Improving Collaborative Filter Using BERT

Riyam Rwedhi
Department of Computer Science, Faculty of Computer science and Mathematics
University of Kufa
Najaf, Iraq
riyamr.alabedi@student.uokufa.edu.iq
Orcid.org/0009-0009-0137-1210

Salam Al-augby
Department of Computer Science, Faculty of Computer science and Mathematics
University of Kufa
Najaf, Iraq
salam.alaugby@uokufa.edu.iq
Orcid.org/0009-0008-7784-1357

DOI: http://dx.doi.org/10.31642/JoKMC/2018/100204

Received Mar. 27, 2023. Accepted for publication May. 3, 2023

Abstract *With the increasing number of published books and the challenge of acquiring adequate research interest, recommendation systems have developed as a key tool for assisting scholars by proposing relevant and valuable books or articles automatically. Recommendation systems have the potential to enhance the accessibility and affordability of books. A recommendation system already exists, but results may be inaccurate in addition to the two main problems of RS: the cold-start problem and the data sparsity problem. In this work, aims to improve the precision of book collaborative filtering through the application of semantic similarity to book summaries. Furthermore, the study seeks to address the significant challenges associated with scalability and sparsity by employing effective techniques. The proposed methodology comprises three stages: preprocessing, building the system, and evaluation. The technologies used in the pre-processing stage included reduction and normalization. The construction system is divided into two phases: semantic similarity and recommendation. The semantic similarity is done by using BERT for sentence embedding and to calculate the similarity between sentences by cosine similarity. During the recommendation phase by using CF based on KNN. In the evaluation stage, classification accuracy metrics had been used. The proposed approach resulted in an enhancement of the accuracy of the book recommendation system, elevating it to 0.89 in contrast to antecedent studies and based on a dataset of 271,000 book summaries. The results obtained from the proposed approach were superior, as it circumvented issues encountered in preceding works, such as sparsity and scalability, by using BERT with CF based KNN. Filtering the data using BERT and the KNN algorithm in the CF added strength to the recommendation, which led to an increase in the accuracy rate.*

Keywords ***Recommendation systems; Semantic similarity; Collaborative filter; BERT; K-Nearest neighbor***

I. INTRODUCTION

Recommendation Systems (RS) can reduce the issue of information overload. The implementation of intelligent recommendation strategies within academic and educational domains significantly improves the efficacy of academic resource allocation, specifically for inexperienced scholars and learners during the search process for publications. It should be known those recommendation outcomes are typically diverse depending on the researchers' wants. To guarantee that RS is personalized and effective, the count of results can be limited and controlled [1]. Most commonly used are collaborative filtering techniques, which do not request prior knowledge of users or items and instead of produce recommendations based their interactions. Despite their efficacy and straightforward, the recommendation systems possess several restrictions such as the issue of cold start, accuracy of predictions, and the incapacity to capture intricate user-item interactions. [2]. When it comes to suggestions from friends for example, as predicted, what one person feels is fantastic may be extremely different from what another person enjoys. As a result, one user's experience with the system may be substantially different from another user's experience. This emphasizes the importance of personalization. Machine learning algorithms are critical for achieving effective personalization at scale [3]. Previous research has used several recommendation systems to generate book recommendations. The most prevalent strategy is content-based filtering. Typically, content-based filtering techniques extract contents in order to construct associations between books. The majority of these systems collect different information from books in order to create a user profile and make suggestions. These techniques based on earlier user profiles will not work effectively for new recommender systems. Furthermore, due to copyright restrictions, all book contents cannot be freely accessible. Furthermore, due to the ambiguity of natural language, content-based filtering cannot accurately capture user interests.

Collaborative filtering is a traditional recommendation method used in book recommendation systems. It generates a book-rating matrix in order to discover links between books [4]. Several challenges had been encountered in this work like: The difficulty of obtaining an integrated database, the lack of an Arabic language database, and the cold start if the recommender is unable to interact with the new members in any way this is beyond the capabilities of the collaborative filter because he is unable to interact with them directly. The remainder of the paper is structured as follows: Section 2 discussed the related works. Section 3 the proposed approach has been present. Section 4 discussed the experimental and result analysis. Finally, Section 5 is conclusion.

