International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.7110
***Article Received**: 20 March 2023 **Revised**: 18 May 2023 **Accepted**: 30 May 2023*

A Novel Approach for Mining Big Data Using Multi-Model Fusion Mechanism (MMFM)

R.Ramprasad[1], C.Jayakumari[2]
[1]Research Scholar
Bharathiar University, Coimbatore, Tamil Nadu 641046.
Email:trramprasadphd@gmail.com
[2]Middle East College, Oman.
Email: jayakumari@mec.edu.om

Abstract—Big data processing and analytics require sophisticated systems and cutting-edge methodologies to extract useful data from the available data. Extracted data visualization is challenging because of the processing models' dependence on semantics and classification. To categorize and improve information-based semantics that have accumulated over time, this paper introduces the Multi-model fusion mechanism for data mining (MMFM) approach. Information dependencies are organized based on the links between the data model based on attribute values. This method divides the attributes under consideration based on processing time to handle complicated data in controlled amount of time. The proposed MMFM's performance is assessed with real-time weather prediction dataset where the data is acquired from sensor (observed) and image data. MMFM is used to conduct semantic analytics and similarity-based classification on this collection. The processing time based on records and samples are investigated for the various data sizes, instances, and entries. It is found that the proposed MMFM gets 70 seconds of processing time for 2GB data and 0.99 seconds while handling 5000 records for various classification instances.

Keywords- Mining, Data analysis, Prediction, Classification, Multi-Model Fusion Mechanism.

I. INTRODUCTION

The big data paradigm, which works with large amounts of stored and exchanged information, has been made possible by the accessibility of and progress in data sources. It leads to using distributed analysis to handle, manage, and analyze big data from various environments [1]. To offer options for real-time development, From the application end, processed data is displayed [2]. The first step in gathering information from different sources is big data acquisition, including businesses, homes, the healthcare sector, applications for road transportation, etc. Such data processing requires extra processing time and support for storage types and hardware architectures [3]. But information processing provides the best answers by visualizing complex data to create useful bits of information. Delays and exceptional device support should be minimized, and huge and raw data should be analyzed and managed using complex methods [4]. The traditional distributed data processing platform needs to be more and more able to handle it. As more data is generated, more apps and real-time demands require sophisticated big-data handling [5]. The amount of data exchanged and processing complexity grow due to the variable application size and user requirements. Another challenge in managing big data is finding and storing data. Different data management strategies are suggested to handle big data analytics' accuracy and delay problems [6].

The data's intensity and necessity are determined by various characteristics in the information gathered from various autonomous sources. In mining apps, the gathered data is displayed as a batch and is represented in various structures. For this reason, data classification is essential [7]. By categorizing data according to its size, characteristics, applications, and services, it can be described in more straightforward structures suitable for retrieval and querying [8]. To query and retrieve information constrainedly, techniques for processing and mining data block entry to these structures, which reduces complex and lengthy processing. Using exact classification techniques developed [9], Access, querying, mining, and information retrieval problems are lessened. Elaggoune et al. [10] used neural learning to categorize data according to the attributes connected to it. Similar to this suggestion, classification is modified using batch processing devices and distributed systems to address the problems with excessive resource consumption. The semantic and grammatical problems in data administration and processing are reduced by using reliable classification. This method is frequently used for information querying in big multi-record storage systems and databases [11]. Fuzzy decision-making and fuzzy logic are used to handle complicated issues with real-time processing. To outperform the current data processing architectures, fuzzy algorithms are used for big data analytics [12]. A knowledge base is needed to pull information from big data analytics for constraint-based analysis, judgment, and data representation. A simple processing system cannot control these characteristics necessitating the addition of algorithms and techniques for information extraction. In big data, fuzzy systems can be applied to various tasks, including reasoning, attribute-based classification, data migration, planning, and display [13]. These decision-making systems establish the basic criteria for

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.7110
***Article Received**: 20 March 2023 **Revised**: 18 May 2023 **Accepted**: 30 May 2023*

evaluating and processing large amounts of data to improve processing accuracy and speed. According to [14] and [15], updated fuzzy systems are beneficial to increase processing levels that can be sustained without using up more resources. These systems perform timely processing, and intricate queries over various accumulated data attributes especially well. The following is the article's contributions:

1) The development of a Multi-model fusion mechanism for data mining (MMFM), which examines inbound semantics to ascertain the relationships and limitations of the data;
2) The semantics of the inbound data is categorized using similarity for the most effective data administration to hasten the coding and processing of the info gathered;
3) Conduct a comparative analysis to assess how long it takes to process and classify big data across various cases and input records. Here, the performance is evaluated with the processing time and data samples.

The work is drafted as section 2 gives wider reviews of prevailing approaches and their pros and cons. The methodology is drafted in section 3, and numerical outcomes in section 4. The work is summarized in section 5 with future research enhancements.

II. RELATED WORKS

Big data analytics gathers a sizable amount of information and technology from various sources to give a company an advantage over competitors. According to [16], big data refers to enormous amounts of empirical data used in decision-making. Big data was defined as real-time or media data in non-traditional formats, IT-driven, social media and enormous data [17]. Numerous studies emphasize the crucial roles of "Veracity" and "Velocity," even though "Variety" and "Volume in Big Data" drawn huge attention. It is critical that the fundamental elements of BDA are analytical instruments and abilities [18]. BDA are described by C. Zhao et al. [19] as gathering, merging, examining, and using Large data collection from various independent sources. This process aims to find trends and other helpful information that can be used to enhance management choices.

Furthermore, BDA is defined by Inkinen et al. [20] as a method for identifying and deriving significant findings from big data to support decision-making. Businesses can use BDA to analyze data from both internal and external sources to spot important patterns. BDA is reportedly being used by many businesses as a potential value-creator to assist in decision-making by Rush et al. [21]. Big data analysts must be able to create intuitions and find consequences [23]. Effective BDA execution requires using suitable analytical resources for the investigation [24]. BDA takes on the most modern systematic approaches for resolving business issues that are impractical previously due to the lack of data or analytical tools.

