Detecting popular subjects and optimizing microblog performance using the Hybrid Hadoop Framework

Dr. A. Ramathilagam[1*], G. Prem Nisha[2], Dr. Sandra Johnson[3], Dr. B. Prathusha Laxmi[4], Dr. V. Vijayaraja[5], Dr. Y. Harold Robinson[6], N. Raghavendran[7], R. Saravanakumar[8]

[1*]Professor, Department of Computer Science and Engineering, P.S.R Engineering College, Sivakasi, Tamil Nadu, India, thilla2012@gmail.com,
[2]Assistant Professor, Department of Computer Science and Engineering, Amrita College of Engineering and Technology, Nagercoil, Tamil Nadu, India, nishainfo92@gmail.com,
[3]Professor, Department of Artificial Intelligence and Data Science, R.M.K. Engineering College, Chennai, Tamil Nadu, India, sjn.ad@rmkec.ac.in,
[4]Professor, Department of Artificial Intelligence and Data Science, R.M.K. College of Engineering and Technology, Chennai, Tamil Nadu, India, hod_ads@rmkcet.ac.in
[5]Professor, Department of Artificial Intelligence and Data Science, R.M.K. College of Engineering and Technology, Chennai, Tamil Nadu, India, vijayarajaads@rmkcet.ac.in
[6]Professor, Department of Computer Science and Engineering, Francis Xavier Engineering College, Tirunelveli, Tamil Nadu, India, yhrobinphd@gmail.com
[7]Assistant Professor, Department of Artificial intelligence and Date Science, RMK College of Engineering and Technology, Chennai, Tamil Nadu, India, ragavendrannv2001@gmail.com
[8]Software Analyst, Iconix Software Solution, Tirunelveli, Tamil Nadu, India,iconixsaro@gmail.com

KEYWORDS
Social Media Networks, Big Data, Hadoop, Mapreduce, Cloud Systems

ABSTRACT
Active users of social media networks all over may dynamically create intolerable stuff. Big data helps to sustain the enormous volume of information on social media channels. Big data is advanced on Hadoop-based cloud systems using its fault-tolerance and dependability. Hadoop is the basic platform for big data analytics. Using Hadoop has a main disadvantage in terms of handling the enormous number of configuration metrics handling. Driven by cloud-based Apache Spark, the hybrid Hadoop Framework is proposed in this paper to enhance big data processing by means of key parameter regulation including workload, response time, network bandwidth, and the hot topic detection mechanism especially tailored for the microblog into the big data. To manage the big volumes, we deliberately construct the MapReduce jobs to precisely identify hot subjects. According to the experimental results, the proposed system's accuracy is quite high when compared to related methods.

1. INTRODUCTION

With iteration, the optimal workloads of Hadoop provide performance metrics that enable the prediction in the acquired dataset as big data related functionality having the main influence in the company [1]. Big data related techniques have many effects on the filtering and analysis of the Hadoop environment that accelerate the functioning of vast data into the centralized framework in data processing [2]. Data quantities are expanding faster than the speed at which the techniques used to handle that data can develop. Right now, there isn't any such technical tool capable of doing the given task. Still, the rise of big data problems has sped up the trend under study [3]. Comparatively to the processing time, the waiting time of the slave nodes in which they get the information from the master node is enormous. The data in the repository is handled equally independent of the demand on the social network forum. Not all the information need to be handled simultaneously. While some data calls for significant attention, others not. We propose an Enhanced Hadoop Framework to help to avoid the following risks by increasing the MapReduce Model's performance. We propose a new method based on the attention level the data acquired during a certain time span "t" falls under. TaskTrackers prioritize the data that reaches the attention_level over those haven't reached the attention_level thereby controlling the reaction time and therefore improving the general performance of the MapReduce Model. Microblog also helps much in processing heated issues as it is very sympathetic to them. Every time the enormous number of users actively engaged in sending and reporting the undesired events, they would find their way to social media. Therefore, the microblog has become the major tool available for individuals to save and analyze the data; every single minute, numerous kinds of issues are published on many micro blogs. The primary contributions of the article are,

- Big data analytics employing the Twitter dataset helps to identify buzz and non-buzz.
- Improved data flow rate, task distribution, and workload balancing are achieved using MapReduce technology.
- Designed to separate buzz and non-buzz materials from the Hadoop cluster in HDFS system, hybrid Hadoop algorithm is created.
- The execution time, general process execution time, Hadoop cluster creation is investigated in the performance assessment and the suggested approach is functioning well in all the criteria.

