Webology (ISSN: 1735-188X)
Volume 18, Number 6, 2021

Extraction Of Medical Information Through Text Mining Method From Aco And Big Data

[1] K Manikandan , [2]Dr. Ramalingam sugumar

[1]Ph.D Research Scholar, PG & Research Department of Computer Science, Christhu Raj College (Affiliated to Bharathidasan University, Tiruchirappalli), Panjapur, Trichy, Tamilnadu.

[2]Professor & Director, PG & Research Department of Computer Science, Christhu Raj College (Affiliated to Bharathidasan University, Tiruchirappalli), Panjapur, Trichy, Tamilnadu.

Abstract

The rapid and precise exploration of texts through vast quantities of academic documents could reveal new information that could help us better understand human diseases and improve food detection, management and therapy. In this research, they built and implemented Text Mining (TM) on Big Data Analytics(BDA) that combines Apache Spark information broadcasting and computer learning technique of a NoSQL dataset. Euclidean method, Ant Colony Optimization (ACO) and multiple estimators were used to develop a prediction model from INSPEC to illustrate its effectiveness in categorizing cancer types.In the 29,437 full-text papers, ACO is able to predict a type of cancer with 93.81 percent accuracy.Text Mining took about 6 minutes to process the information, whereas rival other existing took more than 11 hours.

Keywords: Text mining; prediction; NoSQL database; biomedical research; Ant Colony Optimization

1. Introduction

Using technologies of convergence to Artificial Intelligence (AI), Deep Learning (DL), and database systems, data collection could be used to uncover diagrams in the large-scale dataset [1].TM would be a kind of data analysis that collects data via texts, such as scientific articles [2].By methodically evaluating many comprehensive scientific abstracts and papers, documentary exploration could produce new theories.Biomedical text analysis and its ramifications have been used to promote science in a range of biological applications, including malignity [3].BDA-based TM algorithms can evaluate biomedical articles based on Malignancy's study for convenience [4].

Recognize relevant words or expressions expressed in plain text or normalize them as in a relevant database. Identifiers were a TM application that would be widely used in research text analysis [5]. Gene and protein names, organisms, chemical substances, and ontology concepts like biochemical mechanisms, environmental variables, subatomic activities, illnesses, cellular structures, andphenotypes were examples of biological and biomedical sciences [6]. For Named Entity Recognition (NER), various computational methods and web applications were suggested.Automated TM systems such as STRING and STITCH integrate these additional expected or derived correlations from biological research.All these components could be accessed manually or automatically by means of packages, plugins or application programming interfaces [7].

Another often used method was functional enrichment assessment, which summarises a huge list of genes by evaluating to a pair of gene set descriptor aspectsrepresents a gene ontology aspect, biochemical source, nutrient component, illness, and so on [8]. The probability of significantly enhanced gene annotation groups is then compared to a reference set for identification [9].

2. Related Works

Automatic retrieval of important biological information is difficult due to the diversity of written resources [10].Consequently, word processing and AI methods are merged to extract biological pathways. Gene expression datasets in TM become a comprehensive discipline and specialization in biomedical sciences [11].There are a number of biomedical TM assessments that highlight the technical features. TMs or instruments available or focus on gene and protein-based data, rather than actual research implementations and concerns that go beyond genes, but also the study of proteins [12].The rate of collection of experimental information has increased along with the exponential growth of documentary databases.The administration of hundreds or thousands of genes and proteins is assessed under many experimental conditions through high bandwidth gene expression monitoring or genomic surveys [13,14].

Retrieving the necessary information from bibliographic databases and combining experimental data takes time and requires an accurate selection of keywords and query writing [15].This would often be a staggered and time-consuming operation that would produce inadequate research findings, preventing the full potential of these datasets [16].Researchers could benefit from automated treatment and interpretation of content while analyzing scientific publications [17].Today, TM is expected to be used to solve a wide range of communication problems, from finding drug targets and biomarkers in broadband trials to repositioning drugs. Developing a state-of-the-art summary of a disease or treatment area and developing domain-specific data sets [18].

