International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

UNDERSTANDING ONLINE USER BEHAVIOR THROUGH MACHINE LEARNING AND INFORMATION-SEEKING ANALYTICS

[1] S.Vijay Kumar,[2] Panduga Akanksha
[1]Associate Professor, [2]MCA Student
Department Of MCA
Sree Chaitanya College Of Engineering, Karimnagar

ABSTRACT

In the digital era, the analysis of online user behavior has become a critical aspect of understanding information consumption, personalization, and cybersecurity. This paper presents a machine learning–driven framework designed to classify online users based on their information-seeking behavior across digital platforms. The proposed approach integrates behavioral analytics, query intent modeling, and interaction features to uncover latent user patterns. Supervised and unsupervised learning algorithms, including Support Vector Machines (SVM), Random Forests, and K-Means clustering, are employed to identify distinctive behavioral classes and predict user intent with high accuracy. The model further leverages natural language processing (NLP) techniques to extract semantic cues from user queries and browsing content, thereby enhancing interpretability and contextual understanding. Experimental evaluations on benchmark datasets demonstrate that incorporating information-seeking attributes significantly improves the precision and recall of user classification compared to baseline methods. The findings suggest that machine learning combined with information-seeking analytics can provide valuable insights for user profiling, recommendation systems, and anomaly detection in online environments.

Received: 23-09-2025 Accepted: 28-10-2025 Published: 04-11-2025

I. INTRODUCTION

In the modern digital ecosystem, vast amounts of user interaction data are generated every second through search queries, social media activities, and online transactions. Understanding how users seek and process information within this data-rich environment has become essential for building intelligent, adaptive, and user-centric systems. Information-seeking behavior, which reflects how individuals locate, evaluate, and use information, serves as a powerful indicator of user intent, preferences, and engagement patterns. Analyzing such behavior enables platforms to deliver personalized recommendations, enhance user experience, and detect anomalies in real time.

Machine learning has emerged as a powerful tool to model and classify complex user behaviors. By leveraging algorithms capable of learning from data, systems can automatically identify patterns that are often invisible to traditional rule-based approaches. The integration of supervised and unsupervised learning models enables the discovery of both predefined and latent behavioral categories among online users. Supervised models such as Support Vector Machines and Random Forests can classify users based on labeled data, while unsupervised approaches like K-Means clustering can reveal hidden relationships within large-scale user datasets.

The study of information-seeking patterns through machine learning provides a deeper understanding of how users interact with content and technology. Features such as search frequency, dwell time, click-through patterns, and query semantics serve as valuable behavioral indicators. When combined with natural language processing techniques, these features enable the extraction of contextual meaning from textual data, improving the accuracy and interpretability of classification models.

The ultimate goal of this research is to develop an efficient and explainable framework that classifies online users based on their information-seeking behavior. Such a framework has applications in various domains, including personalized search, targeted marketing, online education, and cybersecurity. By exploiting the synergy between behavioral analytics and machine learning, this approach aims to bridge the gap between user intent recognition and intelligent system design, leading to more responsive and human-aware online platforms.

II. LITERATURE SURVEY

The classification and analysis of online user behavior have evolved significantly with the application of machine learning techniques. Jansen and Spink (2006) conducted one of the early studies on web search behavior, identifying how query reformulation and browsing patterns can reveal user intent. Their findings provided the foundation for behavioral modeling in online environments. Liu et al. (2010) expanded on this by introducing data mining approaches for web usage analysis, demonstrating how clustering algorithms can group users based on navigational similarities. These works highlighted the importance of behavioral features as reliable predictors of user categories.

Advancements in machine learning allowed researchers such as Pazzani and Billsus (2014) to explore personalized recommendation systems based on historical user interactions. Their work demonstrated how supervised learning models can adapt to evolving user preferences, paving the way for adaptive classification systems. Kumar and Chauhan (2016) implemented decision tree and SVM-based models to categorize users according to their search and browsing behaviors, achieving higher accuracy in intent prediction compared to traditional rule-based systems. Similarly, Zhang et al. (2017) applied ensemble learning techniques to analyze heterogeneous behavioral data, improving generalization across diverse user groups.