2. RELATED WORK

Due to the highly competitive nature of banking operations and the abundance of large, mostly unstructured data sets, big data is essential in this industry [4]. As required by the standard, many firms face the difficulty of integrating many data formats, such as spreadsheets, web sources, XML, and traditional DBMSs [5]. A leader in the big data space, the Apache Hadoop data management platform was created especially to handle the volume, variety, and speed of unstructured data generated by several businesses [6]. The framework comes to the conclusion that the unstructured form of the data and the frequent disconnections of new data from files are causing a considerable amount of important data to be lost [7].

According to the most recent framework, managing unstructured data has become more challenging in a number of domains [8]. A new research area in big data analysis has emerged as a consequence of the data showing that integrating large amounts of relational and unstructured data is very difficult [9]. NoSQL databases are becoming synonymous with big data. However, NoSQL has additional drawbacks, such as not supporting SQL, transactions, reporting, and other extra functionalities [10]. Researchers and analysts create enormous amounts of data (such as experiments) every day in fields like high-energy physics, chemistry, engineering, bioinformatics, etc., yet it is almost impossible to extract useful information with RDBMS [11]. The problem is getting useful information out of it [12]. Despite this, businesses leverage the big data concept to make data monitoring easier by using certain mathematical frameworks [13]. The problem of adapting real models and calculations to distinct multidimensional data configurations is still unsolved, despite the extensive recent research on huge information archives [14, 15].

Even with the significant increase in data storage capacity and the growth of data centers throughout the world, it is still very difficult to efficiently confine and aggregate massive data sets, which prevents simple accessibility. It is crucial to differentiate MapReduce from other information transformation frameworks in order to improve understanding of its proper use [16]. It has been noted that traditional information management is not rendered worthless by MapReduce systems [17]. There is a noticeable pattern in the growing number of applications that can handle large amounts of data [18]. Google MapReduce and its open-source cousin Hadoop, for instance, are important tools for the creation of several applications [19]. Topic identification has been covered by a number of well-established techniques, all of which are intended to quickly identify themes and popular topics utilizing a range of content-based temporal criteria. Using the co-occurrence model in a large data setting, the Biterm Topic Model (BTM) [20] has been utilized to find topics within the text. Microblog topic identification from the large number of internet tweets has been accomplished using the topic-based Latent Dirichlet Allocation (TC-LDA) [21]. In order to handle the massive amount of databases employing MapReduce processes for accurate and effective hot topic identification, the parallel 2-phase hot topic detection (TMHTD) [22] has been developed.

3. PROPOSED WORK

The Twitter dataset, including many instances for each observation related to particular topic identification, enables the proposed work to enhance workload management by optimizing response time during task allocation in the MapReduce framework. This is seen in Figure 1.

The objective of the Hybrid Hadoop Framework is to sustain social media analytics to identify socially relevant information, such as tweets. Relevant computing resources are necessary to ascertain the volume of data that is routinely archived. The public's opinion is evaluated to examine the statistics. Social media platforms are used to augment internet marketing and facilitate data modeling technology. The Hybrid Hadoop Framework advocates for the use of data modeling and toll-based data analysis. Twitter utilizes APIs that enable users to limit data for real-time processing. The proposed system may collect up to 100 tweets every second. The whole of Twitter data is systematically structured and maintained via a database management system. A specific file location and suitable indexing techniques are used to store the data acquired from the Tweeters. Daily, the data is maintained in an easily accessible file format. A specialized kind of database is used to execute the data processing activities. The MapReduce method eliminates redundant data. In a distributed architecture, the inquiry is addressed via many parameters.