Organizational performance is achieving objectives, meeting stakeholders' expectations, and remaining competitive [25]. An organization's performance is examined and evaluated regarding its goals and objectives through a procedure that contrasts actual results with desired results, which can also be described as it [26]. The OP compares the organization's real outputs or productivity to its goals or desired outcomes. According to Sun et al. [27], improved performance relies on an organization's capacity to manage innovation, safeguard intangible knowledge assets, and utilize these resources to the organization's advantage. Ensuring that organizational resources are used effectively can also be referred to as OP. It encompasses every action or activity a manager takes at different organizational levels to evaluate how effectively an organization has accomplished its goals.

Knowledge management methods simplify the systematic knowledge generation, acquisition, conversion, and implementation process. As part of KMP, knowledge is gained, stored, distributed, and applied. According to Nonaka and Takeuchi, tactic and explicit knowledge are included in the four phases of conversion that make up knowledge creation [28]. It is a crucial component of KM theories and practices. Understanding is the power that aids in resolving organizational issues. KMP stands for knowledge resource acquisition, conversion, evaluation, retrieval, and dissemination to improve and effectively function organizationally. Additionally, it fosters expansion and competitive advantage. Organizations are very concerned about creating and managing knowledge to enhance company success.

Akter et al. [29] developed deep learning approaches on visualization and data analysis. This approach is used for large-scale medical data management. Deep learning is used to classify the characteristics given for computation, and the assembler is consented to for validation. The Hadoop ecosystem is used in this plan to categorize and share medical data. It is done with big data to increase manufacturing process dependability. It is utilized in engineer-to-order (ETO) production processes for reliable decision-making to enhance product quality. Based on the analytics gathered from the ETO, the manufacturing execution system produces reports that suggest manufacturing support for quality. Big data is a paradigm that Braganza et al. [30] proposed to enhance improvement in acquisition and storage system (ASS) design for industrial information administration. The enhanced ASSs architecture prioritizes shortening the time to process data by taking advantage of storage space constraints.

III. METHODOLOGY

Big data management and handling across various resources taint the accuracy of the research. It is due to a need for more consolidated data or accurate classification. Big data processing becomes unclear as a result of this circumstance. The MMFM technique is developed and discussed in this article to enhance

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.7110
***Article Received**: 20 March 2023 **Revised**: 18 May 2023 **Accepted**: 30 May 2023*

data analytics reliability. The accuracy of the data and granular analysis, in particular, is improved to attain reliability measures. The optimized management technique searches for input semantics classification to ensure accurate data analysis. Fig 1 depicts the data management paradigm for sensor inputs.

The communication and access infrastructure in this MMFM paradigm connects the sensors that collect environmental data toward the manufacturing hub. Within the manufacturing facility, the fuzzy process handles aggregated data. Data centers with sophisticated computing tools, top-tier processing, and ample storage typically serve as processing centers. Here, the gathered data is examined, extracted, and displayed. Weather forecasting facilities and traffic management facilities are a couple of typical instances. The perceived information must be transmitted, which is the responsibility of infrastructure and communication. The fuzzy analysis manages this raw data in the back end for input semantics and classification. The process stated above is further explained in the sections that follow. Fuzzy rules and classifiers are used for analysis and improvement.

a) Model outline

The memory, attention, fusion, gating and aspect-factor representation layer are all parts of the Multi-model fusion mechanism for data mining (MMFM). Fig 1 depicts the MMFM network architecture. The symbols $S = [w_1, w_2, \dots, w_L]$ are symbolized by k −dimensional embedding vector w_i. Here, L represents the sentence's longest word count. Let a_I stand in for the I^{th} aspect of the sentence S while this happens, and let $0 < I < M < L$ stand in for most aspects.