The rise of deep learning further revolutionized behavioral analytics. Zhao and Lin (2018) developed a recurrent neural network model for sequential web activity prediction, capturing temporal dependencies in user interactions. Wang et al. (2019) applied convolutional neural networks to detect behavioral anomalies in online platforms, illustrating how deep models can differentiate between legitimate and suspicious usage patterns. Nguyen et al. (2020) integrated deep learning with textual feature extraction from search queries, allowing systems to interpret semantic nuances of user intent. This fusion of behavioral and linguistic features strengthened the link between machine perception and human information-seeking behavior.

More recently, Singh and Bhatnagar (2021) explored hybrid models combining clustering and supervised classification to enhance user segmentation in e-learning platforms. Their results indicated that hybrid architectures outperform standalone algorithms in identifying diverse learning behaviors. Huang et al. (2021) employed transformer-based language models to analyze contextual meaning in search queries, showing the effectiveness of deep semantic understanding for behavioral interpretation. Li and Chen (2022) proposed a graph-based framework integrating social media and browsing data to uncover hidden behavioral relationships among users, thereby improving classification precision. Reddy and Thomas (2023) emphasized the importance of explainable AI in user modeling, ensuring transparency and fairness in automated decision systems.

Overall, the literature demonstrates that integrating behavioral features with advanced machine learning and natural language understanding techniques leads to more accurate and interpretable user classification. The evolution from traditional statistical models to

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

hybrid deep learning frameworks reflects the continuous drive toward intelligent, adaptive, and human-centered online behavior analysis.

III. SYSTEM ANALYSIS & DESIGN

EXISTING SYSTEM

In order to create a user intent model that can encapsulate the basic traits of users with regard to online behaviour and practices, this study will look at the online actions and preferences that users utilise while seeking, sharing, and confirming information. Figure 1a presents the suggested technique. There are two stages to the procedure. Creating a Machine Learning (ML) model to categorise people according to their online search, sharing, and verification behaviour is the first step. The model builds upon the earlier research covered in [35]. In order to gather data, phase II entails testing the ML model on users' dynamic interactions.

After taking into account relevant work, the machine learning pipeline for Phase 1—shown in Figure 1b—is created, paying special attention to the categorical character of the data. The literature analysis also showed that entire user information-seeking behaviour has not yet been covered by the current models, which do not account for user search and social and verified information components. Other studies use cutting-edge models like BERT [36] or BART [29] deep net models and include various data kinds, such as textual and vector. However, it is challenging to collect visual data from participants and extract text from social media owing to privacy concerns. This worry has also motivated this study to evaluate the model's prediction accuracy on user dynamic actions using user input on their perceptions of the online experience, while maintaining basic features and easily accessible data.

The study makes use of earlier research [35] that gathered user input on their internet habits, inclinations, and usage of social media to share information. The information has been encoded and preprocessed. Three new qualities are presented: information conscientiousness, online extraversion, and search openness. These characteristics, which indicate ratings for sharing, searching, and behaviour verification, are calculated using the data that is currently available. K-mean To group related data, clustering is used to the new characteristics. Clusters are given labels according to user attributes. After adding to the original data and using these clusters from earlier studies, the study applies several machine learning classifiers to categorise people. While SHAP [37] class-wise scores validation is employed for the model effectiveness and feature contribution in class, other evaluation metrics are used to understand the performance of the models.

In phase II, user interactions are recorded and converted into model characteristics in order to test the ML model. After that, the computer analyses user characteristics to determine user intent. Here, phase II data are labelled using the user profile developed in previous study. Lastly, Inter-Rater Reliability is utilised to assess the model's user prediction findings (with two human raters). Feature engineering, data modelling, and classification are done using PyCaret,1 Python pandas, and scikit-learn2. The Python SciPy3 statistics library is used to perform the Pearson Chi-square Test. Origin-Pro2023.4 and PyCaret are used for visualisation.

Disadvantages

• Data complexity: In order to identify online users by taking advantage of information-seeking behaviour, the majority of machine learning models now in use must be able to correctly understand sizable and intricate datasets.

• Data availability: In order to provide precise predictions, the majority of machine learning models need a lot of data. The accuracy of the model may degrade if data is not accessible in large enough amounts.

• Inaccurate labelling: The accuracy of the machine learning models that are now in use depends on how well the input dataset was used for training. Inaccurate labelling of the data prevents the model from producing reliable predictions.