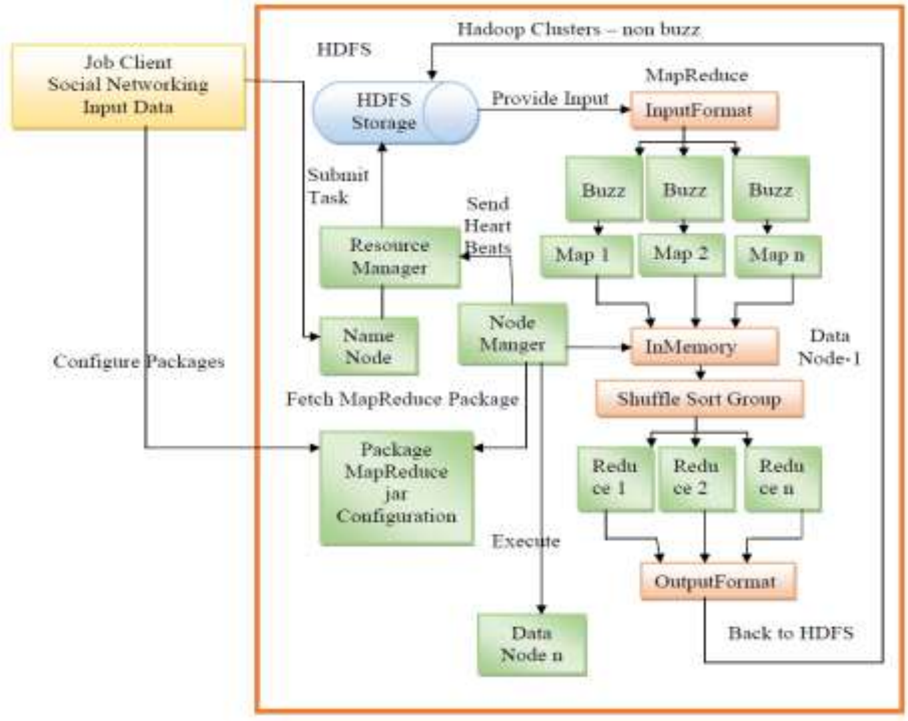

Fig. 1 Proposed approach for the Hadoop social media system

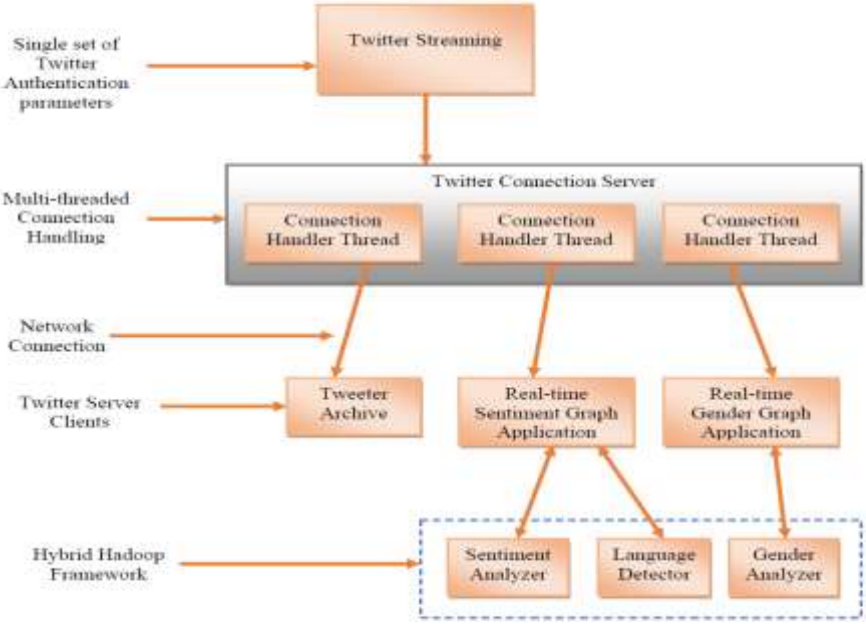
Single set of Twitter Authentication parameters
Twitter Streaming
Multi-threaded Connection Handling
Twitter Connection Server
Connection Handler Thread
Connection Handler Thread
Connection Handler Thread
Network Connection
Twitter Server Clients
Tweeter Archive
Real-time Sentiment Graph Application
Real-time Gender Graph Application
Hybrid Hadoop Framework
Sentiment Analyzer
Language Detector
Gender Analyzer

Fig. 2 Streaming on Twitter using the Hybrid Hadoop Framework

The Hybrid Hadoop Framework for Twitter Streaming is shown in Fig. 2. The network connection is created to link the clients to the server using Tweets, the Twitter authentication parameters are used to supply the input for Twitter Streaming APIs, and the Multi-threaded Connection Handling is used to frame the Twitter Connection Server with the individual Connection Handler Thread. Based on the specific Tweet from the Connection Handler thread, the information is gathered using the Tweet archive, and the created data is saved in the system immediately. Sentiment analysis, language recognition, and gender analysis are all performed using the Real-time Sentiment Graph-based application. Our approach is to use Eq. (1) to calculate the weights w such that it will meet the criterion.