II. RELATED WORK

There are several study provided on the paper's recommendations, all of which aim to deliver the most effective advice. The studies covered in this literature review pertain to two categories: improve collaborative filtering and collaborative filtering. Whereas the researchers in [5, 6, 8, 9, 11, and 13] are working on the same target, which is the enhancement of the collaborative filter (CF) of the recommendation systems. The researchers in [9] instead of evaluating the items and then forming the recommendations, better results were obtained by evaluating the potential recommendation lists. Additionally, genetic-based recommender systems are superior to other methods and are capable of producing predictions that are more accurate, regardless of the quantity of K-neighbors.

As for researchers [7, 10, 12 and 14] they depend on CF, with the researchers in [10] achieving the best results by combining two-level paper-citation links based on citation context in order to recommend articles to customers as paper references.

In 2022, Sallam et al. improved the accuracy of Arabic CF using SA on user reviews. The proposed method is divided into two stages SA and recommendation. The SA phase computes sentiment scores using an Arabic specific lexicon. In the second phase, item-based and singular value decomposition based CF is used. The proposed method improves experiment results by lowering the average of root mean squared and mean absolute errors, as well as dealing with scalability and sparsity issues. They could have gotten much better results if the database had been larger [5].

In 2022, Geng and Li introduced recommendation algorithm that is enhanced based on student behavior data. When creating the behavior graph and behavior route, take into account the behavior sequence. Multidimensional behavior path vectors are compared, and recommendations for collaborative filtering are generated for each dimension separately. Proposed algorithm compared to other recommendation algorithms, and the Recall, Precision, and Root Mean Square Error were calculated of these recommendation algorithms for various recommended users. It demonstrates that the proposed algorithm has greater advantages in three evaluations. If the size of the database had been higher, they would have been able to achieve far more successful outcomes [6].

In 2022, Fkih introduced the similarity measures utilized for CF-based RS. Describe each measure's foundational history and evaluate its performance through an experiment. Experiments conducted on three standard datasets (MovieLens1M, MovieLens100k, and Jester) reveal numerous significant findings. Improved triangle similarity and. Improved Pearson Correlation Coefficient weighted with rating preference behaviorare which similarity metrics are best for a user-based RS. The best choice for an item-based RS is Adjusted Mutual Information, according to the results. By depending on the priority of the user profile, the system faces two problems: sparsity and cold start [7].

In 2021, Zhou et al. presented a CF algorithm based on filling in missing values. By adding user-item matrices, the algorithm resolves the issue of the sparse scoring matrix. The authors suggested an improved approach for optimizing the filling matrix using alternating least squares. Having noticed that various users and objects have unique preferences, they incorporated bias into their rating prediction model. Experiments have shown that the proposed algorithm significantly improves accuracy. Gives priority to the user's profile, and the system suffers from a cold start problem, which is its inability to provide recommendations to the new user [8].

In 2020, Alhijawi et al. introduces a novel genetic established recommendation system dependent on historical rating data and semantic data. The research contributes by evaluating potential recommendation lists rather than evaluating items and then compiling a recommendation list. The results showed that the predictive ability of the system exceeded the other methods in accuracy, whatever the number of K-neighbors [9].

In 2020, Sakib, Rodina, et al. proposed a new method for recommending scientific papers. This approach integrates two-level paper-citation relationships by ascertaining the citation context to recommend papers to users as references. The method was presented by these authors as a new approach to the scientific paper recommendation. Mining individual latent associations while simultaneously computing collaborative similarities produces the best suggestions compared to other approaches. When making recommendations for articles, the algorithm only considers information that may be quickly accessible, such as citation relations. The system runs purely on CF, has a sm all database, and has a cold start issue. If they used larger database or adding filter or algorithm with CF, that greatly improve system performance [10].

In 2019, Neysiani at el. Based on a genetic algorithm, this article presents a method that is more effective than others in producing cred associations rules that have higher performance levels. The MovieLens data set was used for the purposes of conducting evaluations. The following factors will be considered in the evaluation: run time; this study examines the mean values of quality rules, precision, recall, F1-measurement, and accuracy. The empirical assessment of a system founded on their algorithm demonstrates its superior performance compared with the multi-objective particle swarm optimization association rule mining algorithm, ultimately resulting in a 10% reduction in runtime [11].

In 2019, Liu et al. proposed a model for film recommendation by applied the vector's similarity to the CF recommendation algorithm. Experiments demonstrate increase ratio in the precision and recall of the recommendation. The system suffers from a cold start and the sparsity and scalability problems [12].