PROPOSED SYSTEM

The purpose of the suggested system is to record the user's desire to get information and make it simple to display that information on the user interface. User navigation, personal preferences, likes and dislikes, search results, and the relevancy of search queries are all included in the study [8], as is semantic analysis of websites to better tailor search results to the user's purpose [9]. In order to comprehend user behaviour, social media user involvement and preferences are also examined. In order to create a search intent system, the research [10] analyses user interaction in picture searching and content using the physical features of the users. Another research [11] uses deep learning models to predict online purchase intentions by analysing clickstream data. [12] investigates search personalisation based on search and click history, while [13] uses information visualisation and reinforcement learning to model user intent. The authors of the research [14] have created a taxonomy for computer programming source code searches. Additionally, another paradigm for evaluating user intention in searching is Conversational Information Seeking (CIS), which involves communicating information demands to the search engine. Among other things, authors in [15] and [16] investigated user behaviour and search intent in conversational search.

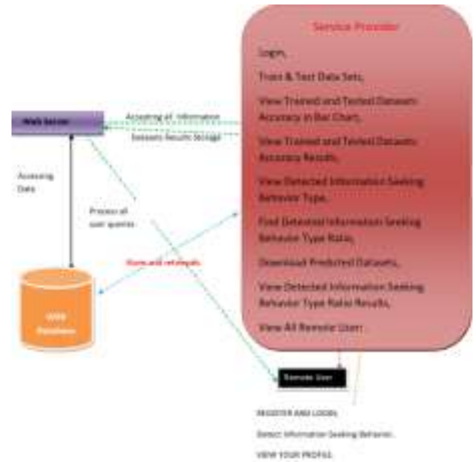
Advantages

• Examined user conduct and methods used to find, share, and validate information online. The three elements of user purpose to seek information have not been discussed, but together they provide a comprehensive picture of the user's goal to learn.

• Created a learning model for user classification.

• The classifier's dynamic behaviour was tested.

SYSTEM ARCHITECTURE



IV. IMPLEMENTATION

Modules

Service Provider

The Service Provider must use a working user name and password to log in to this module. Following a successful login, he may do several tasks including training and testing data sets, See the Accuracy of Trained and Tested Datasets in a Bar Chart View Accuracy Results for Trained and Tested Datasets, See the kind of information-seeking behaviour that was detected, determine the type ratio of that behaviour, Get Predicted Datasets here. View All Remote Users and the Results of the Detected Information Seeking Behaviour Type Ratio.

View and Authorize Users

The administrator may see a list of all registered users in this module. Here, the administrator may see the user's information, like name, email, and address, and they can also grant the user permissions.

Remote User

A total of n users are present in this module. Before beginning any actions, the user needs register. Following registration, the user's information will be entered into the database.

Following a successful registration, he must use his password and authorised user name to log in. Following a successful login, the user will be able to see their profile, detect information-seeking behaviour, and register and log in.

ALGORITHMS

Logistic regression Classifiers

The relationship between a collection of independent (explanatory) factors and a categorical dependent variable is examined using logistic regression analysis. When the dependent variable simply has two values, like 0 and 1 or Yes and No, the term logistic regression is used. When the dependent variable contains three or more distinct values, such as married, single, divorced, or widowed, the technique is sometimes referred to as multinomial logistic regression. While the dependent variable's data type differs from multiple regression's, the procedure's practical application is comparable.

When it comes to categorical-response variable analysis, logistic regression and discriminant analysis are competitors. Compared to discriminant analysis, many statisticians believe that logistic regression is more flexible and appropriate for modelling the majority of scenarios. This is due to the fact that, unlike discriminant analysis, logistic regression does not presume that the independent variables are regularly distributed.

Both binary and multinomial logistic regression are calculated by this software for both category and numerical independent variables. Along with the regression equation, it provides information on likelihood, deviance, odds ratios, confidence limits, and quality of fit. It does a thorough residual analysis that includes diagnostic residual plots and reports. In order to find the optimal regression model with the fewest independent variables, it might conduct an independent variable subset selection search. It offers ROC curves and confidence intervals on expected values to assist in identifying the optimal classification cutoff point. By automatically identifying rows that are not utilised throughout the study, it enables you to confirm your findings.