$$\min_{w} \sum_{i}^{n} \varepsilon_i^2 \qquad (1)$$

The solution of the method of multiple linear regressions is obtained by the least squares methodology. The Decision algorithm is represented using Eq. (2)

$$\sum_{i=1}^{N_D} \left(1 - \sum_{j=0}^{p} w_i x_{ij}\right)^2 + \sum_{i=N_D+1}^{N_D+N_{ND}} \left(\sum_{j=0}^{p} w_i x_{ij}\right)^2 \qquad (2)$$

where $x_{iD} = 1, if\ i = 1,2, \dots, n$. This methodology is implemented in matrix using Eq. (3)

$$\begin{pmatrix} 1_D & X_D \\ 1_{ND} & X_{ND} \end{pmatrix} \begin{pmatrix} w_0 \\ w^T \end{pmatrix} = \begin{pmatrix} I_D \\ 0 \end{pmatrix} \qquad (3)$$

The matrix format is modified using Eq. (4)

$$Xw^T = y^T \qquad (4)$$

The value of X can be computed using Eq. (5)

$$X = \begin{pmatrix} 1_D & X_D \\ 1_{ND} & X_{ND} \end{pmatrix} \qquad (5)$$

If the matrix size is $N_D * (p+1)$ then the value of $X_D$is computed using Eq. (6)

$$X_D = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{ND1} & \cdots & x_{NDp} \end{pmatrix} \qquad (6)$$

If the matrix size is $nN_D * p$ then the value of $X_D$is computed using Eq. (7)

$$X_D = \begin{pmatrix} x_{ND+11} & \cdots & x_{ND+1p} \\ \vdots & \ddots & \vdots \\ x_{ND+ND1} & \cdots & x_{ND+NDp} \end{pmatrix} \quad (7)$$

The matrix form of minimization task is represented in Eq. (8)

$$\min(Xw^T - y^T)^T (Xw^T - y^T) \quad (8)$$

The derivation and equate to 0 is computed using Eq. (9)

$$X^T X w^T = X^T y^T \quad (9)$$

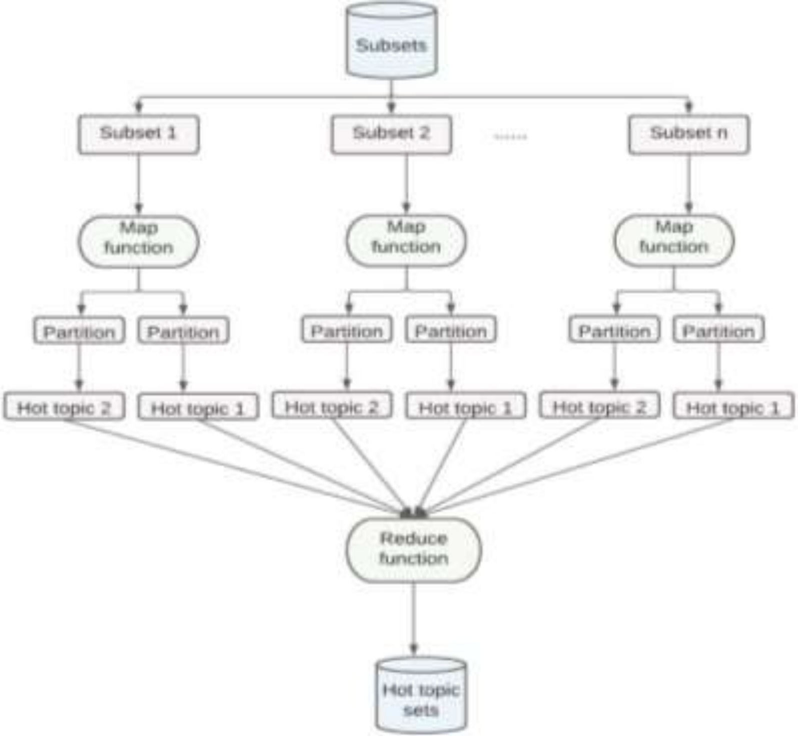

Fig. 3 Hot topics detection framework

The result of reducing the input values is minimized using the MapReduce algorithm. It offers the ability to identify microtopics. Fig. shows the procedure by which the subsets for the Spark platform have been allocated. Within the allotted time, n hot topics are produced in each map function during the map phase. The partition phase computes the produced hot topics 1 and the decreased hot topics 2. The method is used to calculate the hot topics 1 and 2 during the reduction phase. Figure 3 shows the full procedure.