In 2018, Dubey et al. They created a dictionary of feelings to estimate the probability of positive reviews and then determine the feeling values. The utilization of sentimental ratings in a CF system was implemented to enhance recommendations and eliminate items that have received predominantly unfavorable user evaluations. The utilized datasets comprised of an IMDb review dataset that contained 25,000 reviews, which were categorized as either positive or negative. In relation to the usual methods of recommendation, the results of the proposed system were superior. If the size of the database had been higher, they would have been able to achieve far more successful outcomes [13].

In 2017, Haruna, Khalid, et al. study proposed a collaborative strategy for finding undiscovered collections between a needed article and its citations. To locate comparable neighbors, the authors employed a single-level paper-citation connection. However, the utilization of exclusively CF and a citation system comprising a solitary level of relationship may result in a reduction in precision when suggesting scholarly articles. The present methodology proposes the development of a collaborative filtering approach that is sparsity-aware, with the aim of addressing the sparsity issue inherent in conventional collaborative filtering techniques. The approach exceeded the baseline methods in terms of measurement in addition to general execution and the potential to get required and benefits research papers at the top of the suggestion roster. The using of purely CF and a system consists of a single level of citation relationship [14].

III. PROPOSED APPROACH

The proposed system aims to decrease the effects of the majority of problems in the recommender system and improve a CF-based recommendation system, with three stages: the first preprocessing stage consists of reduction and text data cleaning. The second stage is model construction by employing the BERT algorithm and cosine similarity as the first phase, then CF using the KNN algorithm as the second phase. And the third stage is the evaluation of the results of the system using classification techniques. These phases are described in figure (1).

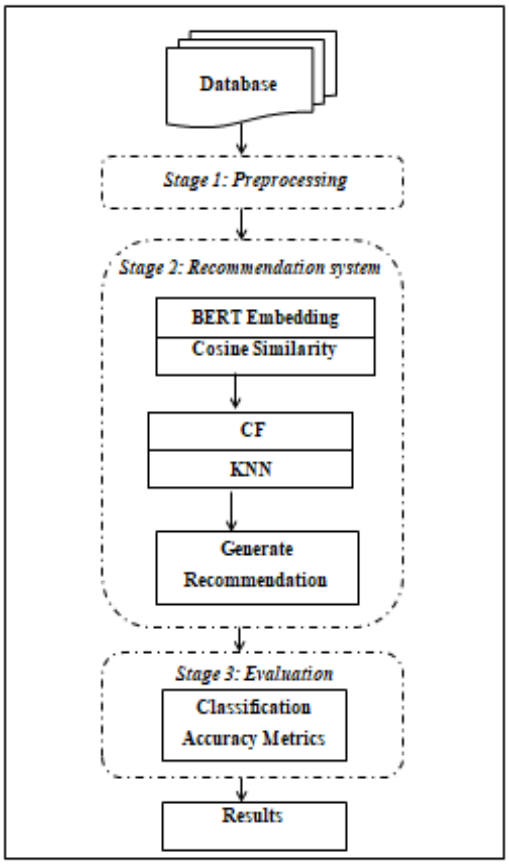

Figure 1: The Proposed system for improvement of the CF recommendation.

A. *Preprocessing*

There are a number of different kinds of data, each of which possesses its own individual set of properties. Data obtained from the real world typically requires some preprocessing, possibly because some of the data is duplicated, noisy, or missing values. In order to use the methods of machine learning in the proposed system, this is a requirement before the data can be used [15]. In Natural Language Processing (NLP), the preprocessing stage usually consists of several steps, including tokenization and stop-word removal [16]. In the proposed system, there was no need for these two steps due to the use of BERT. The stop word in the BERT algorithm is important and makes a difference to the meaning of the sentence, unlike other algorithms, so it is preserved. As for tokenization, it is created automatically in the BERT, so there is no need to do it in the processing stage. In this work, only two types of processing are required: reduction and normalization.

- *Reduction:* The database contains a large amount of redundant or irrelevant data. Data reduction is a technique employed in data mining that aims to decrease the size of a given dataset while preserving the most significant information. The technique of dimensionality reduction has been employed, wherein the superfluous features in a given dataset are eliminated in order to streamline the analysis. Within the reduction

process, books that were rated less than 100 times and users who rated less than 100 books were excluded in this way the Sparsity had been solved.

