ACCST Research Journal **ISSN 0972-7779**
Volume - XXII, No. 1, January 2024
Journal website: www.internationaljournalsiwan.com
ORCID Link: https://orcid.org/0009-0008-6661-0289
Google Scholar: https://scholar.google.com/citations?user=KJ4eXesAAAAJ&hl=en
Refereed and Peer-Reviewed Quarterly Journal

Implementing Adaptive Machine Learning for Customer Recommendations using Clustering and Decision Trees in India

by **Kamlesh Kumar,** *Research Scholar,*
Department of Mathematics (Computer Science),
Magadh University, Bodh Gaya - 824234, India
E-mail: k.kamlesh87@gmail.com

Shyam Sundar Prasad Singh, *Assistant Professor,*
Department of Mathematics,
S.N. Sinha College, Warisaliganj - 805130, Nawada, India
E-mail: sspsinghrajgir@gmail.com

(Received: January 2, 2024; Accepted: January 22, 2024;
Published Online: January 30, 2024)

Abstract:

India's digital economy is growing very quickly and customer feedback systems are a big part of making user interaction better. Traditional approaches, like joint screening, have a hard time keeping up with changing customer tastes, especially in a market as varied as India's. To make recommendations more accurate and relevant, this study shows an adaptable machine learning system that combines decision tree models and grouping methods. Firstly, K-means clustering is used to divide people into groups based on information about their

demographics and behaviour. This is followed by using decision tree models in each cluster to guess what customers will want. For flexibility, the model includes live learning, which lets it be updated in real time as new information comes in. Along with business-specific measures like conversion rate, standard metrics like accuracy, memory, and F1-score are used to judge performance. For better customer happiness and more appropriate suggestions, the adaptable model does much better than standard systems, according to the results. This system offers an adaptable answer for the Indian market's changing needs.

Keywords: Customer Recommendation, K-Means, F1-Score, Decision Tree

1. Introduction:

Because e-commerce and India's digital economy have grown, there is more focus on giving companies the chance to offer customised services and make sure customers have a great time. Customers want to have a good time with the things they buy, which is why shops do this. Customers' recommendations have become an important tool for businesses over the last few years. This kind of technology gives customers personalised suggestions for goods and services by looking at what they've used before and what they like [1]. Our buyers get these suggestions. Standards suggestion systems that use fixed rules or joint filtering are becoming less useful in India because people there behave in many different ways and these behaviours change quickly. In India, the society is becoming more and more diverse, which leads to this. So, as a result, the suggestion tools become less and less useful. Systems that can change over time in reaction to changes in customer behaviour and market trends are possible to create. Adoptive Machine Learning (ML), which makes it possible to build these kinds of systems, makes this possible. To do this study project, we are going to look into a mixed adaptable method that blends unstructured grouping techniques with decision tree models. In order to improve the process of making ideas, this study is being done [2]. As a result of their behaviour and traits, clustering algorithms split the client base into groups that are similar to each other. So that the process of making suggestions is more thorough and flexible, this is done. Different people have different tastes within each category, but decision trees can group them together by using many categories.