3. Methodology

Figure 1 depicts the basic foundation of Text. It consists of smart machines such as Natural Language Processing (NLP), DL, BDA and decentralized NoSql datasets structure. It retains natural textual data and the capabilities generated by biological information. Present the model and text growth used to extract disease data in 19,681 abstracts and 29,437 full-text research journals, and then create predictive models to categorize the data [19].First, transformed summaries and/or real articles into a structure that Machine Learning (ML) algorithms and classification techniques might comprehend. To assess the relevance of words to parameter estimation, a Bag-of-Words (BoW) formulation using Term Frequency - Inverse Document Frequency (TD-IDF) was used. The probability of the happening of every particular phrase and TD-IDF to create bag-of-words, description. It scores were to be used as feature values [20]. Equation (1) calculates the TF-IDF weighting value as follows:

$$Q_{x,d} = \left(TF_{x,d}\right) \times \log_{10}\left({}^{M}/_{IDF_x}\right) \quad (1)$$

The TF, would be the probability with which the phrase x, M denotes the several papers to the dataset, and IDF denotes the several reports that comprise the phrase x. $Q_{x,d}$ is commonly utilized in text analysis and information extraction methods. The elimination of extraneous characteristics was a significant advantage of utilizing $Q_{x,d}$. A database, for example, has 1,000 pages. To evaluate the relevance of the phrase "nearly" to a character in the dataset, the TF-IDF grading rating would be computed as follows: Supposing the regularity to the phrase "almost" in the first essay is 58 and the phrase "nearly" happen in 1,000 papers in the database, the TF-IDF grading rating was computed in Equation 2.

$$Q_{almost,1} = (58) \times \log_{10}\left({}^{1000}/_{1000}\right) \quad (2)$$

Humans used three distinct classification techniques to develop a prediction model based on abstracts and whole text reports recovered by INSPEC, including the Euclidean method, ACO, and multiple optimizers. The primary work of the prediction model would be to allocate abstracts/articles to several designated groups: carcinoma, lung carcinoma and prostate carcinoma. The developer's malignancy classifications were compared to those defined using medical topic terminology. We recognize that MeSH phrases were generated procedurally using more advanced techniques, but we're utilizing them as a gold standard to see if a BDA could replicate forecasts that meet this gold standard [21]. A BDA infrastructure was used to build the proposed evolutionary framework, which included an Apache Hadoop group, Apache Spark elements, and NoSQL data set.