- *Normalization:* All uppercase letters can be changed to lowercase, symbols, signs, figures, or non-English characters can be removed, and irrelevant words or letters and punctuation that do not contribute to the text's content can be eliminated [17].

B. Building recommendation system

In order to improve the performance of CF and make the system provide more accurate recommendations, the proposed system uses BERT to filter the data as an initial stage, and then CF is applied to those results. This section goes over the details of the BERT and CF.

- *Bidirectional Encoder Representations from Transformers (BERT):* BERT is the encoder part of the transformer model, which is based on a self-attention mechanism. Language models are limited to reading textual input in a sequential manner, either from left-to-right or right-to-left, but not both simultaneously. BERT possesses a distinctive feature of bidirectional reading capability. The ability to process information in both forward and backward directions, commonly known as bi-directionality, has been made feasible with the advent of transformers [18]. The utilization of BERT embeddings for the purpose of extracting features from textual data, specifically pertaining to word and sentence embedding vectors. The vectors are employed as inputs with superior quality for models downstream. For instance, when it comes to encoding words, there have been two options available: one-hot encoding or neural word embeddings. where context-free, that means generate single word embedding representation for each word in the vocabulary in feature embeddings that were created by models such as Word2Vec or FastText [19]. After the pre-processing is done, the corpus is ready to apply the BERT algorithm. Firstly, data is embedded using the sentence transformer (bert-base-nli-mean-token). Secondly, calculate the similarity between the recommended book and other books using cosine similarity. Then, the similarity score should be sorted. Finally, obtain the highest N. The figure (2) shows steps to get books with most similarity to the query.

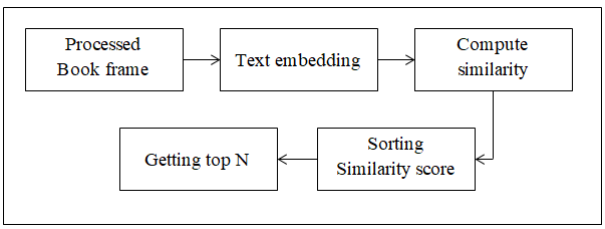

Figure 2: Similarity process

The most common inner product family metric is cosine similarity, which gauges how similar two vectors are. It may be used to show how similar two documents are within the context of text classification. It accepts values in the range of 0 and 1, with 0 denoting complete lack of resemblance and 1 denoting complete similarity between the documents [20].

$$\mathrm{Sim}(\mathrm{doc}\,1, \mathrm{doc}\,2) = \frac{\mathrm{doc}\,1 \cdot \mathrm{doc}\,2}{\|\mathrm{doc}\,1\|\|\mathrm{doc}\,2\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

Given two documents, where Ai and Bi stand for the elements of the corresponding vectors doc1 and doc2.

When calculating the semantic similarity between the required query and the rest of the database and then taking a specific part of it, with this process the number of data was reduced and irrelevant was excluded in this way the scalability had been solved.

- *Collaborative Filtering (CF):* In order to determine how to propose a product, CF first collects and analyzes a significant quantity of data on user priorities, attitudes, or activities [21]. The idea of CF is if customers A and B rate identical things, then their preferences will be seen as similar. Client A may be prescribed certain items if they are found in Client B's record but not in Client A's. In other words, CF is a method of making recommendations based on the evaluations of other clients [22]. The types of CF are: item and user based. The item-based CF assumes that users will like items that are like to items that the user has previously liked. The User-Based CF approach is a method in which to predict the preferences of a target user for a individual item by utilizes ratings provided by similar users. The main benefit of this strategy is that user reviews can be used to gauge an item's originality [23].
- *The KNN Collaborative Filtering algorithm:* The K-Nearst Neaghbor (KNN) is well-known in recommendation systems due to its fast predictive nature and short calculation time. KNN assigns any unlabeled class to its appropriate class based on a similarity measure [24]. KNN algorithm identifies the nearest neighboring data points from a given training dataset in relation to a query. The nearest data points are determined by calculating their distances from the query point. The algorithm employs a majority voting mechanism to ascertain the class that exhibits the highest frequency subsequent to identifying the k nearest data points [25]. The proposed system employs the KNN algorithm to compute the distance between the target book and the remaining books in the dataset with parameters " metric = 'cosine', algorithm='brute' ". Subsequently, the system arranges the top k books that are in close proximity to the target book by employing the cosine distance metric. The KNN collaborative filtering algorithm is a hybrid approach that integrates the KNN algorithm with collaborative filtering. It employs the KNN algorithm to select neighbors for recommendation purposes. The fundamental procedures of the algorithm involve the computation of user distance, selection of KNN nearest neighbor, and calculation of prediction score, as shown in figure (3).