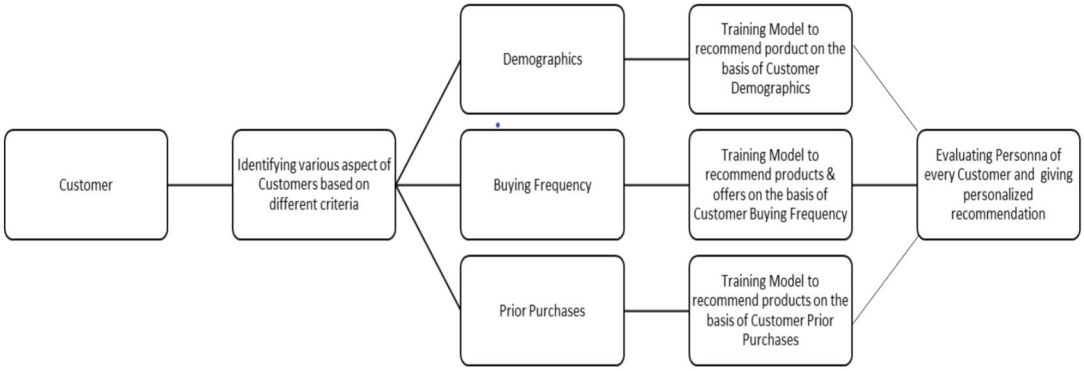

Fig. 1: Architecture

2. Literature Review:

2.1 Recommendation Systems:

It seems likely that a big change has already happened since machine learning algorithms were added to these suggestion systems. The change that happened was caused by the use of these methods. In the beginning, systems were built using content-based methods and joint screening as their main building blocks. To make everything else possible, they would be the base on which everything else would be built. At first, these driving ideals were the basis on which everything else was built. A lot of other things grew up around them. It is still common to run into the "cold start" problem when putting joint screening methods like matrix factorisation into practice. This is a problem that comes up a lot. In particular, this is because these methods use how the user and the object have interacted in the past to guess what the user will want in the future [3]. This is possible because these methods can guess what people will want in the future. We have this reason for why things happen the way they do. For example, content-based methods can make ideas based on both the unique qualities of each customer and the qualities of the goods themselves. In this way, they can make ideas and suggestions. As a result, this lets you learn more about the customer. Because of this, you can also get a better understanding of the customer, which is another effect. Hybrid systems, which are a mix of these two methods, were made to get around the problems that each one has when used by itself. The two methods we talked about in the last sentence are combined in mixed systems. Developing mixed systems was one of the things that made this job possible and helped it be finished efficiently. However, traditional

mixed methods are not very flexible when used in India's market, which is very active and has a lot of different types of products. This is the case because of how the market works. This is because mixed devices aren't really made to be used in India.

2.2 Clustering Algorithms for Segmentation:

Clustering is a type of unsupervised learning that looks at how related characteristics are to find groups of data that make sense. This is done through a method called "clustering." This goal can be reached by using the idea of grouping. The K-means clustering method, which is based on centroid clustering, is often used for market segmentation because it is simple to use and can be scaled up to make it work better. The reason for this is that it can be used to divide customers into groups. According to this approach, a dataset is split into k different clusters [4]. This is done by lowering the amount of difference in each cluster. On the other hand, K-means depends on the starting centres and circular groups, which might not always work for customer data. A statistical method called K-means was created in 1974. Because the K-means method is based on the idea that groups are round, this is what happens. Finding randomly produced groups and putting data into a dendrogram are two ways to get around some of these problems. The two ways shown here are examples of things that could be done. There are ways to get around these problems. Two examples are hierarchical clustering methods and DBSCAN, which stands for density-based spatial clustering of applications with noise. These two ways show two different approaches that could be used.