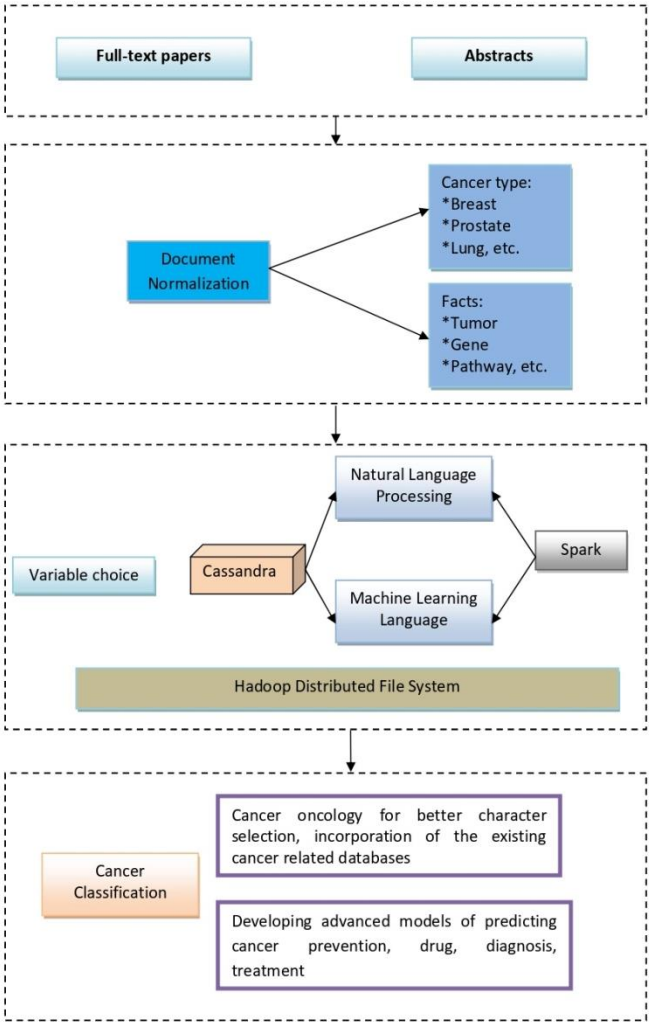

Figure 1: TM proposed architecture

3.1 Text selection

First, create phrases using pre-categorized summaries or full-text publications. Replaced special characters such as inverted commas and other exclamation points with empty spaces to generate standardized remarks, and all sentences were written in lower case. After that, it dissected phrases in specific terms. After deleting unusual expressions and user-defined stop-words, the Stemmer Porter technique was used to restrict all expressions.

3.2 Extraction of features

IDF was calculated by dividing the total of summaries or full-text papers in the training sample by the number of summaries or complete documents that make up the phrase. The TF-IDF balancing approach, which integrates these variables, was a very good balancing strategy in text categorization. For bigrams, trigrams, unigrams, and other forms of words, this grading method worked perfectly. During these procedures, humans converted all conceptions or full-text journals into "equal-length" mathematical feature representation, with each feature providing the TF-IDF

of a unigram and/or bigrams in a detailed report or idea recurrence. When using unigrams and/or bigrams in multiple databases, there were thousands to tens of thousands of characteristics, including summary and/or full-text. All summaries or full-text publications accompanying feature representation were organized using a BoW method. A basic example of a BoW structure could be seen in Table 1.

Table 1: Representation Style

Article Number	Bio-log	Bio-psi	Bio-lab	Bio-tin	Nearly	cancer	Stages of cancer	Article classification
12	13	2	3	11	0	2	5	BC
15	11	2	0	4	0	7	2	BC
76	5	2	2	2	0	29	0	BC
215	5	0	0	0	0	19	8	BC
332	0	2	0	10	0	21	2	BC
985	8	3	0	15	0	12	6	BC
4032	4	2	4	2	0	20	9	LC
5326	5	5	3	7	0	15	12	LC
5970	9	0	5	10	0	10	18	LC
40261	2	0	0	12	0	22	2	PC
51191	70	0	6	15	0	12	2	PC
82038	7	2	2	18	0	20	0	PC
93851	2	2	9	18	0	18	4	PC

Note: BC- Breast Cancer LC- Lung Cancer PC- prostate-cancer

4. Training and Evaluation

Following the processes above, all summaries or full-text papers were converted to an articulation system suited to computer learning approaches. To enhance and develop a forecasting model, used three well-known classification techniques: ACO, Multiple optimizer, and Euclidean algorithm. The Apache Spark created such as ACO, Multiple optimizers, Euclidean method classifiers, and scalable Apache Spark MLlib classification components. Information is labeled in a classification issue by being allocated to a particular caste. After that, the modeled judgments are to give categories to additional unlabeled data. This is a discrimination issue in which the similarities and differences between categories are modeled. Recognizing a minimize of a convex function f that would be attached with a parameters matrix w that has d number of attributes could be structured as a convex optimization problem, which would be the job of recognizing a minimize of convex function f that would be attached with a parameters matrix w that has d amount of materials. This remark could be summarized optimization problems.L2 regularization is used to default in the Apache Spark ACO method of Equation (3). The element generates a binary multiple optimizer model for a binary classification issue. This model could make

forecasts using logistic analysis, Equation (4) provided a data point, represented by x.

$$\frac{1}{2}\|x\|_2^2 \quad (3)$$

$$f(x) = \frac{1}{1+e^{-x}} \quad (4)$$

We used the default values for the Apache Spark MLlib categorization elements ACO, Logistic Regression, and Euclidean method. Humans investigate their characteristics using both default and nondefault variables to better empirically evaluate the proposed methodology, as shown in the findings. In our evaluation, they ran ten 5-fold cross-evaluation tests looking at the detection performance of both summaries and full-text documents. In an experiment, the dataset was divided into five comparable resamples at the chance. One of the five subsamples was kept as the validation data, while the others were utilized as the training databases to develop the model. The process of cross-validation was carried out five times. The average predictive performance across all tests was then calculated.