Algorithm – Hot topic detection algorithm

```
Begin Procedure Hot_topic_detection ()
    Subsets as the input for detection
    for every subset from 1 to n
        Map subsets as $< i, subset_i >$
        Generate the hot topics from $< i, subset_i >$
        Produce hot topic1 ← hot topic set
    end for
    Combine the topic with reduce function
    return hot topic2
End Procedure
```

4. PERFORMANCE EVALUATION

Because the large size of blocks and the relocated time of data govern the entire process time and the IO is a key factor in the execution time, the distribution functionality is implemented to evaluate the performance for the utilization of the IO devices in a data-centric application. The enormous amount can be limited for delays. While it is needed, MapReduce simply handles the jobs of analyzing the nodes and the load data from limited areas in the remote disk. It is necessary to categorize the workloads in order to avoid the initial time. Performance will decrease as waiting times grow because the CPU does not receive the most recent job until the remaining resources have been stored in memory. A cluster including several nodes with varying data sizes is used for the experiment; table 1 shows the experimental setup and performance measure.

Table 1 Experimental setup

Parameter	**Value**
Environment	Spark 1.1.0
Cluster	1 master, 100 slaves
Processor	Pentium Dual-Core 3.20 GHz
Memory	32 GB
Operating System	Ubuntu 12.04.4
Total Microblog data	16 million

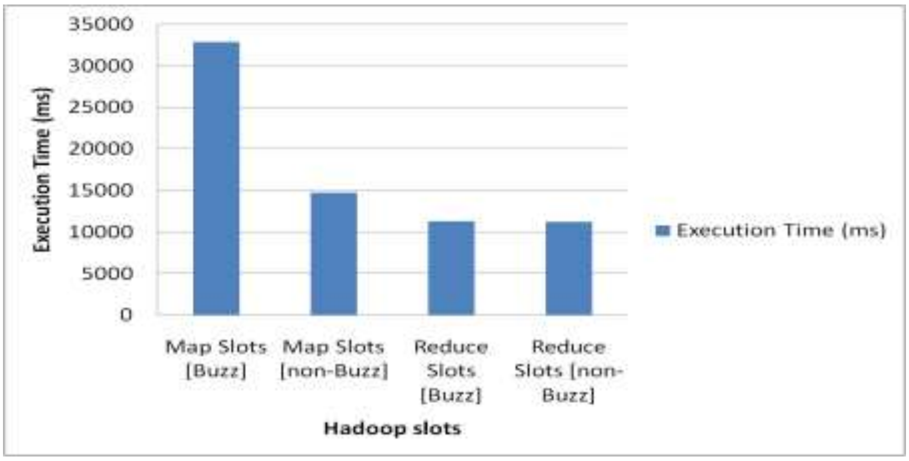

Fig. 4 Execution of map and reduce slots in Hadoop

This reduction shows that the input and output metrics of map and reduce tasks are managed, and it significantly reduces the response time. Fig. 4 shows the total number of MapReduce slots assigned for the dataset with an execution period of 32831 ms and 11312 ms, while the total number of MapReduce slots assigned for the data with an execution period of 14742 ms and 11221 ms.

The findings showed that the suggested approach has the least amount of runtime when compared to other strategies. Fig. 5 shows the runtime for a limited number of samples.