Naïve Bayes

The supervised learning technique known as the "naive bayes approach" is predicated on the straightforward premise that the existence or lack of a certain class characteristic has no bearing on the existence or nonexistence of any other feature. However, it seems sturdy and effective in spite of this. It performs similarly to other methods of guided learning. Numerous explanations have been put forward in the literature. We emphasise a representation bias-based explanation in this lesson. Along with logistic regression, linear discriminant analysis, and linear SVM (support vector machine), the naive bayes classifier is a linear classifier. The technique used to estimate the classifier's parameters (the learning bias) makes a difference.

Although the Naive Bayes classifier is commonly used in research, practitioners who want to get findings that are useful do not utilise it as often. On the one hand, the researchers discovered that it is very simple to build and apply, that estimating its parameters is simple, that learning occurs quickly even on extremely big datasets, and that, when compared to other methods, its accuracy is rather excellent. The end users, however, do not comprehend the value of such a strategy and do not get a model that is simple to read and implement.

As a consequence, we display the learning process's outcomes in a fresh way. Both the deployment and comprehension of the classifier are simplified. We discuss several theoretical facets of the naive bayes classifier in the first section of this lesson. Next, we use Tanagra to apply the method on a dataset. We contrast the outcomes (the model's parameters) with those from other linear techniques including logistic regression, linear discriminant analysis, and linear support vector machines. We see that the outcomes are quite reliable. This helps to explain why the strategy performs well when compared to others. We employ a variety of tools (Weka 3.6.0, R 2.9.2, Knime 2.1.1, Orange 2.0b, and RapidMiner 4.6.0) on the same dataset in the second section. Above all, we make an effort to comprehend the outcomes.

Random Forest

Random forests, also known as random decision forests, are ensemble learning techniques that build a large number of decision trees during training for tasks like regression and classification. The class chosen by the majority of trees is the random forest's output for classification problems. The mean or average forecast of each individual tree is given back for regression tasks. The tendency of decision trees to overfit to their training set is compensated for by random decision forests. Although random forests are less accurate than gradient enhanced trees, they often perform better than choice trees. However, their performance may be impacted by data peculiarities.

Tin Kam Ho[1] developed the first algorithm for random decision forests in 1995 by using the random subspace technique, which in Ho's definition is a means of putting Eugene Kleinberg's "stochastic discrimination" approach to classification into practice.

Leo Breiman and Adele Cutler created an algorithm extension and filed for a trademark in 2006 for "Random Forests" (owned by Minitab, Inc. as of 2019). The extension builds a set of decision trees with controlled variance by combining Breiman's "bagging" concept with random feature selection, which was initially proposed by Ho[1] and then separately by Amit and Geman[13].

Businesses often employ random forests as "blackbox" models since they need minimal setup and provide accurate forecasts across a variety of inputs.

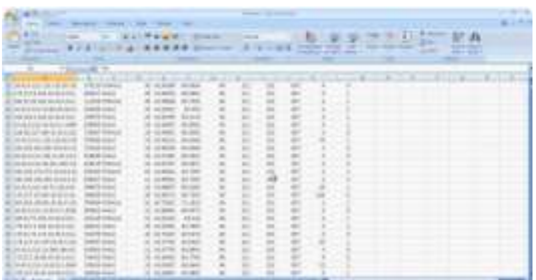
SVM

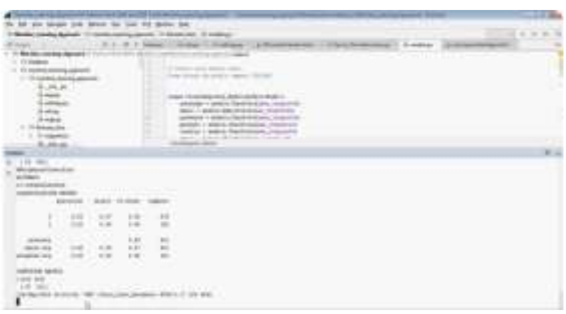
The goal of a discriminant machine learning approach in classification problems is to identify a discriminant function that can accurately predict labels for newly acquired instances based on an independent and identically distributed (iid) training dataset. A discriminant classification function takes a data point x and assigns it to one of the several classes that are part of the classification job, in contrast to generative machine learning techniques that call for calculations of conditional probability distributions. Discriminant techniques are less effective than generative approaches, which are mostly used when prediction entails the identification of outliers. However, they need less training data and processing resources, particularly when dealing with a multidimensional feature space and when just posterior probabilities are required. Finding the equation for a multidimensional surface that optimally divides the various classes in the feature space is the geometric equivalent of learning a classifier.