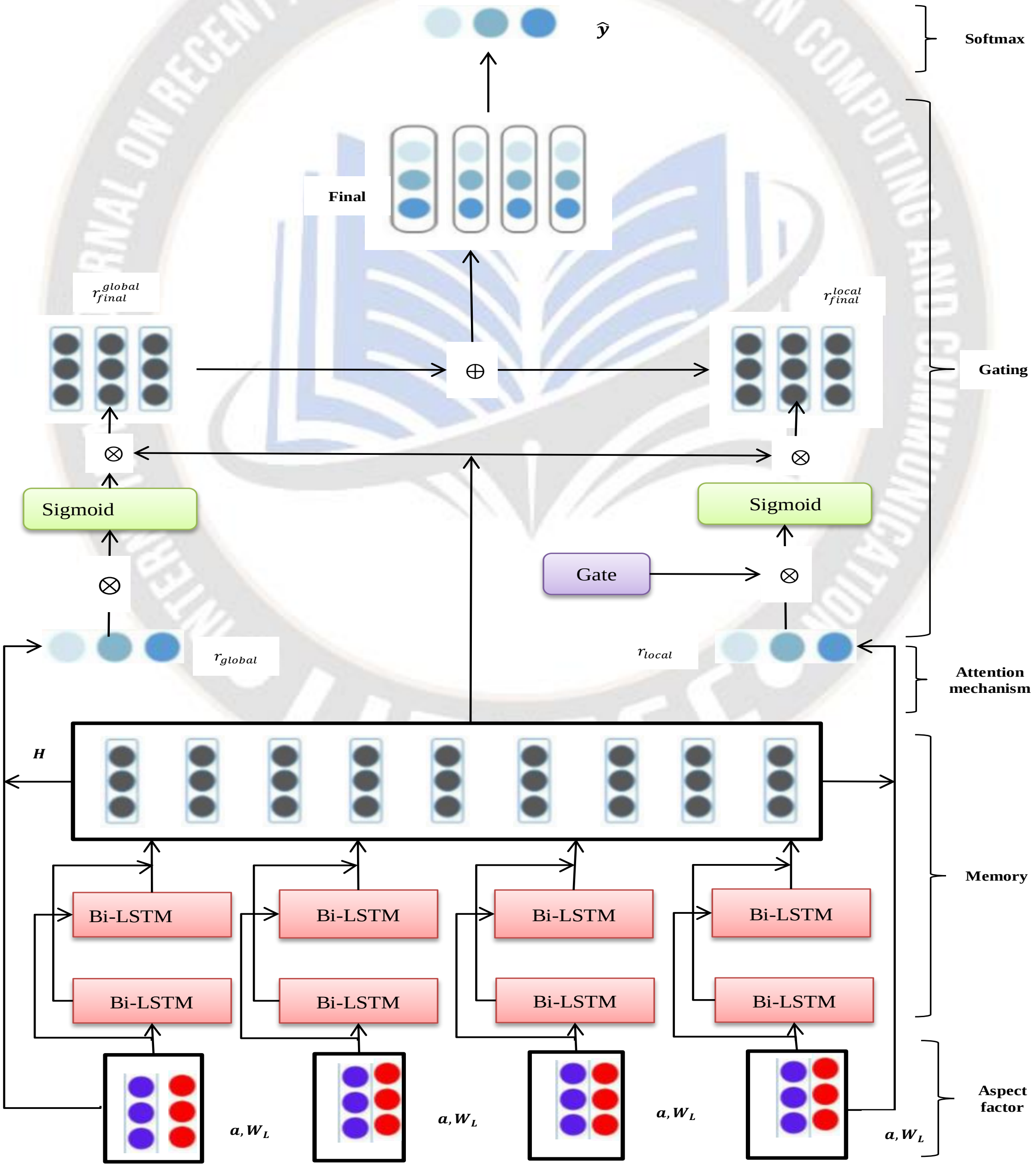

Fig 1 Network framework

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.7110
***Article Received**: 20 March 2023 **Revised**: 18 May 2023 **Accepted**: 30 May 2023*

b) Data representation

Only some words in a phrase have the sentimental content necessary to convey a specific idea, and some words, like stop words, lack sentimental content entirely. As a result, we only need one sentence to convey a feeling about a particular feature. MMFM first joins every word w_i in the phrase S to aspect a_i. Only some words in a phrase have the emotional content necessary to convey a specific idea, and some words, like stop words, lack sentimental content entirely. As a result, we only need one sentence to convey a feeling about a particular feature. MMFM first joins every word w_i in the phrase S to aspect a_i:

$$S_{a_1} = [[w_1; a_I], [w_2; a_I], \dots, [w_L; a_I]] \quad (1)$$

Where the concatenate procedure is indicated by $[;]$.

c) Vector representation

The MMFM model transmits word vectors into Bi-directional Long Short-Term Memory network to maintain context (Bi-LSTM). Its concealed layer size is D_0. The word embedding for forward $LSTM$ is S_{a_I}, and the concealed state at time step t_1 is $\vec{h}_{t-1}$. The secret state can be found $\vec{h}_t$ at time step t can be found as follows:

$$\vec{h}_t = \overrightarrow{LSTM}\,(s_{a_I,} \vec{h}_{t-1}) \quad (2)$$

Here, the input sequence is fed into the backward LSTM in the opposite direction from how it is fed into the forward LSTM. By using the hyperbolic tangent activation function on the merged output of the LSTMs for forward and reverse, MMFM creates the ultimate hidden state.

$$h_1 = \tanh\left(\vec{h}_t; \overleftarrow{h}_t\right) \quad (3)$$

At step t, the forward and backward LSTMs' respective results are $\vec{h}_t$ and $\overleftarrow{h}_t$. These two outputs are represented in the formula ht i (3). Here, $H = \{h_1, h_2, \dots, h_L\}$ represents the Bi-LSTM's final result, where $\overleftarrow{h}_t \in \mathbb{R}^{2D_0}$.

d) Relation establishment

This layer's main objective is to understand the semantic relationships in a particular sentence component using data from global and local views.

e) Aspect vector

Using the goal aspect vector a_I and the sentence representation, the attention value α_i is calculated for each word representation $H = \{h_1, h_2, \dots, h_L\}$.

$$m_i = W_{att2}^T \tanh(w_{att1}[h_i; a_I] + b_{att1}) \quad (4)$$

$$\alpha_i = \frac{\exp(m_i)}{\sum_{j=1}^{L} \exp(m_j)} \quad (5)$$

Where $W_{att1} \in \mathbb{R}^{(2D_0+k)*(2D_0+k)}$ and $W_{att2} \in \mathbb{R}^{(2D_0+k)}$ are weight vectors used in training and $b_{att1} \in \mathbb{R}^{2D_0+k}$. Weighted sum of hidden state h_i about its attention score i is then created as the global-attention-based representation $r_{glo} \in \mathbb{R}^{2D_0}$:

$$r_{glo} = \sum_{i=1}^{L} h_i \propto_i \quad (5)$$

f) Data dependencies

We must initially select words from the semantic information close to the specific objective because only a portion of the words in a sentence receive local focus. The dependency tree, which includes a wealth of linguistic data between words, can capture the connections between the target word's environment and its syntactic dependencies. Consequently, we present the dependency tree-based syntactic-based distance. Each word in a phrase S is represented by a node in the dependency tree D. A distance of one is considered to exist between any two linked nodes. To determine the distance between each remaining word and the chosen objective, we traverse D. Based on the target word's syntactic distance from it. A word is chosen for local focus. The difference between position-based distance and syntactic-based distance is shown in Fig 1. Syntactic-based word distance works better when combining semantic data than position-based word distance. We decide on terms that are grammatically t steps or less away from the goal term and use $LS(t)$ to represent them. Then, we give the words that match $LS(t)$ the following attention:

$$n_i = W_{att4}^T \tanh(W_{att3}[h_i; a_I] + b_{att2}) \quad (7)$$

$$\beta_i = \frac{\exp\,(n_i)}{\sum_{j \in LS(t)} \exp\,(n_j)} \quad (8)$$

Where, $i \in LS(t)$, $W_{att3} \in \mathbb{R}^{(2D_0+k)*(2D_0+k)}$ and $W_{att4} \in \mathbb{R}^{(2D_0+k)}$ are weight vectors and $b_{att2} \in \mathbb{R}^{2D_0=k}$ is bias. When there are numerous words in the target, only the words closest to each other by $t-$step are chosen. We give each context word a local attention value as an example:

$$\beta = \begin{cases} 0 & i \neq LS(t) \\ \dfrac{\exp\,(n_i)}{\sum_{j \in LS(t)} \exp(n_j)} & i = LS(t) \end{cases} \quad (9)$$