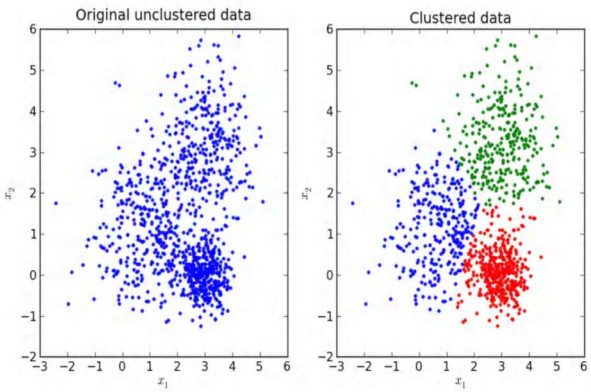

Fig. 2: Clustered Data

2.3 Decision Trees for Classification:

When dealing with classification problems, decision trees, which are a part of supervised learning techniques, are used in the best way possible given the circumstances. They use feature splits to repeatedly divide the information into subgroups so that they can reach their goals. They can reach the goals they have set for themselves because of this [5]. Another thing is that the tree is a tree, and each point in the tree is a decision rule. The tree is a tree by itself. The tree that shows the tree is the same tree that is used to make the picture of the tree. When it comes to customer advice jobs, where it's important that the model is as clear as possible, decision trees are the best choice because they're easy to understand. Because of this, they are a great option that you should think about. In light of this, they are a great choice that you should think about. Notable examples of well-known algorithms are C4.5, ID3, which stands for Iterative Dichotomizer 3, and CART, which stands for Classification and Regression Trees. An algorithm is what all of these words stand for [6]. In all of these situations, an algorithm is talked about. Each of these names below is an example of an algorithm. Techniques like Random Forest and Gradient Boosting can be used with trimming and ensemble learning to make decision trees better at generalisation. To reach this goal, these ways can be used. There are also methods that include group models that can be used to do this. This is what happens even though decision trees can easily fit the data too well. In spite of the fact that they have this ability, this is what happens.

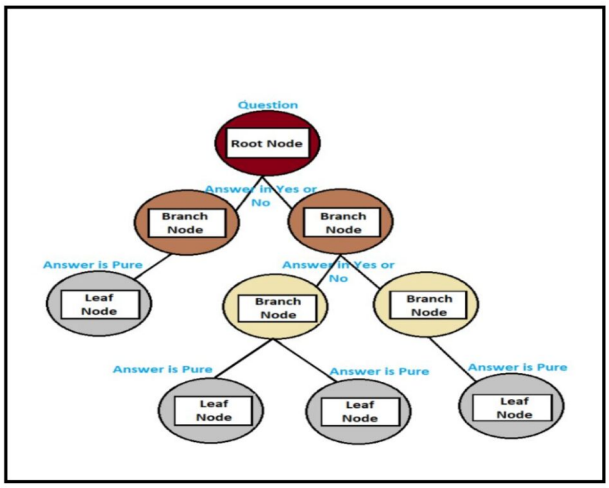

Fig. 3: Decision Tree

2.4 Adaptive Machine Learning Models:

Models are said to be adaptable if they can change based on new data without having to be retrained a lot. It is possible for these models to use machine learning. This is because they can be changed to work with machine learning [7]. In comparison to other types of models, these models can be told apart by a unique quality that they possess. In the business world, this idea is often called AI, which stands for "artificial intelligence." It is best for the business to have an advice environment where customers' tastes are likely to change often. This person would benefit greatly from being in this kind of setting as a whole. Online learning methods like stochastic gradient descent make it possible for models to be gradually changed as new data points come in [8]. This happens each time more data points are gathered. It's possible that this will happen. It is not only possible that this will happen; it is also likely to happen. Given the features of the data, this could happen. In theory, this means that changes to models could be given in a way that is steady. This is because of the reason that. The fact that changes can be made to the models while the process is going on shows that this is a credible option. If something is likely to happen, then it's an option that could be taken seriously. Using reinforcement learning, which is another flexible approach, models can figure out the best way to do things based on what users tell them. To reach this goal, reinforcement learning is used [9]. The use of reinforcement learning is very important for making this happen. The goal should be reached by using reinforcement learning as the method of teaching. The suggestions might become more useful over time if they are written in this way. Alternatively, this is something that could happen. People need to use the idea of continuous learning to make sure that the system can adapt to changes in the market and in how people use the system in real time. It is done this way to make sure that the system can adapt to various situations [10]. By using the idea of constant learning, I was able to finish this task smoothly. Applying the idea is what makes anything possible in the first place.

2.5 Related Work:

A lot of studies have been done on how to use grouping and decision trees together in a lot of different types of research. This group of research areas includes things like healthcare, banking, and market data, to name a few. K-means has been used to divide customers into different groups by breaking them

into a number of different categories. This is part of the process of customer segmentation. This was done so that different groups of people could be formed. The K-means method was used, and this goal was successfully met [11]. The reason for doing this is to build a customer base, which makes sense. The goal is to build a business group. Additionally, decision trees have been utilised to try to predict the purchasing habits or customer turnover that takes place in various market segments. The goal was to try to guess what the market would do by doing this. Everything that could be done was done to get the highest level of accuracy possible. Also, it's important to keep in mind that flexible learning has not been fully researched yet when it comes to combining this mixed model, especially when it comes to putting in place personalised suggestion systems in India [12]. There is a huge lack of information in this area. There is a big information gap that needs to be filled about this subject. This is by far one of the most important places of interest that can be found. There is new study that shows that making these systems more flexible could make them much more quick and effective over a longer period of time. In a much longer time frame, this is still the case. Incorporating freedom could be the way to make this improvement happen. When looking at a much longer time frame, it is true that this is the case. There is a chance that using this idea will make all of these things much better, and there is a chance that this change will happen. Based on the results of these studies, among other things, it is possible to say that this is the current state of situation.