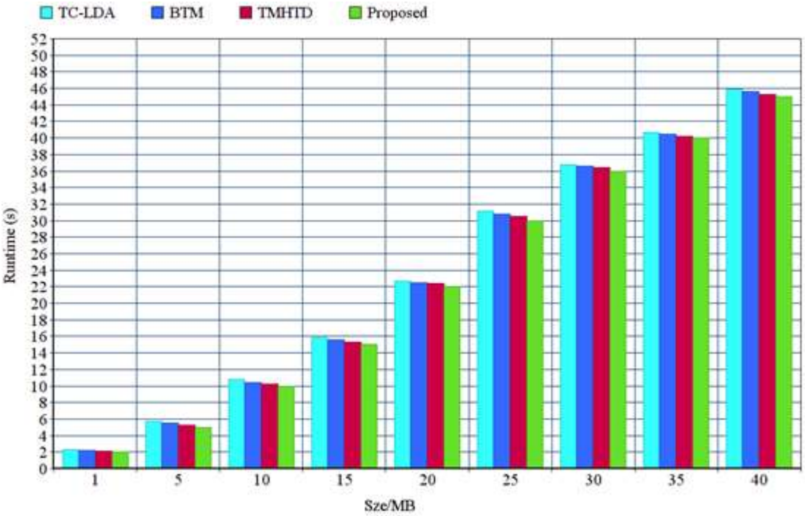

Fig. 5 Runtime for small amount of datasets

Figure 6 shows that the suggested strategy has a shorter runtime than similar techniques for a large number of samples. Every stage of the procedure takes more time since the dataset is so large. The suggested method is designed to execute the job dynamically in order to balance task management.

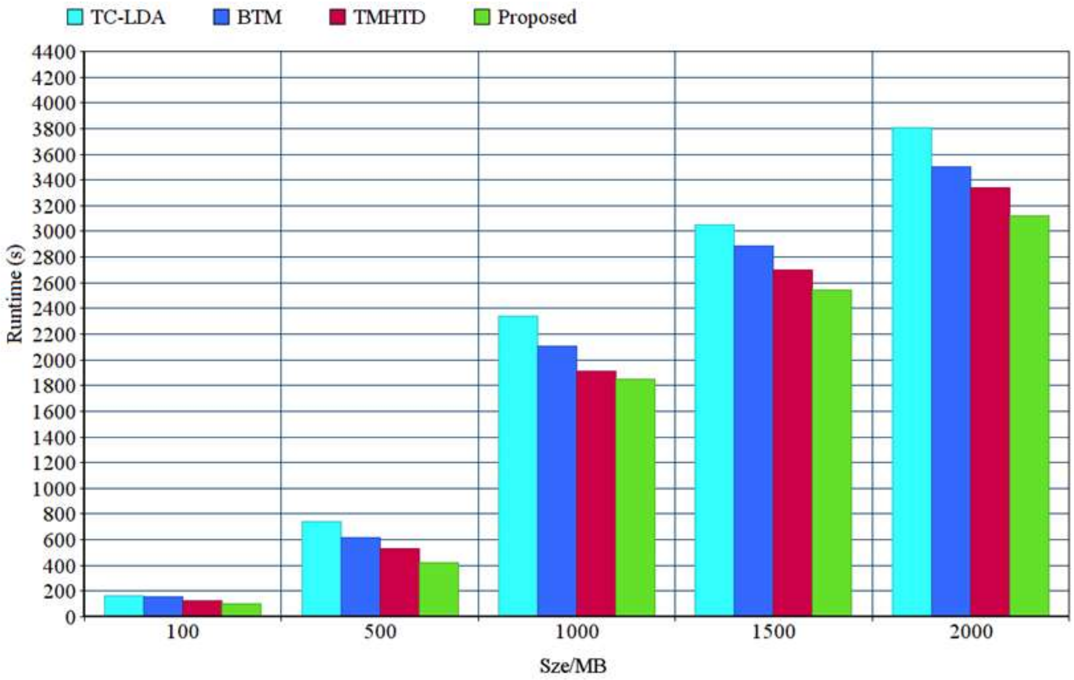

Fig. 6 Runtime for large amount of datasets

The accuracy quantification is measured using the $F_{measure}$ value in Eq. (10).

$$F_{measure} = \frac{(1+\gamma^2)\,(pr+re)}{\gamma^2} \qquad (10)$$

where pr is the exact detection precision value, re is the successfully hot topic detection, γ is the value within the pr and re. The accuracy of the suggested method has been assessed by contrasting it with comparable methods. Using varying dataset sizes, the tests have been conducted using analogous methodologies. Using clusters with different datasets, Spark 1.1.0 is implemented; the experimental outcome is shown in Fig. 7. The $F_{measure}$ value for the proposed technique is higher than the related techniques. Whenever the data size is bigger, the accuracy has been reduced accordingly.