The representation of local attention is computed as follows:

$$r_{loc} = \sum_{i=1}^{L} h_i \beta_i \quad (10)$$

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.7110
***Article Received**: 20 March 2023 **Revised**: 18 May 2023 **Accepted**: 30 May 2023*

g) Attention mechanism

With local and global vectors, MMFM employs gating layer to merge certain target data from attention results. Eq. (11) uses global and local attention vectors to compute information gate g:

$$g = sigmoid\ (W_{gate}(r_{glo} + r_{loc})) \tag{11}$$

Where $W_{gate} \in \mathbb{R}^{2D_0 * 2D_0}$ is acquired through training and $g \in \mathbb{R}^{D_0}$, each word embedding dimension may represent various interpretations of a word's meaning. As a result, we employ a gated unit expressed by a vector instead of a scalar. We postulate that each gating unit dimension governs a unique aspect of the attention vector. The situation is then shown as follows:

$$r = r_{glo} + r_{loc} \odot g = \sum_{i=1}^{L} h_i \propto_i + \left(\sum_{i=1}^{L} h_i \beta_i)\right) \odot g \tag{12}$$

Where the Hadamard substance is represented by $\odot$. Moreover, the local concentration equation can be changed as follows:

$$\tau_{ij} = \beta_i . g_j \tag{13}$$

The hyperbolic tangent function might impact the value of g_j, where g_j stands for the gating unit j^{th} dimension, to be negative. Since the focus is defined as being positive, τ_{ij} might be negative. We employ Eq. (14) to maintain the attention non-negative and give the attention number significance.

$$t_{ij}^{loc} = sigmoid\ (\beta_i . g_j) \tag{14}$$

Then, after adding the attention ratings of each word in the j^{th} dimension, a normalization function is used to guarantee that the result is identical to 1:

$$\gamma_{ij}^{loc} = \frac{t_{ij}^{loc}}{\sum_{k=1}^{L} t_{kj}^{loc}} \tag{15}$$

Where w_i normalized attention value on the j^{th} dimension is represented by τ_{ij}. Finally, the following formula can be used to determine the situation that local attention represents:

$$r_{final}^{loc} = \sum_{i=1}^{L} h_i \odot \gamma_i^{loc} \tag{16}$$

The sigmoid function is also used to attain global focus $t_{ij}^{glo} = sigmoid(\propto_i)$. To acquire global attention representation, we can repeat the above procedures:

$$r_{final}^{glo} = \sum_{i=1}^{L} h_i \odot \gamma_i^{glo} \tag{17}$$

The ultimate context representation is created by combining local attention representation and the global attention representation's context representations:

$$r_{final} = r_{final}^{loc} + r_{final}^{glo} \tag{18}$$

h) Target representation

Softmax classifier size C (where $'C'$ specifies total categories) then generates final sentiment classification after getting target feature representation produced by gating layer:

$$\hat{y} = softmax\ (W_s rfinal + b_s) \tag{19}$$

Here, $W_s \in \mathbb{R}^{2D_0 * c}$ and $b_s \in \mathbb{R}^C$ are the softmax layer's training parameters, and $\hat{y}$ results from the mood polarity prediction.

i) Training process

Here, y and the actual name by $\hat{y}$ denote the predictive label. We train the network for 30 repetitions using category cross entropy and an $L2$-regularizer as the loss function:

$$L = -\frac{1}{N} \sum_{p=1}^{N} \sum_{q=1}^{C} y_{pq} \log(\hat{y}_{pq}) + \lambda \left|\left|\theta\right|\right|_2 \tag{20}$$

The entire parameter set that needs to be learned into the network is indicated by the letters N, where N stands for the total sample count, p for the sample index, q for the classification group, and w for the regularization weight. The ADAM algorithm is the most effective method for the MMFM model in the interim.

IV. NUMERICAL RESULTS AND DISCUSSION

This part discusses various metrics that can be used to validate the proposed MMFM. Sensor data for climate detection is used to verify MMFM. Establishing connections and boundaries for the incoming data allows for classification. The constraints are created, and the links are modeled using the environmental images taken. The instances of weather forecasting are used to create the modeling constraints. Temperature, relative humidity, perception, time of day, wind direction, and other sensor data are evaluated. A relationship of chained instances is created by assimilating the necessary elements. Learning-based generalization is created for data analysis using MATLAB. Here, 20,000 entries and 65,000 real-time images are included in the sensor data. The collection of gathered sensor data and images can be stored in a maximum of 14.2 GB. The simulator processes between 5 and 30 categorization instances. A CPU running at 3 Hz, 4 × 4 GB of physical RAM, and 250 GB of space for local storage are used in this experiment. The data are prepared to calculate the classification, processing time, and index. MMFM is evaluated

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.7110
***Article Received**: 20 March 2023 **Revised**: 18 May 2023 **Accepted**: 30 May 2023*

to the previously stated p-Cloud, B-DMS, and H-DM to make it simple to confirm reliability.

Different records affect the processing time of efficient data processing instances. Analysis of processing time is shown in Fig. 2 to Fig 10. MMFM first builds $\mathbb{R}$ and chooses k for data administration. According to (1) and (2), the processed data satisfies the real relationship. This identification helps classify d. The detecting method β uses i and j to separate the assimilation above. These verifications and d estimate R and categorize the data as overlapping or non-overlapping. MMFM only enhances classified d analysis, unlike the other techniques' joint relationship and data validation processes. Compared to other methods, this independent study and attribute classification is quicker.

The processing durations for various records and data sizes are examined in Fig 12 to Fig 15. The period of processing time includes similarity verification and categorization. As was already mentioned, this technique takes less time for classification. By utilizing α and β, the objective of increasing data is realized in the similarity proof case. The classification of borders using the proposed MMFM aids in processing accurate data while excluding d. Then, selecting best processing instance for different records and data quantities is necessary. It is easy to complete these groups. As a result, it is simpler to determine other related data. It contains the sensor data and the statistics that were used. Table 1 displays a contrast of processing rates for various data sizes. The Table above shows that, regardless of the data size, the recommended MMFM requires less computation time. Different data sizes are used uniformly to verify the relationship and constraints simply.