This stage has only used the results of the BERT phase. Initially, the BERT result was combined with the rating frame. The items value evaluated by a pair of users is utilized to compute the measure of distance between them. Each user represents an item's score with an X-dimensional vector. After it has been determined the user's neighbors, now the score can be predicted based on the score of the item's neighbor. The following is the procedure for calculating the user's prediction score:

Step 1: Create a user-item two-dimensional matrix of the score as a pivot matrix, consisting of the book's title, user ID, and book rating value and convert it to array.

Step 2: Use the cosine distance principle to figure out how far away each user is from every other user and make the user distance matrix.

Step 3: Using the results from Step 2, find the M number of scores with the minimum weight; the corresponding M users are the user's neighbors.

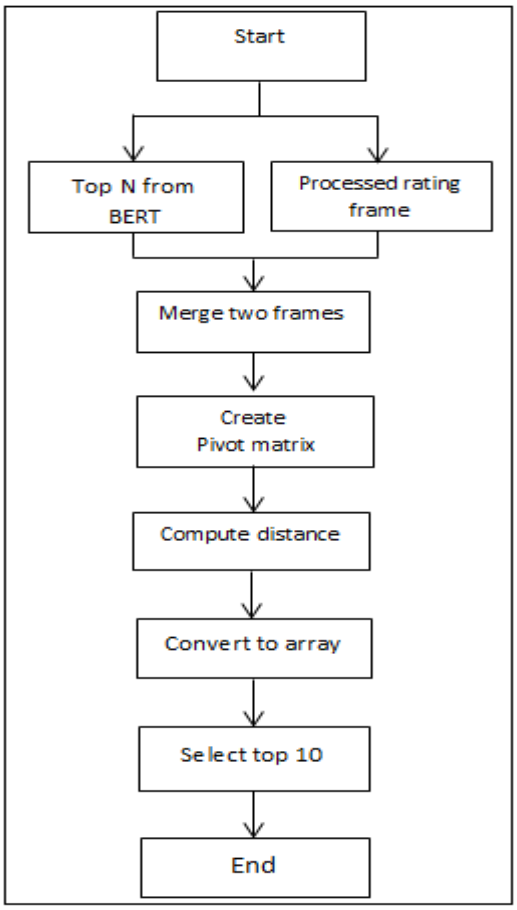

Figure 3: CF based on KNN

C. *Evaluation*

The precision, recall, F1-Score, and accuracy were used to evaluate the proposed system based on labeling. The data set is split into two parts: training uses 80% of it, and testing uses 20%. The system includes evaluation twice: once for the first evaluation stage of the CF in the recommendation prediction process, and once for the proposed system BERT with CF in the recommendation prediction process. For compare the results of the system in the normal case and our proposed system and prove its effectiveness. Each suggestion falls into one of the following categories:

- True Positive (TP): a suitable book is recommended.
- False Positive (FP): an unsuitable book is recommended.
- True Negative (TN): an unsuitable book that is not recommended.
- False Negative (FN): a suitable book is not recommended.

The Eq. (2) refers to the proportion of relevant things suggested relative to the total number of products recommended. The term "precision" refers to the likelihood that an item that has been suggested is applicable [26].

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

The proportion of relevant things advised to the overall number of relevant products is the definition of the recall equation, which is written as (3). The term "recall" refers to the likelihood that an item of relevance will be suggested [26].

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

F-measure is the total rate of accuracy and recall, which is defined by the equation (4). The F-measure is intended to provide a representation of the middle ground between accuracy and recall [27].

$$\text{F-measure} = 2 * \text{Precision} * \text{Recall} / (\text{Precision} + \text{Recall}) \qquad (4)$$

Accuracy (Ac) in (5) estimating the algorithm's effectiveness by displaying the likelihood of the true value of the class label; in other words, it evaluates the algorithm's overall effectiveness [27].

$$Ac = TP / (TP + FP + TN + FN) \qquad (5)$$

IV. EXPERIMENTAL AND RESULT ANALYSIS

A. *Dataset*

In this work, the proposed recommendation system has been performed on The Book-Crossing dataset, which consists of three frames: the first is a book frame with 271360 books. The second is a user frame with 278858 users. The third component is a rating frame with 1149780 ratings expressed on a scale of 1 to 10. The description of the database is shown in table (1).