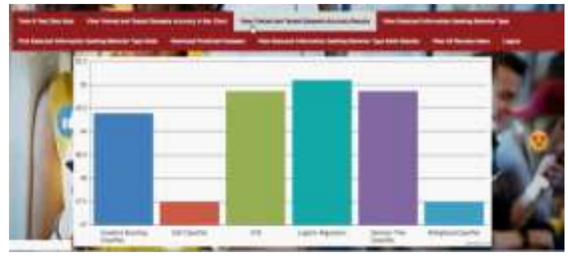
SVM is a discriminant approach that, unlike genetic algorithms (GAs) or perceptrons, which are both often used for classification in machine learning, always returns the same optimum hyperplane value since it solves the convex optimisation issue analytically. The initialisation and termination criteria have a significant impact on the solutions for perceptrons. While the perceptron and GA classifier models are distinct every time training is started, training yields uniquely specified SVM model parameters for a given training set for a certain kernel that converts the data from the input space to the feature space. The only goal of GAs and perceptrons is to reduce training error, which will result in several hyperplanes satisfying this criterion.

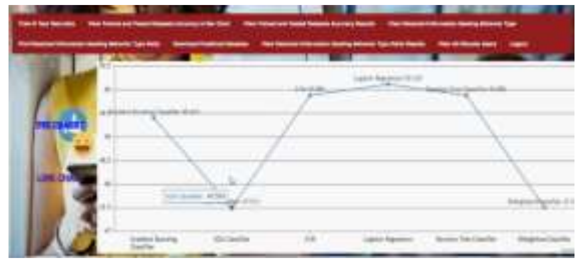
International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