Table 1 Processing time comparison in seconds

Data size (GB)	P-cloud	B-DBMS	H-DM	F-DM	MMFM
2	151	134	115	78	70
4	256	193	173	86	74
6	269	215	185	89	75
8	274	218	198	92	89

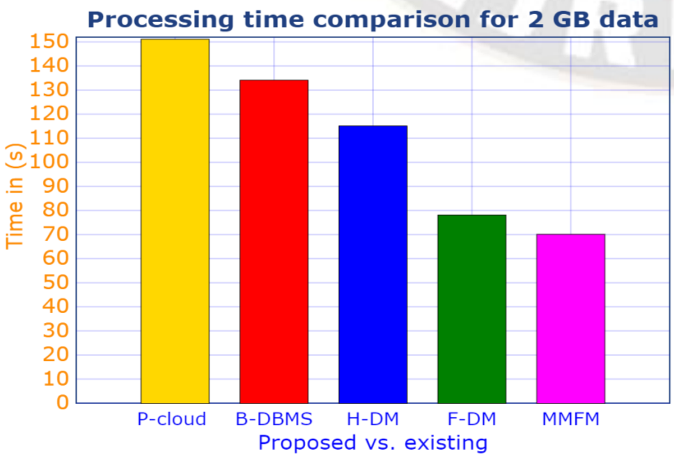

Fig 2 Processing time comparison with 2GB data

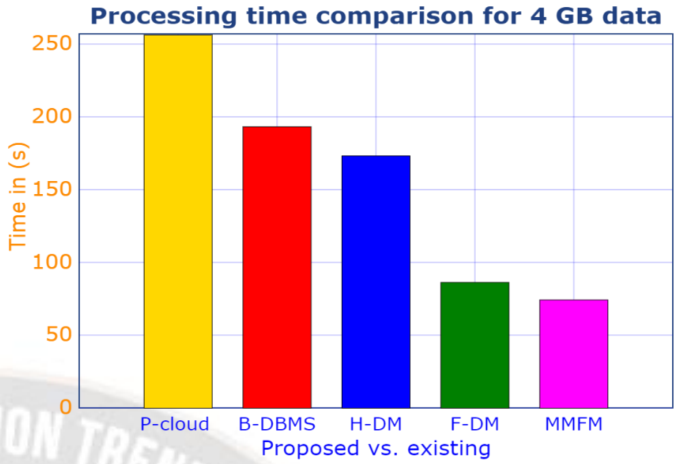

Fig 3 Processing time comparison with 4GB data

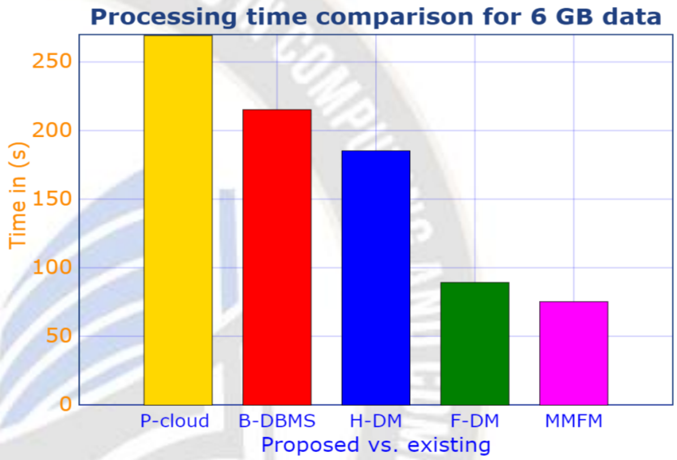

Fig 4 Processing time comparison with 6GB data

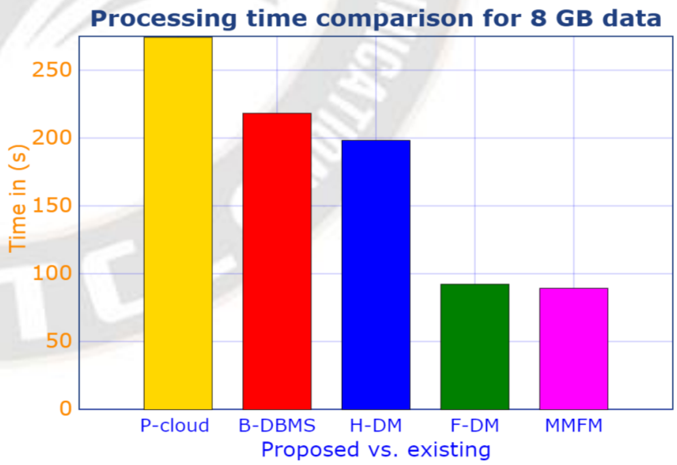

Fig 5 Processing time comparison with 8GB data

The number of classification instances needed rises as data size changes. Regardless of the size of the data, Instances of categorization are required to speed up processing. The quantity of processed data indirectly reduces the analysis's intricacy. When overlapping and non-overlapping clusters are used, the processing instances acquire the defined value. This case representation is classified using a learning-based

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.7110
***Article Received**: 20 March 2023 **Revised**: 18 May 2023 **Accepted**: 30 May 2023*

generalization. The goal is to choose data jointly or individually from the overlapping and non-overlapping groups. R has improved *I* by processing the similarity index and confirming that x_i and y_i satisfy a certain condition.

Consequently, the data classification using the proposed model is universally successful for all types of data and records. This results in processing taking less time during the validated data interval. A comparison of the processing periods is shown in Table [] for the different classification instances. Results of Table 2 comparison for Processing time demonstrate that MMFM achieves a lower processing time in both small and large classification instances. It is due to the early step of relationship and constraint classification.