Table 1: Data description

Description	value
Ratings of books	1149780
No. of Unique users	278858
No. of Unique books	271360

The dataset includes 1,000,000 ratings (on a scale of 1 to 10) given by 278,000 users to 271,000 different books. In this case, the summary will be taken to be the term of the book in order to identify similarities, and the book data frame contains a summary of the characteristics of the book.

B. Experimental results and discussion

The overall experimental result is shown in Table 2 and Figure 4 based on the outcome, we perceive that proposed approach outperforms the other approaches. With the Book-Crossing dataset, the experimental outcome shows that the lowest precision, recall and f1-score rate of 0.65, 0.62 and 0.63 is achieved with CF, and the highest precision, recall and f1-score rate of 0.89 for all is accomplished using BERT with CF based KNN. While CF based KNN achieved an average rate between the highest and lowest approaches to the precision, recall and f1-score of 0.66, 0.76 and 0.70.

Table 2: Evaluation value

Evaluation	CF	CF based KNN	BERT with CF based KNN
Precision	0.65	0.66	0.89
Recall	0.62	0.76	0.89
F1-Score	0.63	0.70	0.89
accuracy	0.62	0.76	0.89

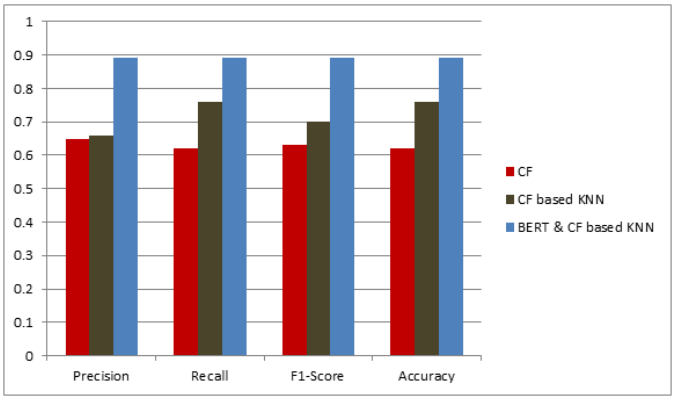

Figure 4: Evaluation ratio

V. CONCLUSIONS

A large amount of effort has been directed towards enhancing recommendation systems due to the emergence of opinion mining and sentiment analysis methodologies. The present study introduced a collaborative filtering approach that relies on semantic similarity in order to furnish recommendations. The approach that was proposed resulted in an enhancement of the accuracy of the book RS, with an increase to 0.89. Compared to the previous works on a dataset comprising 271,000 book summaries, proposed methodology has demonstrated superior outcomes by circumventing issues that were encountered in earlier studies, such as scalability and sparsity. This was achieved through the utilization of BERT in conjunction with CF-based KNN.

For future work, deep learning approaches can be used in recommender systems such as Gated Recurrent Units Long Short-Term Memory etc. to improve recommendations. A hybrid recommendation can be made using the current system and combined with a second type of recommendation filters in case there are another details about user profile and product profile. And execution module for different datasets (hotels, restaurants, travel destinations, music, and Twitter followers), if available, in addition to rich data with content and rating characteristics.

REFERENCES

[1] Zhu, Yifan, et al. "Recommending scientific paper via heterogeneous knowledge embedding based attentive recurrent neural networks." Knowledge-Based Systems 215 (2021): 106744.

[2] Nassar, Nour, Assef Jafar, and Yasser Rahhal. "A novel deep multi-criteria collaborative filtering model for recommendation system." Knowledge-Based Systems 187 (2020): 104811.

[3] Jannach, D., Manzoor, A., Cai, W., & Chen, L. "A survey on conversational recommender systems." ACM Computing Surveys (CSUR) 54.5 (2021): 1-36.

[4] [-31 Hui, B., Zhang, L., Zhou, X., Wen, X., & Nian, Y. "Personalized recommendation system based on knowledge embedding and historical behavior." *Applied Intelligence* (2022): 1-13.

[5] Sallam, Rouhia Mohammed, Mahmoud Hussein, and Hamdy M. Mousa. "Improving collaborative filtering using lexicon-based sentiment analysis." International Journal of Electrical and Computer Engineering 12.2 (2022): 1744.