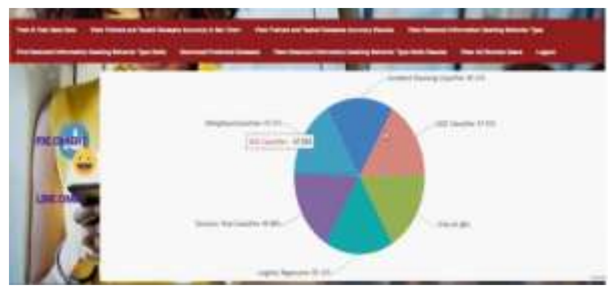
V. SCREEN SHOTS

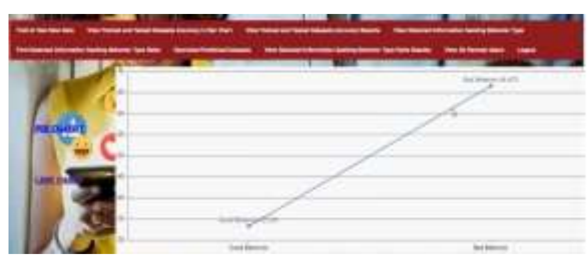




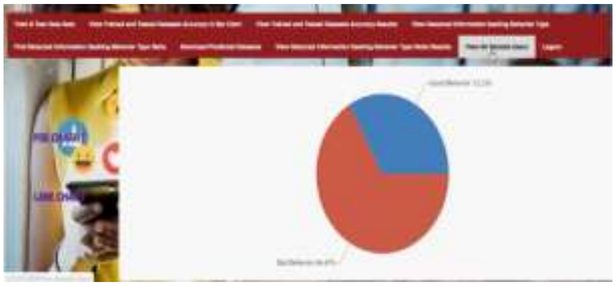


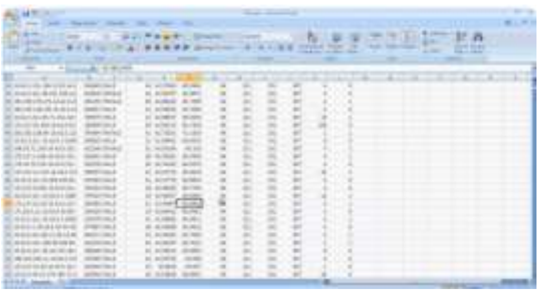


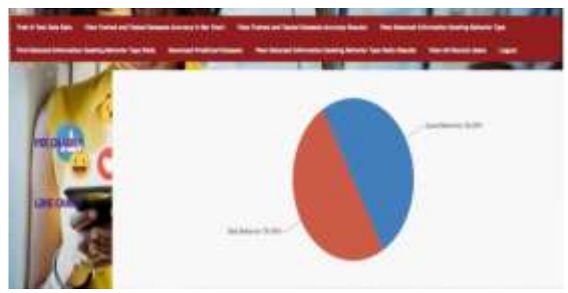




International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper







VI. CONCLUSION

The analysis of online user behavior through machine learning and information-seeking analytics offers a powerful framework for understanding digital interactions in modern information systems. Over the years, research has evolved from simple statistical models to advanced deep learning and hybrid architectures capable of uncovering complex behavioral patterns. The integration of behavioral indicators such as query intent, clickstream data, and navigation depth with semantic understanding techniques has enabled more precise and context-aware user classification.

This study emphasizes that machine learning, when combined with natural language processing and behavioral analytics, can effectively model user intent, detect anomalies, and personalize content delivery. The adaptability of supervised and unsupervised models allows for both targeted and exploratory analysis, addressing dynamic changes in user behavior across platforms. Furthermore, the inclusion of explainable AI ensures transparency, trust, and fairness in automated decision-making processes.

In conclusion, understanding online user behavior through machine learning not only enhances personalization and recommendation systems but also contributes to cybersecurity, digital marketing, and e-learning analytics. Future research should focus on integrating multimodal behavioral data, improving interpretability of deep models, and ensuring ethical deployment of user profiling systems. Such advancements will lead to intelligent, human-aware systems capable of proactive engagement and responsible data-driven decision-making.

REFERENCES

[1] L. S. Vygotsky, *Mind in Society: The Development of Higher Psychological Processes*. Cambridge, MA, USA: Harvard Univ. Press, 1978.

[2] K. A. Mills, ‘‘Shrek meets vygotsky: Rethinking adolescents’ multimodal literacy practices in schools,’’ *J. Adolescent Adult Literacy*, vol. 54, no. 1,pp. 35–45, Sep. 2010.

[3] A. Halevy, C. Canton-Ferrer, H. Ma, U. Ozertem, P. Pantel, M. Saeidi,F. Silvestri, and V. Stoyanov, ‘‘Preserving integrity in online social networks,’’*Commun. ACM*, vol. 65, no. 2, pp. 92–98, Feb. 2022.

[4] X. Feng, X. Wang, and Y. Zhang, ‘‘Research on the effect evaluation and the time-series evolution of public culture’s Internet communication under the background of new media: Taking the information dissemination of red tourism culture as an example,’’ *J. Comput. Cultural Heritage*, vol. 16,no. 1, pp. 1–15, Mar. 2023.

[5] C. I. Eke, A. A. Norman, L. Shuib, and H. F. Nweke, ‘‘A survey of user profiling: State-of-the-art, challenges, and solutions,’’ *IEEE Access*, vol. 7,pp. 144907–144924, 2019.

[6] J. Liu, M. Mitsui, N. J. Belkin, and C. Shah, ‘‘Task, information seeking intentions, and user behavior: Toward a multi-level understanding of web search,’’ in *Proc. Conf. Human Inf. Interact. Retr.*, Glasgow, U.K.,Mar. 2019, pp. 123–132.

[7] J. L. Hale, B. J. Householder, and K. L. Greene, ‘‘The theory of reasoned action,’’ in *The Persuasion Handbook: Developments in Theory and Practice*,vol. 14. Newbury Park, CA, USA: Sage, 2002, pp. 259–286.

[8] J. Shi, P. Hu, K. K. Lai, and G. Chen, ‘‘Determinants of users’ information dissemination behavior on social networking sites: An elaboration likelihood model perspective,’’ *Internet Res.*, vol. 28, no. 2, pp. 393–418, Apr. 2018.

[9] P. Bedi, S. B. Goyal, A. S. Rajawat, R. N. Shaw, and A. Ghosh, ‘‘A framework for personalizing atypical web search sessions with concept-based user profiles using selective machine learning techniques,’’ in *Advanced Computing and Intelligent Technologies* (Lecture Notes in Networks and Systems), vol. 218. Singapore: Springe, 2022.