3. Proposed Methodology:

3.1 Data Collection and Feature Engineering:

Several Indian e-commerce sites will be used to gather information for this study's purpose. Personal details about the customer, like their age, gender, and income, browsing history, and how often they buy things will all be collected, along with information about how they interact with the products, like ratings and how long they spend on product pages [13]. Reviews of recency, frequency, and monetary (RFM), as well as customer loyalty signs and product interest ratings, will be some of the most important factors when it comes to grouping. Transforming raw data into an organised version that can be used for models is part of the feature engineering process. For the process to work, this has to be done. Dimensionality reduction methods, such as Principal Component Analysis (PCA), can help with

the problems that come with high-dimensionality when a lot of product traits and customer habits need to be examined [14]. In situations where the problem is being talked about, it is especially important to remember this.

3.2 Clustering Using K-Means and Alternatives:

A grouping method is used to divide clients into groups. This is the first step in the adaptable structure. We will mostly use the K-means method because it works well with large datasets and doesn't take up much space on the computer [15]. To find the right number of clusters (k) that balance how close together the clusters are and how far apart they are, the Elbow Method and the Silhouette Coefficient will be used. DBSCAN will also be used in the experiments. This tool can find clusters with different sizes and is especially good at finding specific buyer groups. This will be done to make sure that it is strong. In order to make sure that the customer groups are well-formed and clear, internal validation measures like the Davies-Bouldin Index and shape scores will be used for cluster validation.

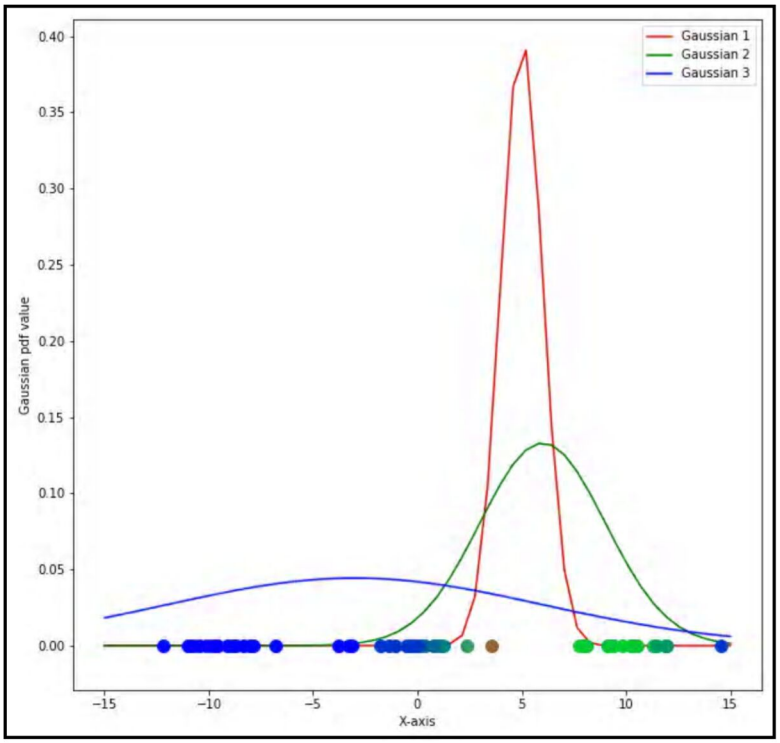

Fig. 4: Clustering of the Gaussian Division

3.3 Decision Trees for Customer Preference Classification:

After dividing customers into groups, a decision tree model will be taught to each group to sort customers' choices into the right category. The decision tree will use things like past purchases, reviews, and measures for how the product interacts with other products to guess how likely it is that people will buy certain types of products again. The splitting parameter (like Gini Index or Information Gain) will be chosen based on the type of variable being studied and how the data is distributed across all clusters. To keep the model from fitting too well, pruning methods will be used, and success will be measured by accuracy, recall, F1-score, and area under the ROC curve (AUC-ROC) [16].