Table 2 Processing time comparison based on various samples

Samples	P-cloud	B-DBMS	H-DM	F-DM	MMFM
5	184	167	123	109	100
10	184	187	142	111	105
15	196	202	185	114	110
20	214	209	198	116	112
25	236	220	199	121	115
30	242	224	201	141	135

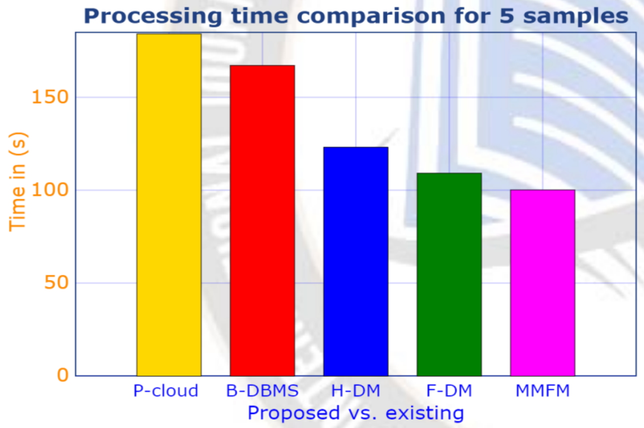

Fig 6 Processing time comparison for 5 samples

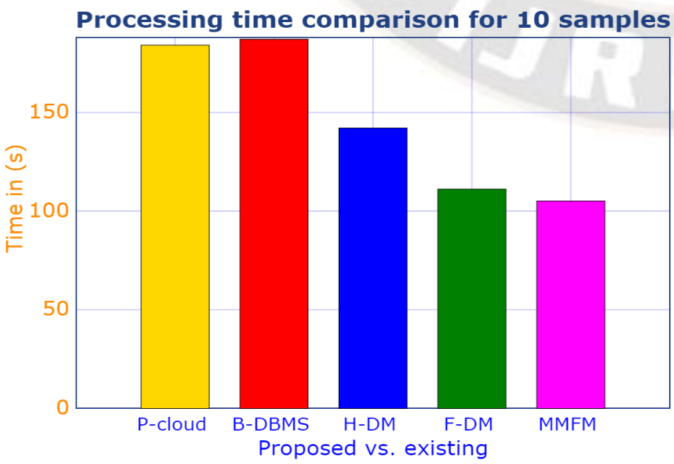

Fig 7 Processing time comparison for 10 samples

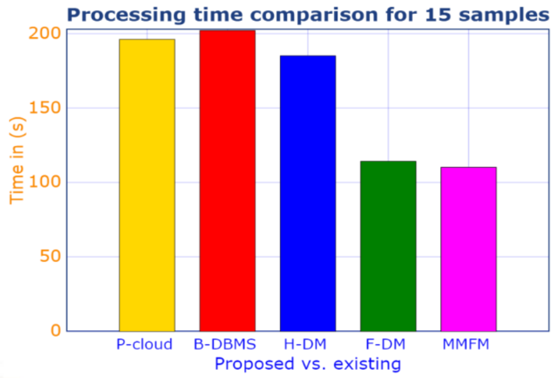

Fig 8 Processing time comparison for 15 samples

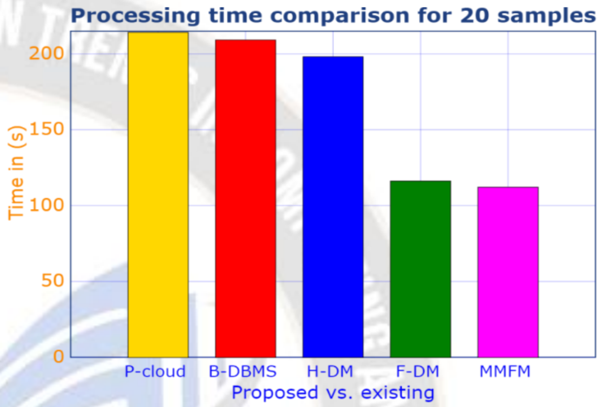

Fig 9 Processing time comparison for 20 samples

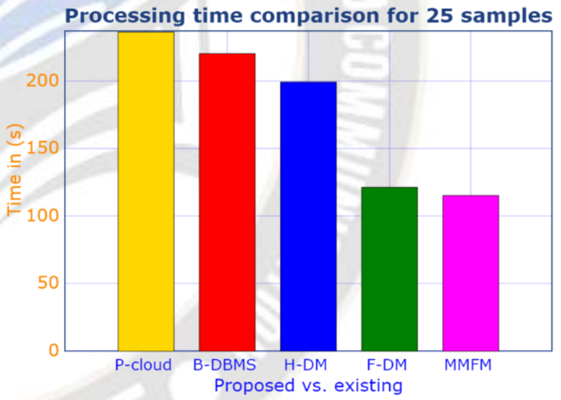

Fig 10 Processing time comparison for 25 samples

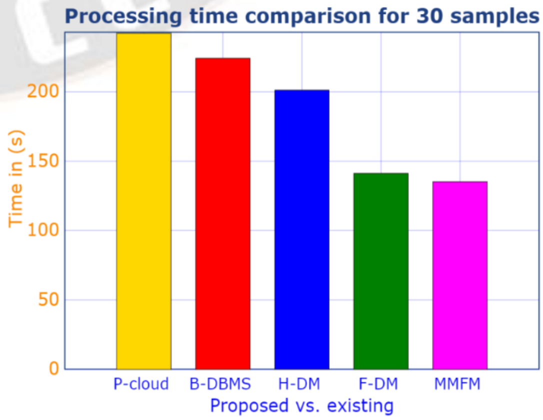

Fig 11 Processing time comparison for 30 samples

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.7110
***Article Received**: 20 March 2023 **Revised**: 18 May 2023 **Accepted**: 30 May 2023*

The effects of different records and classification cases are examined to confirm the processed d's similarity index. The similarity score is evaluated in this instance on two distinct scales. A threshold is first established to compare with the conserved d as efficiently as possible. It pertains to every CD song. The proposed is the earliest method of identifying the variance in d. Using $\propto$ and β correspondingly, they found data boundaries assessed during the generalization process. All classification instances handle the resulting $\propto$ and β using estimation. These conditions dictate whether $d \subseteq R$ or $d \subseteq k$ satisfies S_d. The input semantics verification uses the pre-classification of $d \subseteq \{\emptyset\}$ to independently identify the restrictions of $[P(R) \cap P(k)]$. The chosen metrics aid in reducing processing time and negative effects on the complete data collection.