4.1 BDA using Spark

Biomedical text analytics could produce novel hypotheses by methodically evaluating a huge number of summaries and/or full-text pieces through scientific journals. A fundamental difficulty with the utilization of BDA reported to biochemical publications is the proper administration, storage, and retrieval of vast amounts of data. If data sizes reach a terabyte, the information must be divided into digestible chunks so that it can be accessed and interpreted utilizing distributed calculation techniques. To address the difficulties of big-scale text categorization, humans constructed the proposed toolbox using Apache Spark and the Apache NoSQL dataset. Apache Spark would be a free Distributed Data processing platform that focuses on speed, reliability, reusable, and complicated analytics.It provides simple and strong programming languages that serve a variety of applications, including ETL, computer vision, processing capabilities, and chart computation. It's also a scalable system for BDA, with a high-level application programming interface and a standard design concept. In 2009, Spark was developed at the University of California Berkeley's AMPLab, allowing more than 80 high-level specialists to design distributed applications. A NoSQL dataset would be a decentralized database that could manage massive volumes of data and is open source. Content provides flexibility and accessibility because it lacks a single point of failure.

Algorithm: ACO spark with BDA

Step 1: The ratio of the two community BDA sizes is calculated as

$$rpq = \min(|Cp|, |Cq|) / \max(|Cp|, |Cq|) \quad (5)$$

Step 2: If $r_{pq} < 1$, which means that there is no significant difference in size between the two communities, calculate the overlapping proportion measure of BDA (δ_{pq}) using

$\delta pq = | cp \cap cq | / | cp \cup cq |$ (6)

$\beta = (| cp \cup cq | - (| NcpUNcq | / 2)) / 10$ (7)

where, Cp and Cq are the pth and qth overlapping communities

NCp and NCq are the set of neighbor nodes that directly connect with the nodes in Cp and Cq

Otherwise, go to step 4.

Step 3: If $\delta_{pq} > \beta$, the two communities should be combined; otherwise, there is no combination operation executed. Go to step 6.

Step 4: If $r_{pq} > 1$, which means that the size of one community is much smaller than the other,

$\delta pq = | cp \cap cq | / (| cp \cup cq |)(|Cp \cap cq |)$ (8)

is used to calculate the overlapping proportion measure.

Step 5: If $\delta_{pq} > \beta$, the two communities should be combined; otherwise, there is no combination operation executed. Go to step 6.

Step 6: Output the community combination result.

Step 7: Merge small regions with the best solution.

5. Results and discussions

To create both training and testing databases, they obtained summaries and full-text papers from INSPEC. Table 2 lists the datasets and their properties. When assessing text classification methods, segregating a database into "training data" and "test samples" is crucial. A training set would be used to develop forecasting models in this database, while a testing set is used to assess the system. For this purpose, they used 5-fold cross-validation for each database in Table 2, employing 80 percent of the objects to train forecasting models and the remaining 20% to verify it. They employed the 64-bit Linux CentOS computer system on a cluster configuration with 20 data nodes, every to 6 GB storage, two CPUs, as well as 1 TB of hard disc space.

True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN)are four possible results in an Confusion Matrix. The numerical findings for Text's accuracy, accuracy, and Recall are shown in Table 2. The results of the three separate databases, researchers used three classification algorithms in this research. Default configuration integrity to the multiple optimizers, Apache Spark of Euclidean algorithm classifiers, ACO, and MLlib elements, were used to produce the results. It looked at the Receiver Operating Characteristics bend for a comparative evaluation of the three distinct ML approaches' estimation methods. The ACO classifier's area under the curve indicated a reasonable test, however, the Euclidean algorithm classifier's region contrasted poorly to the other two classification techniques. On the dataset, the ACO classifier outperformed the Multiple Optimizer and the Euclidean algorithm.

Table 2: Parameters used in TM

Data set source	Year in range	instance	BC	LC	PC
Article1	2017-2021	19,702	6,236	7,206	6,879
Article 2	2017-2021	12,872	4,951	4,285	4,315
Article 3	2017-2021	30.125	9,167	9,872	9,792

Compared the validity of the proposed system utilized summaries and entire-text papers and discovered to the forecast models utilizing summaries were more accurate than entire-text reports. The counterintuitive, one plausible justification could be the characteristic space's size, which appears to be too huge for full-text papers. Paper 1 may be more useful for text classification than full-text scientific papers. Full-text paper, on the other hand, was anticipated to be a fuller delivery of data than summaries alone the challenge of information processing and knowledge extraction. As a result, future research will try to use a feature extraction algorithm to optimize forecasting models based on full-text documents.