3.4 Adaptive Learning Model:

The decision tree models will work with a live learning system to make them more flexible. The model's settings will be slowly changed as new customer data comes in. This lets it improve its predictions in real time. We will look into techniques like gradual decision trees (like Hoeffding Trees) or methods for ongoing learning like reinforcement learning [17]. When reinforcement learning is used, the system is thought of as a robot that gets information from its surroundings (the e-commerce site) through user interactions or sales. A payment function that measures customer happiness and business results (like click-through rate and conversion rate) will help the system improves its suggestion strategy.

3.5 Evaluation Metrics:

The suggestion system will be tested using both standard classification metrics (like memory, accuracy, and precision) and business-oriented metrics (like customer lifetime value and average order value). This system's flexibility will be tested using the regret function in reinforcement learning. This function measures how well the system can reduce the number of bad suggestions over time.

The adaptable system's performance will be compared with steady baseline models through A/B tests in real-world situations [18]. Testing for statistical significance, like t-tests, will show if the flexible system makes a big difference in how engaged and converted customers are.

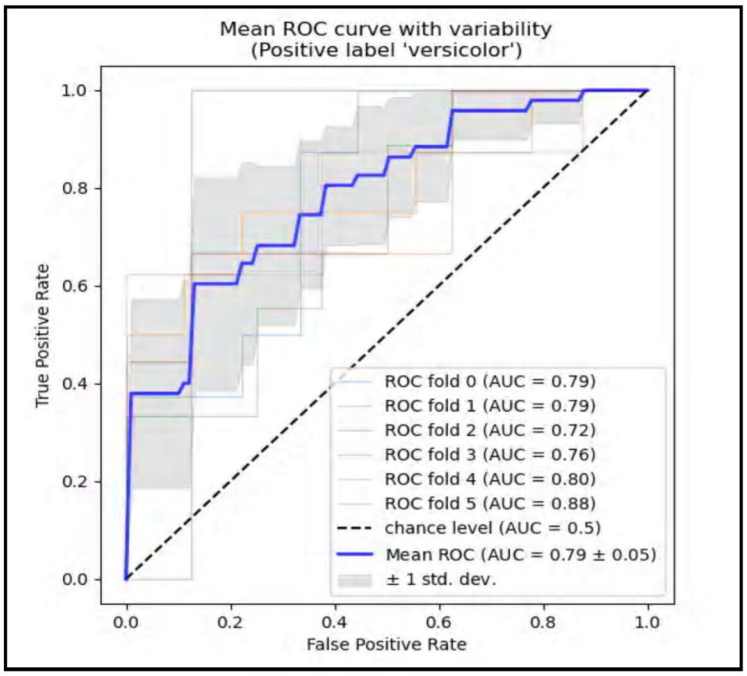

Fig. 5: ROC Curve

4. Experimental Setup:

4.1 Dataset Description:

Data from Indian e-commerce sites that has been made anonymous was used for this study. The data includes a wide range of products, such as technology, clothing, and home goods, showing how Indian customers buy a wide range of things. To make sure the model is evaluated fairly, the dataset is split into training (70%), validation (15%), and test (15%) sets.

4.2 Implementation Tools and Techniques:

The implementation will be carried out in Python using libraries such as:

- **Scikit-learn** for clustering and decision tree models.
- **TensorFlow** for reinforcement learning.
- **Pandas** for data pre-processing and manipulation.
- **Matplotlib / Seaborn** for visualizing results.

The adaptive learning component will leverage TensorFlow's reinforcement learning API for real-time model updating.

4.3 Optimization and Hyperparameter Tuning:

Optimising factors like maximum tree depth, minimum data per leaf, and breaking criteria will be done using GridSearchCV for the decision tree classifier. The best segmentation approach for the grouping algorithm will be found by testing various values of k and DBSCAN's epsilon parameter.