Table 3 Processing time comparison based on various samples

Records	P-cloud	B-DBMS	H-DM	F-DM	MMFM
5000	0.84	0.86	0.92	0.98	0.99
10000	0.79	0.82	0.89	0.97	0.98
15000	0.63	0.75	0.87	0.95	0.96
20000	0.69	0.82	0.88	0.92	0.93

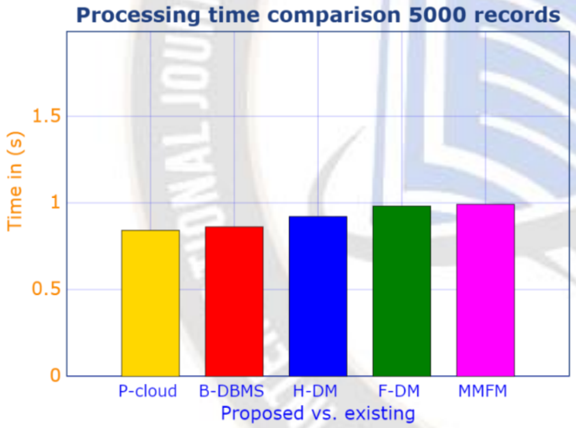

Fig 12 Processing time comparison for 5000 records

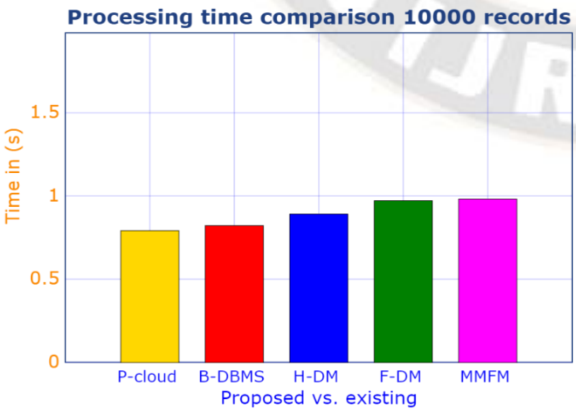

Fig 13 Processing time comparison for 10000 records

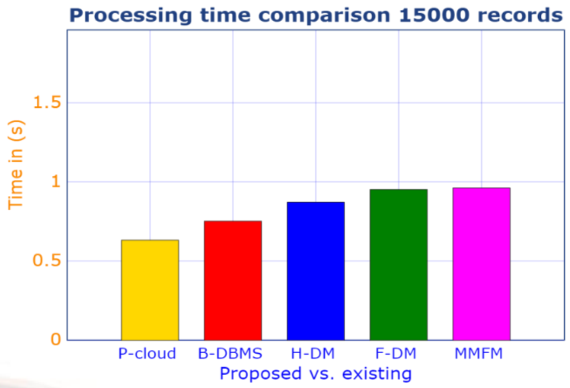

Fig 14 Processing time comparison for 15000 records

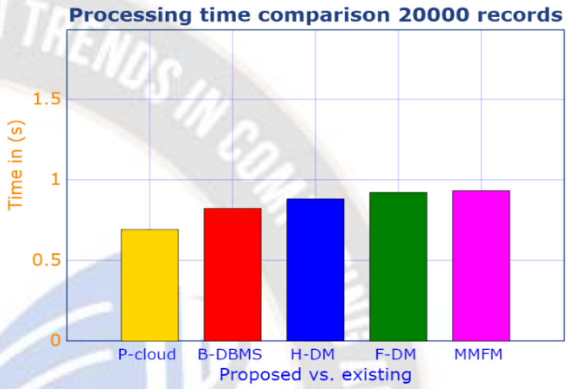

Fig 15 Processing time comparison for 20000 records

Consequently, when mapping d to related variables, there are noticeably fewer errors, enabling the greatest attainment of S_d. The ratings of the comparative similarity measure for every entry are shown in Table 3. Table 3 depicts that MMFM constantly outperforms in various data and classification scenarios. By changing the solution at various levels, the similarity is increased.

V. CONCLUSION

The MMFM technique was covered in this piece as a way to enhance analytics using huge amounts of data. The sensed data were classified using the relationship and constraints according to the semantics of the incoming attributes. The info was combined into a singular entity by comparing the classification data's learning generalization-based similarity scores. By creating classification boundaries for the incoming data, generalization influences several characteristics. With superior processing time, the samples from input data speed up the processing time. The suggested method's performance assessment showed that it carried out data analysis to the best of its ability resulting in 70 seconds of processing time for 2GB data and 0.99 seconds while handling 5000 records for various cases of categorization. However, the sensitivity characteristic of the accumulated data must be taken into account by the suggested model MMFM, requiring longer processing times that will be reduced by using hybrid data classification models.

REFERENCES

[1] Wang F, Zhang P, Wang X, Hu J. Clinical risk prediction by exploring high-order feature correlations. AMIA Annu Symp Proc. 2014;2014:1170–9.

[2] Vie LL, Scheier LM, Lester PB, Ho TE, Labarthe DR, Seligman MEP. The US army person-event data environment: a military-civilian big data enterprise. Big Data. 2015;3(2):67–79.

[3] Luo J, Wu M, Gopukumar D, Zhao Y. Big data application in biomedical research and health care: a literature review. Biomed Inform Insights. 2016;8:1–10.

[4] Obermeyer Z, Emanuel EJ. Predicting the future - big data, machine learning, and clinical medicine. N Engl J Med. 2016;375(13):1216–9.

[5] Doll KM, Rademaker A, Sosa JA. A practical guide to surgical data sets: surveillance, epidemiology, and results (SEER) database. JAMA Surg. 2018;153(6):588–9.

[6] Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3: 160035.

[7] Unnam, A. K. ., & Rao, B. S. . (2023). An Extended Clusters Assessment Method with the Multi-Viewpoints for Effective Visualization of Data Partitions. International Journal of Intelligent Systems and Applications in Engineering, 11(1s), 51–56. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2476.

[8] Bhatti, M. I. U. Lali, B. Shahzad, F. Javid, F. U. Mangla, and M. Ramzan, “Leveraging the big data produced by the network to take intelligent decisions on flow management,” IEEE Access, vol. 6, pp. 12197–12205, Feb. 2018.