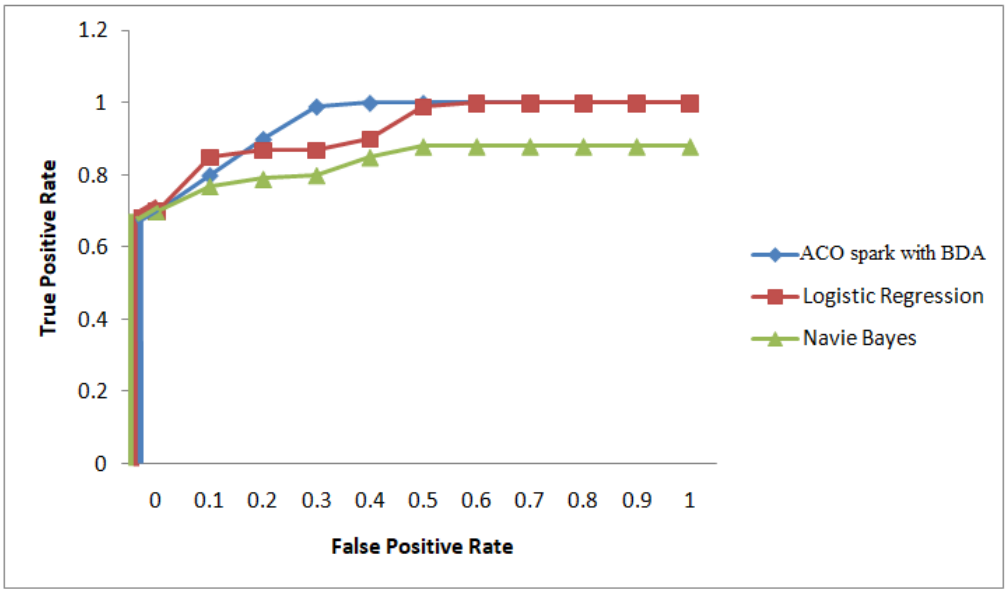

Figure 2: ROC curves of dataset

The text was also calculated to two Tag Helper Tools, open-source toolsets, Weka Library, and for reliability, accuracy, and recall. On several databases, including summaries and full-text papers, Figure 3 demonstrates quantitative comparisons of estimation methods utilizing the ACO, Multiple optimizer, and Euclidean algorithm. The variance between the precision assessed by the Text structure and the precision predicted by two well-known Tag Helper and Tools Weka Library is less than 1%, as shown in this research. As a consequence, the Text framework produces fair results in terms of forecasting performance. Text's performance was encouraging, as the Tag Helper tools and Weka Library had used to handle text data categorization difficulties for some years. Text performed substantially better than the two most commonly used text mining libraries in terms of effectiveness, precision, and memory, as demonstrated below.