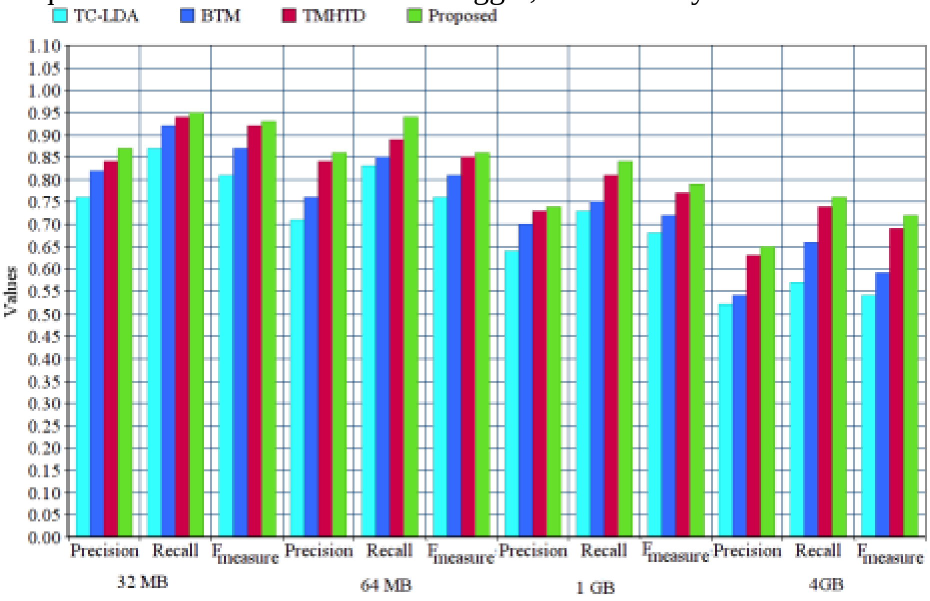

Fig. 7 Accuracy comparison from various datasets

Figure 8 shows that, in comparison to the pertinent methodologies, the suggested methodology has achieved improved hot topic identification.

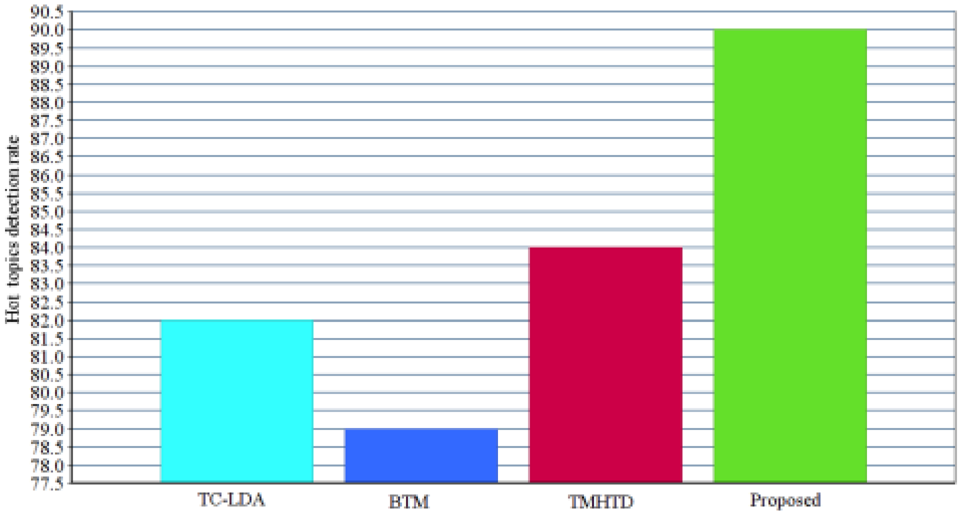

Fig. 8 Hot topics detection rate

5. CONCLUSION

Big data is used to preserve vast amounts of information and to provide fault tolerance and dependability. Semantic data may be dynamically created by social media networks. Using key indicators like cloud, response time, and bandwidth for hot topic discovery, the Hybrid Hadoop Framework has been used to improve data processing. When using MapReduce with the Hadoop setup, the Hadoop framework's performance enhancement may operate well in Hadoop-related clusters from the Twitter dataset. By enhancing all performance measures, the testing findings demonstrated that the suggested approach has the ability to improve performance with multi-node clusters.

REFERENCES

[1] Oussous, A., Benjelloun, F.Z., Ait Lahcen, A., Belfkih, S.: Big data technologies: a survey. J. King Saud Univ. - Comput. Inf. Sci. (2017).

[2] M. Shakarami, I.F. Davoudkhani, Wide-area power system stabilizer design based on grey wolf optimization algorithm considering the time delay, Electr. Power Syst. Res. 133 (2016) 149–159.

