[12] "https://www.tableau.com/products/public," TABLEAU SOFTWARE, 2022. [Online]. Available: https://www.tableau.com/products/public.

[13] A. Ohmann and M. Floyd, Creating Data Stories with Tableau Public, Packt Publishing, ISBN: 9781849694766, 2015.

[14] "https://wiki.yellowfinbi.com," Yellowfin, 12 6 2021. [Online]. Available: https://wiki.yellowfinbi.com/display/yfcurrent/REST+API.

[15] www.yellowfinbi.com, "https://www.yellowfinbi.com/," 2022. [Online]. Available: https://www.yellowfinbi.com/.

[16] https://community.sap.com, "https://community.sap.com," 2022. [Online]. Available: https://community.sap.com/topics/cloud-analytics#:~:text=SAP%20Analytics%20Cloud%20is%20a,you%20with%20your%20learning%20journey.

[17] https://docs.oracle.com, "https://docs.oracle.com/en/cloud/paas/analytics-cloud/visualize-data.html," 6 2022. [Online]. Available: https://docs.oracle.com/en/cloud/paas/analytics-cloud/visualize-data.html.

[18] https://www.domo.com/, "https://www.domo.com/solution/data-visualization-software#:~:text=Domo's%20data%20visualization%20software%20lets,faster%2C%20better%2Dinformed%20decisions.," [Online]. Available: https://www.domo.com/.

[19] A. S. Rob Collie, "Power Pivot and Power BI," in Power Pivot and Power BI, Yayınevi: Holy Macro! Books / ISBN: 9781615473496, 2015.

[20] sistembul.com, "https://www.sistembul.com/cozumler/tibco-spotfire," 21 3 2022. [Online]. Available: https://www.sistembul.com/cozumler/tibco-spotfire.

[21] globaltechmagazine.com, "https://www.globaltechmagazine.com," 2022. [Online]. Available: https://www.globaltechmagazine.com/2016/06/06/sas-viya-devrim-yaratacak-yeni-analitik-mimarisi/.

[22] newgenapps.com, "https://www.newgenapps.com/mk/blogs/what-is-the-google-data-studio-2/," 18 12 2019. [Online]. Available: https://www.newgenapps.com/mk/blogs/what-is-the-google-data-studio-2/.

[23] "https://home.vizlib.com/vizlib-library-for-qlik-sense/," vizlib, 2022. [Online]. Available: https://home.vizlib.com/vizlib-library-for-qlik-sense/.

[24] "https://aws.amazon.com/quicksight/," aws.amazon.com, 2022. [Online]. Available: https://aws.amazon.com/quicksight/.