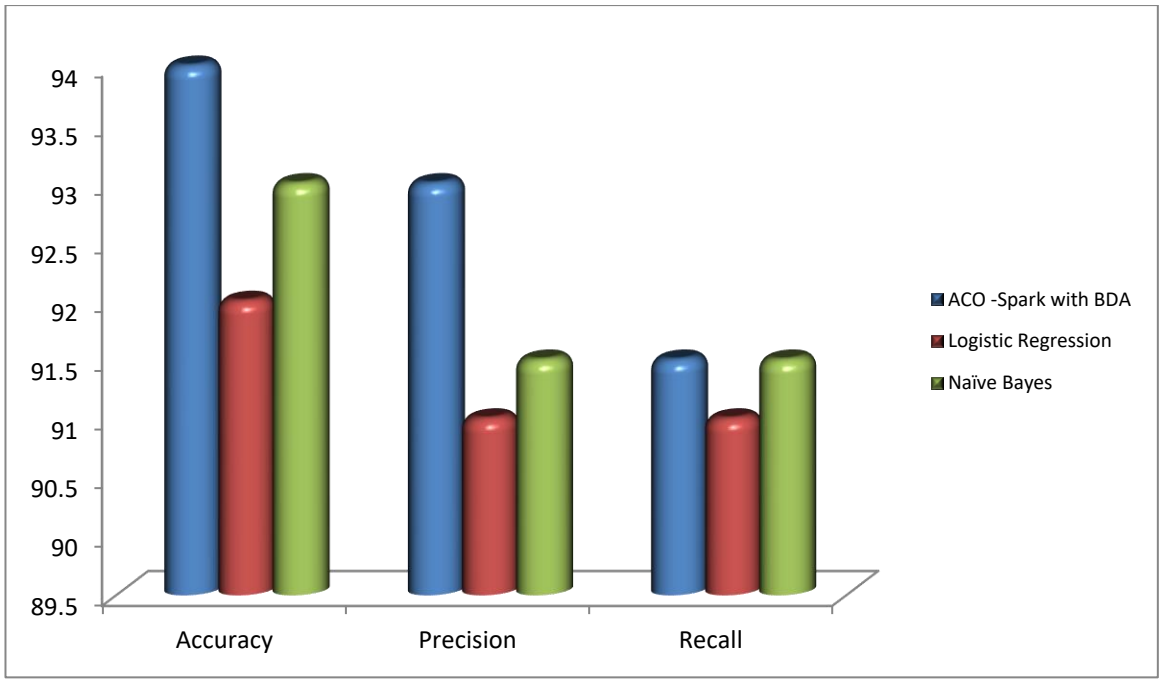

Figure 3: TM values of proposed and existing methods

The goal of this study was to look into huge-scale biochemical phrase categorization using enormous databases collected to INSPEC. We used NLP, BDA to develop and construct a parallel computing structure for extracting data like cancer type for chest, prostate, and/or lung diseases, and then developing a forecasting model of classifying data retrieved through tens of thousands of summaries and entire-text papers uploaded from INSPEC by affiliated MeSH aspects. A BDA including an Apache Hadoop group, an Apache Spark element, and NoSQL data was used to construct the TM. ACO's Spark with BDA accuracy in detecting malignancy type using summaries was 94.63 percent, whereas it was 93.81 percent utilizing the 29,437 entire-text papers. The suggested toolkit is mined to a big database of 29,437 entire-text papers more than 130 times quicker than other available approaches.

6. Conclusions

To categorize three types of cancer, the direct research used three different computer vision techniques offered from Apache Spark elements and ten thousand documents acquired by INSPEC. We did not examine whether a particular malignant tumor or a group of malignant tumors were published in this study. Proposed a new method of ACO Spark with BDA to evaluate the precision, adaptability and performance of text execution. We should emphasize feature extraction and sparse representation to provide meaningful features, which would compress all the features of text classification methods. It want to concentrate on multi-dimensional categorization jobs to classify information gathered from science publications by malignancy category, but also malignancy therapies, recognition, and protection categorization, to improve malignancy research knowledge and information. We also plan to create novel machine learning techniques for uncovering correlations across gene and aliment, gene and drug dosage, and affiliations to promote personalized treatment. This research shows how a ACO spark with BDA can be used for large-scale TM with real-time medical application prove that it predict the disease with more accuracy than existing methods.

References

[1] Jalali, S. M. J., Park, H. W., Vanani, I. R., & Pho, K. H. (2021). Research trends on big data domain using text mining algorithms. Digital Scholarship in the Humanities, 36(2), 361-370.

[2] Kushwaha, A. K., Kar, A. K., & Dwivedi, Y. K. (2021). Applications of big data in emerging management disciplines: A literature review using text mining. International Journal of Information Management Data Insights, 1(2), 100017.

[3] Lyu, F., & Choi, J. (2020). The forecasting sales volume and satisfaction of organic products through text mining on web customer reviews. Sustainability, 12(11), 4383.

[4] Kitsios, F., Kamariotou, M., Karanikolas, P., & Grigoroudis, E. (2021). Digital marketing platforms and customer satisfaction: Identifying eWOM using big data and text mining. Applied Sciences, 11(17), 8032.