[6] Geng, Li. "The Recommendation System of Innovation and Entrepreneurship Education Resources in Universities Based on Improved Collaborative Filtering Model." Computational Intelligence and Neuroscience 2022 (2022).

[7] Fkih, Fethi. "Similarity measures for Collaborative Filtering-based Recommender Systems: Review and experimental comparison." Journal of King Saud University-Computer and Information Sciences 34.9 (2022): 7645-7669.

[8] Zhou, Xin, and Wenan Tan. "An Improved Collaborative Filtering Algorithm Based on Filling Missing Data." Human Centered Computing: 6th International Conference, HCC 2020, Virtual Event, December 14–15, 2020, Revised Selected Papers 6. Springer International Publishing, 2021.

[9] Alhijawi, Bushra, and Yousef Kilani. "A collaborative filtering recommender system using genetic algorithm." Information Processing & Management 57.6 (2020): 102310.

[10] Sakib, Nazmus, Rodina Binti Ahmad, and Khalid Haruna. "A collaborative approach toward scientific paper recommendation using citation context." IEEE Access 8 (2020): 51246-51255.

[11] Neysiani, B. S., Soltani, N., Mofidi, R., & Nadimi-Shahraki, M. H. "Improve performance of association rule-based collaborative filtering recommendation systems

using genetic algorithm." International Journal of Information Technology and Computer Science 11.2 (2019): 48-55.

[12] Liu, Gaojun, and Xingyu Wu. "Using collaborative filtering algorithms combined with Doc2Vec for movie recommendation." 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). IEEE, 2019.

[13] Dubey, A., Gupta, A., Raturi, N., & Saxena, P. "Item-based collaborative filtering using sentiment analysis of user reviews." International Conference on Application of Computing and Communication Technologies. Springer, Singapore, 2018.

[14] Haruna, K., Akmar Ismail, M., Damiasih, D., Sutopo, J., & Herawan, T. "A collaborative approach for research paper recommender system." PloS one 12.10 (2017): e0184516.

[15] Rajasundari, T., P. Subathra, and P. N. Kumar. "Performance analysis of topic modeling algorithms for news articles." Journal of Advanced Research in Dynamical and Control Systems 11 (2017): 175-183.

[16] Al-augby, Salam, and Kesra Nermend. "USING RULE TEXT MINING BASED ALGORITHM TO SUPPORT THE STOCK MARKET INVESTMENT DECISION." Transformations in Business & Economics 14 (2015).

[17] Tong, Zhou, and Haiyi Zhang. "A text mining research based on LDA topic modelling." International conference on computer science, engineering and information technology. 2016.

[18] N. Adaloglou, "Transformers in Computer Vision," https://theaisummer.com/, 2021.

[19] C. McCormick and N. Ryan, "BERT Word Embeddings Tutorial," 2019, [Online]. Available: http://www.mccormickml.com.

[20] Park, Kwangil, June Seok Hong, and Wooju Kim. "A methodology combining cosine similarity with classifier for text classification." Applied Artificial Intelligence 34.5 (2020): 396-411.

[21] Gasmi, Sara, Tahar Bouhadada, and Abdelmadjid Benmachiche. "Survey on Recommendation Systems." Proceedings of the 10th International Conference on Information Systems and Technologies. (2020).

[22] Bai, Xiaomei, et al. "Scientific paper recommendation: A survey." Ieee Access 7 (2019): 9324-9339.

[23] Madia, Nidhi, Amit Thakkar, and Kamlesh Makvana. "Survey on recommendation system using semantic web mining." International Journal of Innovative and Emerging Research in Engineering 2.2 (2015).

[24] Gupta, M., Thakkar, A., Gupta, V., & Rathore, D. P. S. "Movie recommender system using collaborative filtering." 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC). IEEE, 2020.

[25] Uddin, S., Haque, I., Lu, H., Moni, M. A., & Gide, E. "Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction." Scientific Reports 12.1 (2022): 1-11.

[26] Chicco, Davide, and Giuseppe Jurman. "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation." BMC genomics 21 (2020): 1-13.

[27] Jumadi, J., Maylawati, D. S., Pratiwi, L. D., & Ramdhani, M. A. "Comparison of Nazief-Adriani and Paice-Husk algorithm for Indonesian text stemming process." IOP Conference Series: Materials Science and Engineering. Vol. 1098. No. 3. IOP Publishing, 2021.