[9] J. Chen, Z. Lv, and H. Song, “Design of personnel big data management system based on blockchain,” Future Gener. Comput. Syst., vol. 101, pp. 1122–1129, 2019

[10] Cozar, A. Fernandez, F. Herrera, and J. A. Gamez, “A hierarchical meta rule decision system to design robust fuzzy classifiers based on data complexity," IEEE Trans. Fuzzy Syst., vol. 27, no. 4, pp. 701–715, Apr. 2019.

[11] Elaggoune, R. Maamri, and I. Boussebough, “A fuzzy agent approach for smart data extraction in big data environments,” J. King Saud Univ. Comput. Inf. Sci., vol. 32, pp. 465–478, 2019.

[12] Goudarzi, “Heterogeneous architectures for big data batch processing in MapReduce paradigm,” IEEE Trans. Big Data, vol. 5, no. 1, pp. 18–33, Jan. 2019

[13] Hernández, E. Zamora, H. Sossa, G. Téllez, and F. Furlán, “Hybrid neural networks for big data classification,” Neurocomput., vol. 390, pp. 327–340, 2019.

[14] Kozjek, R. Vrabiˇc, B. Rihtaršiˇc, and P. Butala, “Big data analytics for operations management in engineer-to-order manufacturing,” Procedia CIRP, vol. 72, pp. 209–214, 2018.

[15] Sherubha, "Graph-Based Event Measurement for Analyzing Distributed Anomalies in Sensor Networks", Sådhanå(Springer), 45:212, https://doi.org/10.1007/s12046-020-01451-w

[16] Stojanovic, N. . (2020). Deep Learning Technique-Based 3d Lung Image-Based Tumor Detection Using segmentation and Classification. Research Journal of Computer Systems and Engineering, 1(2), 13:19. Retrieved from https://technicaljournals.org/RJCSE/index.php/journal/article/view/6

[17] Sherubha, “An Efficient Network Threat Detection and Classification Method using ANP-MVPS Algorithm in Wireless Sensor Networks”, International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Volume-8 Issue-11, September 2019

[18] Sherubha, “An Efficient Intrusion Detection and Authentication Mechanism for Detecting Clone Attack in Wireless Sensor Networks”, Journal of Advanced Research in Dynamical and Control Systems (JARDCS), Volume 11, issue 5, Pg No. 55-68

[19] Padillo, J. M. Luna, and S. Ventura, "A grammar-guided genetic programming algorithm for associative classification in big data," Cogn. Comput., vol. 11, no. 3, pp. 331–346, 2019.

[20] Degadwala, D. S. ., & Vyas, D. . (2021). Data Mining Approach for Amino Acid Sequence Classification . International Journal of New Practices in Management and Engineering, 10(04), 01–08. https://doi.org/10.17762/ijnpme.v10i04.124.

[21] Q. T. Tran, S. D. Nguyen, and T. Seo, “Algorithm for estimating online bearing fault upon the ability to extract meaningful information from big data of intelligent structures,” IEEE Trans. Ind. Electron., vol. 66, no. 5, pp. 3804–3813, May 2019

[22] Yang and T. Chen, “Analysis and visualization implementation of medical big data resource sharing mechanism based on deep learning,” IEEE Access, vol. 7, pp. 156077–156088, Oct. 2019

[23] C. Zhao, L. Ren, Z. Zhang, and Z. Meng, “Master data management for manufacturing big data: A method of evaluation for data network,” World Wide Web, vol. 23, pp. 1407–1421, 2019.

[24] Inkinen HT, Kianto A, Vanhala M. Knowledge management practices and innovation performance in Finland. Baltic J Manage. 2015;10(4):432–55.

[25] Steven Martin, Thomas Wood, María Fernández, Maria Hernandez, .María García. Machine Learning for Educational Robotics and Programming. Kuwait Journal of Machine Learning, 2(2). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/179.

[26] Rush, S. Chopra, and J. Weston, A neural attention model for abstractive sentence summarization, in Proc. 2015 Conf. Empirical Methods in Natural Language Processing, Lisbon, Portugal, 2015, pp. 379–389.

[27] Wang, J. Li, S. S. Li, Y. Y. Kang, M. Zhang, L. Si, and G. D. Zhou, Aspect sentiment classification with both word-level and clause-level attention networks, in Proc. 27th Int. Joint Conf. Artificial Intelligence, Stockholm, Sweden, 2018, pp. 4439–4445.

[28] Li, Z. Zhao, R. F. Hu, W. S. Li, T. Liu, and X. Y. Du, Analogical reasoning on Chinese morphological and semantic relations, in Proc. 56th Annu. Meeting of the Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 138–143.

[29] Tang, B. Qin, X. C. Feng, and T. Liu, Effective LSTMs for target-dependent sentiment classification, in Proc. 26th Int. Conf. Computational Linguistics: Technical Papers, Osaka, Japan, 2016, pp. 3298–3307.

[30] Dhablia, A. (2021). Integrated Sentimental Analysis with Machine Learning Model to Evaluate the Review of Viewers. Machine Learning Applications in Engineering Education and Management, 1(2), 07–12. Retrieved from http://yashikajournals.com/index.php/mlaeem/article/view/12.

[31] Bouazizi and T. Ohtsuki, Multi-class sentiment analysis on Twitter: Classification performance and challenges, Big Data Mining and Analytics, vol. 2, no. 3, pp. 181–194, 2019
[32] Kude T, Hoehle H, Sykes TA. Big data breaches and customer compensation strategies: personality traits and social influence as antecedents of perceived compensation. Int J Oper Prod Manage. 2017;37(1):56–74.
[33] Isabella Rossi, Reinforcement Learning for Resource Allocation in Cloud Computing , Machine Learning Applications Conference Proceedings, Vol 1 2021.
[34] Sun J, et al. Leverage RAF to fnd domain experts on research social network services: a big data analytics methodology with MapReduce framework. Int J Prod Econ. 2015;165:185–93.
[35] Costa V, Monteiro S. Key knowledge management processes for innovation: a systematic literature review. VINE J Inf Knowl Manage Syst. 2016;46(3):386–410.
[36] Akter S, et al. How can big data analytics capability and business strategy alignment improve firm performance? Int J Prod Econ. 2016;182:113–31.
[37] Braganza A, et al. Resource management in big data initiatives: processes and dynamic capabilities. J Bus Res. 2017;70:328–37