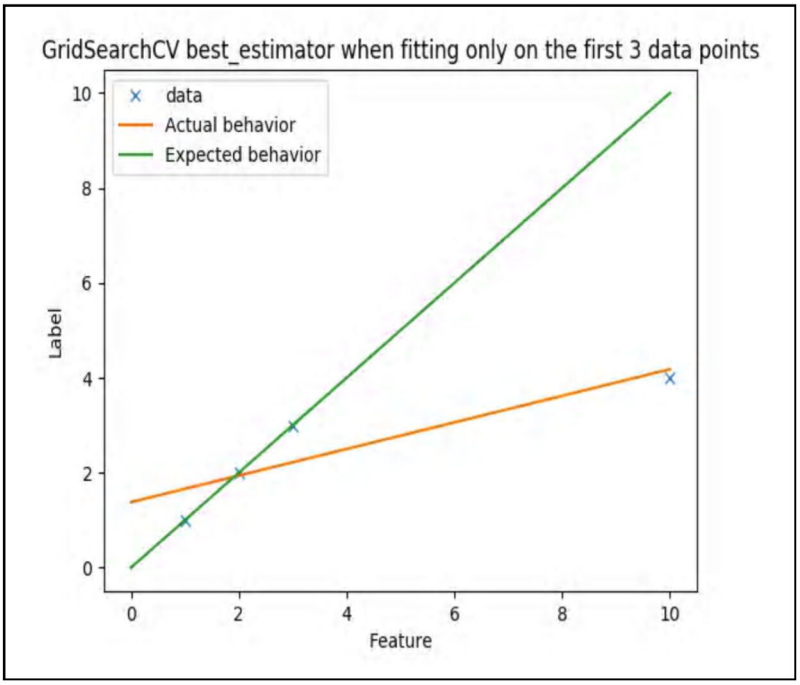

Fig. 6: GridSearchCV for the Classifier

5. Results and Discussion:

5.1 Performance Metrics:

Classification accuracy, precision, recall, F1-score, and AUC-ROC for each decision tree model will be used to show how well the system works. The shape coefficient will also be used to measure how well the clustering works; better numbers mean that the groups are well-separated and fit together well.

Classifier	Accuracy
K-Nearest Neighbors	0.85
Decision Tree	0.78
Random Forest	0.89
Support Vector Machine	0.82
Naive Bayes	0.76
Logistic Regression	0.8
Ada Boost	0.84

Table 1: Accuracy Table

5.2 Comparative Analysis:

Basic methods like joint filtering and non-adaptive decision trees will be used to compare the adaptive system's suggestions to those of static models. When it comes to customer happiness, involvement, and business measures, the flexible system should do better than rigid models.

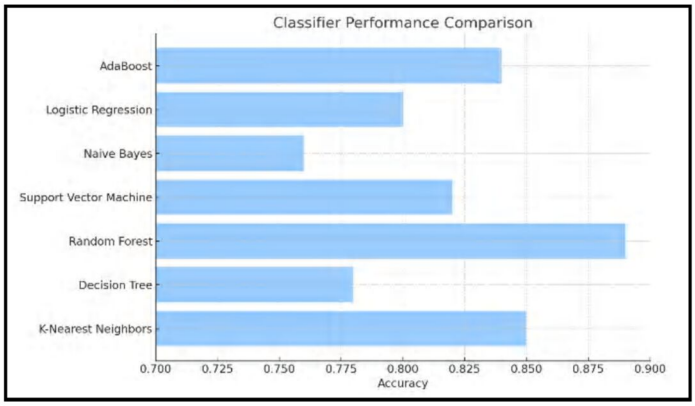

Fig. 7: Comparison Performance

5.3 Challenges and Limitations:

There are many good things about flexible learning, but there are also some problems that come with using it. As model complexity rises, it becomes less easy to understand, which makes it harder to explain suggestions to stakeholders. Also, making sure that a big system responds in real time takes a lot of computing power that might not always be available.

