Machine Learning Approach to Classification of Online Users By Exploiting Information Seeking Behavior

[1]Mrs. RENUKA B.N [2] RACHANA A JAIN

[1]Assistant Professor, Department of MCA, BIET, Davanagere
[2] Student,4th Semester MCA, Department of MCA, BIET, Davanagere

ABSTRACT

With the exponential growth of digital content, the internet has become a vast repository of unstructured and continuous data originating from diverse sources. Advanced algorithms are increasingly employed to deliver relevant information based on a user's intent. Online user behaviour involves various actions such as searching, sharing, and verifying information; however, a holistic analysis of these behaviours remains relatively unexplored. This research presents a machine learning-based approach to classify users according to their intent, derived from their online activities across three key dimensions: search, share, and verification behaviours. A comprehensive dataset was collected through questionnaires involving participants of varying age groups, occupations, and genders.

Keywords: Online user behavior, machine learning, user intent classification, information seeking, search behavior, sharing behavior, verification behavior, dynamic interaction analysis.

1. INTRODUCTION

In the digital age, the internet has become an indispensable source of information, with users constantly engaging in activities such as searching, sharing, and verifying content. These actions form a pattern of behavior that reflects a user's intent and preferences while interacting with online platforms. However, understanding this behavior in a structured and meaningful way remains a significant challenge due to the volume, variety, and complexity of user data generated across digital platforms.

This project aims to bridge that gap by leveraging machine learning techniques to classify online users based on their information-seeking behavior. By analyzing how individuals search for, disseminate, and authenticate information, the system seeks to identify distinct user profiles. These profiles are formed using clustering algorithms and later classified using supervised learning methods to predict user intent with high accuracy.

The classification of users is not only beneficial for understanding online behavior but also holds immense value in areas such as personalized content delivery, targeted marketing, and enhancing search engine results. The study further explores dynamic user interactions to validate the model's effectiveness in real-time scenarios. Through this work, a comprehensive model is developed that considers all three critical aspects of user engagement—searching, sharing, and verifying—offering a holistic approach to online behavior analysis.

2. LITERATURE REVIEW

"A machine learning approach to user profiling for data annotation of online behavior," CMC-Comput. Mater. Con., vol. 74, no. 1, pp. 123–135, Feb. 2024. This study presents a machine learning-based framework for user profiling aimed at improving data annotation in online behavior analytics. The proposed method uses clustering and classification algorithms to analyze user activities and group behavioral patterns. By leveraging historical data and user interactions, the model is trained to identify relevant attributes that distinguish user types. This allows for more accurate labeling and segmentation of user profiles, which is essential for targeted marketing, personalized recommendations, and anomaly detection. M. Kanwal, N. A. Khan, and A. A. Khan.[1]

"Explainable AI for machine fault diagnosis: Understanding features' contribution in machine learning models for industrial condition monitoring," J. Ind. Inf. Integr., vol. 40, pp. 100–115, Feb. 2023. This article emphasizes the importance of explainability in artificial intelligence applications for industrial fault diagnosis. The authors propose a framework that combines machine learning algorithms with interpretability techniques, allowing engineers and operators to understand how and why decisions are made. E. Brusa, L. Cibrario, C. Delprete, and L. G. Di Maggio.[2]

"Data clustering: Application and trends," Artif. Intell. Rev., vol. 55, no. 9, pp. 7453–7479, Nov. 2022. This comprehensive review explores the evolving applications and methodologies of data clustering in artificial intelligence. Clustering is an unsupervised learning technique that partitions datasets into meaningful groups based on similarity metrics. The article reviews traditional clustering methods such as K-means and hierarchical clustering, as well as modern approaches including DBSCAN, spectral clustering, and density-based techniques. G. J. Oyewole and G. A. Thopil.[3]

"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, Oct. 2022. This research investigates the dynamic process and impact of online public culture communication in the context of new media. Using red tourism culture as a case study, the authors analyze how cultural narratives evolve over time through digital platforms. The methodology involves time-series analysis combined with natural language processing and sentiment analysis to measure the reach, reception, and transformation of cultural content. X. Feng, X. Wang, and Y. Zhang.[4]

"Social media behaviour analysis in disaster-response messages of floods and heat waves via artificial intelligence," AI Soc., vol. 38, no. 2, pp.