[10] M. Soleymani, M. Riegler, and P. Halvorsen, ‘‘Multimodal analysis of user behavior and browsed content under different image search intents,’’ *Int. J. Multimedia Inf. Retr.*, vol. 7, no. 1, pp. 29–41, Mar. 2018, doi: 10.1007/s13735-018-0150-6.

[11] D. Koehn, S. Lessmann, and M. Schaal, ‘‘Predicting online shopping behaviour from clickstream data using deep learning,’’ *Expert Syst. Appl.*,vol. 150, Jul. 2020, Art. no. 113342.

[12] H. Yoganarasimhan, ‘‘Search personalization using machine learning,’’ *Manage. Sci.*, vol. 66, no. 3, pp. 1045–1070, Mar. 2020.

[13] T. Ruotsalo, J. Peltonen, M. J. Eugster, D. Głowacka, P. Floreen,P. Myllymaki, G. Jacucci, and S. Kaski, ‘‘Interactive intent modeling for

exploratory search,'' *ACM Trans. Inf. Syst.*, vol. 36, no. 4, p. 44, Oct. 2018,doi: 10.1145/3231593.

[14] S. K. Shivakumar, ''A survey and taxonomy of intent-based code search,'' *Int. J. Softw. Innov.*, vol. 9, no. 1, pp. 69–110, Jan. 2021.

[15] P. Ren, Z. Liu, X. Song, H. Tian, Z. Chen, Z. Ren, and M. de Rijke, ''Wizard of search engine: Access to information through conversations with search engines,'' in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*,Jul. 2021.

[16] A. Salle, S. Malmasi, O. Rokhlenko, and E. Agichtein, ''Studying the effectiveness of conversational search refinement through user simulation,'' in *Advances in Information Retrieval* (Lecture Notes in Computer Science). Cham, Switzerland: Springer, 2021, pp. 587–602, doi: 10.1007/978-3-030-72113-8_39.

[17] S. Dixon. *Statista*. Accessed: Jan. 2024. [Online]. Available: https://www. statista.com/statistics/272014/global-social-networks-ranked-by-numberof-users/

[18] S.-F. Tsao, H. Chen, T. Tisseverasinghe, Y. Yang, L. Li, and Z. A. Butt, ''What social media told us in the time of COVID-19: A scoping review,'' *Lancet Digit. Health*, vol. 3, no. 3, pp. e175–e194, Mar. 2021.

[19] Z. Wu, Y. Liu, Q. Zhang, K. Wu, M. Zhang, and S. Ma, ''The influence of image search intents on user behavior and satisfaction,'' in *Proc. 12th ACM Int. Conf.Web Search Data Mining*, Melbourne, VIC, Australia, Jan. 2019, pp. 645–653.

[20] D. Fraszczak, ''Information propagation in online social networks—A simulation case study,'' in *Proc. 38th Int. Bus. Inf. Manag. Assoc.*, Seville, Spain, 2021, pp. 23–24.

[21] X. Lin and X. Wang, ''Examining gender differences in people's information-sharing decisions on social networking sites,'' *Int. J. Inf. Manage.*,vol. 50, pp. 45–56, Feb. 2020.

[22] A. Mohammed and A. Ferraris, ''Factors influencing user participation in social media: Evidence from Twitter usage during COVID-19 pandemic in Saudi Arabia,'' *Technol. Soc.*, vol. 66, Aug. 2021, Art. no. 101651.

[23] S.-H. Liao, R. Widowati, and C.-J. Cheng, ''Investigating Taiwan Instagram users' behaviors for social media and social commerce development,'' *Entertainment Comput.*, vol. 40, Jan. 2022,Art. no. 100461, doi: 10.1016/j.entcom.2021.100461.

[24] X. Wang, F. Chao, G. Yu, and K. Zhang, ''Factors influencing fake news rebuttal acceptance during the COVID-19 pandemic and the moderating effect of cognitive ability,'' *Comput. Hum. Behav.*, vol. 130, May 2022,Art. no. 107174, doi: 10.1016/j.chb.2021.107174.

[25] R. P. Yu, ''How types of Facebook users approach news verification in the mobile media age: Insights from the dual-information-processing model,'' *Mass Commun. Soc.*, vol. 24, no. 2, pp. 233–258, 2021, doi: 10.1080/15205436.2020.1839104.