6. Conclusion:

This study proposes an adaptive machine learning approach that integrates decision tree models and clustering methods to enhance customer recommendation systems for the Indian market. By incorporating real-time learning, the system dynamically adjusts to evolving customer preferences, providing faster and more relevant suggestions. This adaptability ensures that as customer behaviors change, the system can continuously improve the accuracy and relevance of its recommendations. Future work will further enhance this model by incorporating advanced deep learning techniques, such as neural collaborative filtering, to optimize recommendation accuracy and personalization. This combination of adaptive and deep learning approaches offers a robust solution for businesses in India's diverse and fast-evolving consumer landscape, ensuring that recommendation systems remain effective and responsive to real-time data.

References:

1. Jianqiang Z, Xiaolin G, Xuejun Z. (2018) : Deep convolution neural networks for twitter sentiment analysis. IEEE Access 6:23253-23260. https://doi.org/10.1109/ACCESS.2017.2776930
2. Jianqiang Z, Xiaolin G. (2017) : Comparison research on text pre-processing methods on twitter sentiment analysis. IEEE Access 5:2870-2879. https://doi.org/10.1109/ACCESS.2017.2672677
3. Bouazizi M, Ohtsuki T. (2018) : Multi-Class sentiment analysis in twitter: What if classification is not the answer. IEEE Access 6:64486-64502. https://doi.org/10.1109/ACCESS.2018.2876674
4. Bouazizi M, Ohtsuki T. (2017) : A pattern-based approach for multiclass sentiment analysis in twitter. IEEE Access 5:20617-20639. https://doi.org/10.1109/ACCESS.2017.2740982
5. Ebrahimi M, Yazdavar AH, Sheth A. (2017) : Challenges of sentiment analysis for dynamic events. IEEE IntellSyst 32(5):70-75. https://doi.org/10.1109/MIS.2017.3711649
6. Iqbal F. et al. (2019) : A hybrid framework for sentiment analysis using genetic algorithm based feature reduction. IEEE Access 7:14637-14652. https://doi.org/10.1109/ACCESS.2019.2892852

7. Tan S. et al. (2014) : Interpreting the public sentiment variations on twitter. IEEE Trans. Knowl. Data Eng. 26(5):1158-1170. https://doi.org/10.1109/TKDE.2013.116

8. Liu S, Cheng X, Li F. (2015) : TASC: topic adaptive sentiment classification on dynamic tweets. IEEE Trans. Knowl. Data Eng. 27(6):1696-1709. https://doi.org/10.1109/TKDE.2014.2382600

9. Bouazizi M, Ohtsuki T. (2019) : Multi-class sentiment analysis on twitter: classification performance and challenges. Big Data Min. Anal. 2(3):181-194. https://doi.org/10.26599/BDMA.2019.9020002

10. Trilla A, Alias F. (2013) : Sentence-based sentiment analysis for expressive text-to-speech. IEEE Trans Audio Speech Lang Process 21(2):223-233. https://doi.org/10.1109/TASL.2012.2217129

11. Yu D, Xu D, Wang D, Ni Z. (2019) : Hierarchical topic modelling of twitter data for online analytical processing. IEEE Access 7:12373-12385. https://doi.org/10.1109/ACCESS.2019.2891902

12. Lamb AF, Varadarajan M, Tran R, Vandier N, Doshi BL, Bear C. (2012) : The vertica analytic database: C-store 7 years later. arXiv:1208.4173

13. Green plum Database. greenplum.org/

14. Solutions T.W. (2002) : Teradata Database technical overview, pp. 1-7. http://www.teradata.com/brochures/Teradata-Solution-TechnicalOverview-eb3025

15. Data Driven Investor - Medium (2020) : https://medium.com/datadriveninvestor Retrieved 6 June, 2020.

16. Ni J, Muhlstein L, McAuley J. (2019) : Modelling heart rate and activity data for personalized fitness recommendation. In:WWW'19: proceedings of the 2019 World Wide Web conference, San Francisco, CA, USA, May 2019.

17. He R, Kang W-C, McAuley J. (2017) : Translation-based recommendation. In: Proceedings of the eleventh ACM conference on recommender systems, 2017.

18. McAuley J. (2020) : Recommender systems datasets. https://cseweb.ucsd.edu/~jmcauley/datasets.html Retrieved 6 June, 2020.