[3] Susmitha, K., Jayaprada, S.: Smart cities using big data Analytics. Int. Res. J. Eng. Technol. (IRJET), p. 1617 (2017).

[4] Hamid, R., Xiao, Y., Gittens, A., DeCoste, D.: Compact random feature maps. In: Proceedings of the 31st International Conference on Machine Learning, vol. 32, ser. ICML 2014. JMLR.org, pp. II–19–II–27 (2014).

[5] Eessaar, E., Saal, E.: Evaluation of different designs to represent missing information in SQL databases. In: Elleithy, K., Sobh, T. (eds.) Innovations and Advances in Computer, Information, Systems Sciences, and Engineering, pp. 173–187. Springer, New York (2013).

[6] Aggarwal Charu C, Zhai Cheng Xiang. Mining Text Data. Springer New York Dordrecht Heidelberg London: Springer Science+Business Media, LLC'12; 2012.

[7] J. Santoso, E. M. Yuniarno, M. Hariadi, Large scaletext classification using map reduce and naive bayes algorithm for domain specified ontology building, in:Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2015 7th International Conference on, Vol. 1,2015, pp. 428–432.

[8] M. Wasi-ur-Rahman et al., ``High-performance RDMA-based design of Hadoop MapReduce over InniBand," in Proc. IEEE 27th Int. Parallel Distrib. Process. Symp. Workshop Ph.D. Forum (IPDPSW), May 2013, pp. 1908-1917.

[9] Moura, J., Serrão, C.: Security and privacy issues of big data. In: Handbook of Research on Trends and Future Directions intelligence, vol. 2, pp. 20–52 (2015).

[10] Olofson, C.W., Vesset, D.: Big Data: Trends, Strategies, and SAP Technology, White Paper IDC, pp. 1–16, August 2012.

[11] Ibrahim, H., Victor, C., Nor Badrul, A., Kayode, A., Ibrar, Y., Abdullah, G., Ejaz, A., Haruna, C.: The role of big data in smart city, p. 2, University of Malaya (2016).

[12] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up?: sentiment classification using machine learning techniques,” in Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002, pp. 79–86.

[13] R. Pal, M. Saraswat, Data clustering using enhanced biogeography-based opti-mization, in: 2017 Tenth International Conference on Contemporary Comput-ing, IC3, IEEE, 2017, pp.1–6.

[14] Syed Thouheed Ahmed. (2017) “A Study on multi objective clustering techniques for medical datasets” In proceedings of International Conference on Intelligent Computing and Control Systems, pp 174-177, IEEE.

[15] Shinde, G., Deshmukh, S.N.: Sentiment TFIDF feature selection approach. Int. J. Comput. Commun. Eng. (2016)

[16] C.-M. Huang, C.-H. Shao, S.-Z. Xu, and H. Zhou, ``The social Internet of Thing (S-IOT)-based mobile group handoff architecture and schemes for proximity service," IEEE Trans. Emerg. Topics Comput., vol. 5, no. 3, pp. 425437, Jul./Sep. 2017.

[17] Zhou, Z., Jin, X.L., Vogel, D.R., Fang, Y., Chen, X., 2011. Individual motivations and demographic differences in social virtual world uses: an exploratory investigation in second life. Int. J. Inf. Manag. 31 (3), 261–271.

[18] Wang, G., Gunasekaran, A., Ngai, E. W., & Papadopoulos, T. (2016). Big data analytics in logistics and supply chain management: Certain investigations for research and applications. International Journal of Production Economics, 176, 98–110.

[19] Liang, Po-Wei, and Bi-Ru Dai. Opinion Mining on Social Media Data. Mobile Data Management (MDM), 2013 IEEE 14th International Conference on. Vol. 2. IEEE, 2013.

[20] X. Cheng, X. Yan, Y. Lan, and J. Guo. Btm: Topic modeling over short texts. IEEE Transactions on Knowledge and Data Engineering, 26(12):2928–2941, 2014.

[21] G. Ge, L. Chen, and J. Du. The research on topic detection of microblog based on tc-lda. In Communication Technology (ICCT), 2013 15th IEEE International Conference on, pages 722–727. IEEE, 2013.

[22] W. Ai and D. Li, "Parallelizing hot topic detection of microblog on spark," 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Changsha, 2016, pp. 1461-1468.