[5] Rodríguez-Rodríguez, I., Rodríguez, J. V., Shirvanizadeh, N., Ortiz, A., & Pardo-Quiles, D. J. (2021). Applications of artificial intelligence, machine learning, big data and the internet of things to the covid-19 pandemic: A scientometric review using text mining. International Journal of Environmental Research and Public Health, 18(16), 8578.

[6] Park, J. H., Lee, H., & Cho, H. (2021). Analysis of the supportive care needs of the parents of preterm children in South Korea using big data text-mining: Topic modeling. Child Health Nursing Research, 27(1), 34.

[7] Lerena, O., Barletta, F., Fiorentin, F., Suárez, D., & Yoguel, G. (2021). Big data of innovation literature at the firm level: a review based on social network and text mining techniques. Economics of Innovation and New Technology, 30(2), 134-150.

[8] Wu, B., Wang, L., Lv, S. X., & Zeng, Y. R. (2021). Effective crude oil price forecasting using new text-based and big-data-driven models. Measurement, 168, 108468.

[9] Uthayakumar, M., Karnan, B., Slota, A., Zajac, J., & Davim, J. P. (2019). Performance Study of LaPO4-Y2O3 Composite Fabricated by Sol-Gel Process Using Abrasive Waterjet Machining. In Handbook of Research on Green Engineering Techniques for Modern Manufacturing (pp. 143-161). IGI Global.

[10] Kumar, S., Kar, A. K., & Ilavarasan, P. V. (2021). Applications of text mining in services management: A systematic literature review. International Journal of Information Management Data Insights, 1(1), 100008.

[11] Minu, M. S., Chevanan, S., & Samuel, J. V. (2021). A Unique and Integrated Approach about the facts of Big Data Cloud Adoption using Data Mining. Annals of the Romanian Society for Cell Biology, 13627-13634.

[12] Saritas, O., Bakhtin, P., Kuzminov, I., & Khabirova, E. (2021). Big data augmented business trend identification: the case of mobile commerce. Scientometrics, 126(2), 1553-1579.

[13] Garikipati, P., & Balamurugan, K. (2021). Abrasive Water Jet Machining Studies on AlSi 7+ 63% SiC Hybrid Composite. In Advances in Industrial Automation and Smart Manufacturing (pp. 743-751). Springer, Singapore.

[14] Aroulanandam, V. V., Latchoumi, T. P., Balamurugan, K., & Yookesh, T. L. (2020). Improving the Energy Efficiency in Mobile Ad-Hoc Network Using Learning-Based Routing. Rev. d'Intelligence Artif., 34(3), 337-343.

[15] Ahn, S. J., Yoon, H. Y., & Lee, Y. J. (2021). Text mining as a tool for real-time technology assessment: Application to the cross-national comparative study on artificial organ technology. Technology in Society, 66, 101659.

[16] Romero-Silva, R., & De Leeuw, S. (2021). Learning from the past to shape the future: A comprehensive text mining analysis of OR/MS reviews. Omega, 100, 102388.

[17] Gopalakrishnan, K., Agrawal, A., & Choudhary, A. (2017). Big Data in building information modeling research: survey and exploratory text mining. MOJ Civil Eng, 3(6), 00087.

[18] Garikapati, Pruthviraju, Balamurugan, K., Latchoumi, T. P., Malkapuram, Ramakrishna (2021). A Cluster-Profile Comparative Study on Machining AlSi7/63% of SiC Hybrid Composite Using Agglomerative Hierarchical Clustering and K-Means. Silicon, 10.1007/s12633-020-00447-9

[19] Asogwa, D. C., Anigbogu, S. O., Onyenwe, I. E., & Sani, F. A. (2021). Text Classification Using Hybrid Machine Learning Algorithms on Big Data. arXiv preprint arXiv:2103.16624.

[20] Rockwell, G., & Berendt, B. (2016). On big data and text mining in the humanities. Data mining and learning analytics: Applications in educational research, 29-40.

[21] Sung, Y. A., Kim, K. W., & Kwon, H. J. (2021). Big data analysis of Korean travelers' behavior in the post-COVID-19 era. Sustainability, 13(1), 310.