521–539, Jun. 2022. This article explores the application of AI in analyzing social media communication during natural disasters, specifically floods and heat waves. The authors develop a framework that uses natural language processing and sentiment analysis to categorize messages related to disaster response. Data from Twitter and other platforms are collected and filtered using AI classifiers to detect urgency, location relevance, and emotional tone. V. Ponce-López and C. Spataru.[5]

"A survey of human-in-the-loop for machine learning," Future Gener. Comput. Syst., vol. 135, pp. 328–346, May 2022. This survey provides a comprehensive analysis of human-in-the-loop (HITL) techniques in machine learning, emphasizing the synergy between human expertise and automated learning systems. The authors categorize HITL methodologies based on the degree and stage of human involvement, including data labeling, model training, evaluation, and refinement. X. Wu, L. Xiao, S. Yixuan, and L. He.[6]

"Fake news detection based on news content and social contexts: A transformer-based approach," J. Inf. Technol. Polit., vol. 19, no. 1, pp. 42–59, Jan. 2022. This paper presents a novel transformer-based model for fake news detection, integrating both content features and social context to enhance classification accuracy. Unlike conventional methods that focus solely on textual content, this approach incorporates metadata such as user profiles, posting patterns, and propagation networks. The model uses a transformer architecture—specifically BERT—to capture the semantic relationships within the text, while graph-based features help understand how misinformation spreads across platforms. S. Raza and C. Ding.[7]

"This is fake! Shared it by mistake: Assessing the intent of fake news spreaders," in Proc. Int. AAAI Conf. Web Social Media, Apr. 2022. This study investigates the underlying motivations of users who share fake news online, distinguishing between malicious intent and accidental sharing. The authors develop a machine learning framework to classify intent based on linguistic features, sharing patterns, and user metadata. By analyzing user behavior across multiple platforms, the study identifies traits associated with intentional spreaders versus those misled by deceptive content. X. Zhou, K. Shu, V. V. Phoha, and R. Zafarani. [8]

"A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects," Eng. Appl. Artif. Intell., vol. 110, pp. 104743, Apr. The authors categorize clustering techniques into partition-based, hierarchical, density-based, grid-based, and model-based methods. Each category is analyzed in terms of computational complexity, scalability, and suitability for different types of data. A. E.-S. Ezugwu, A. M. Ikotun, O. Olaide, and A. Akinyelu.[9]

"Preserving integrity in online social networks," Commun. ACM, vol. 65, no. 2, pp. 92–98, Feb. 2022. This paper explores techniques to preserve information integrity in online social networks, a pressing issue amid rising misinformation and coordinated manipulation campaigns. The authors outline architectural and algorithmic solutions that enhance the trustworthiness of content shared on platforms. The paper emphasizes a user-centric design philosophy, advocating transparency in moderation policies and user access to content history and verification tools. A. Halevy, C. Canton-Ferrer, H. Ma, and V. Stoyanov.[10]

3. METHODOLOGY

3.1 Remote User

The **Remote User (Client Side)** interacts directly with the system to perform actions such as searching for information, sharing content, and verifying data. As part of the process, the user fills out behavior-related questionnaires that help in gathering insights into their online habits. Additionally, the system observes the user's real-time activities, including how they search and share information. Based on this behavior, the system analyzes and classifies the user into specific intent categories such as Searcher, Sharer, or Verifier, providing a personalized and adaptive experience.

4. EXISTING SYSTEM

This study focuses on examining users' online behaviors and preferences related to how they search for, share, and verify information. The goal is to construct a user intent model that captures common behavioral patterns in digital environments. As depicted in Figure 1a, the research process is divided into two primary phases.

Involves developing a Machine Learning (ML) model to categorize users based on their online activities concerning searching, sharing, and verifying information. This work builds upon the foundation laid in a previous study. The model is tested against real-time user interactions to evaluate its ability to identify behavior

The machine learning pipeline designed for Phase I is shown in Figure 1b. It was developed with careful consideration of the categorical nature of the input data, as highlighted in the literature review. Previous studies have not comprehensively addressed user behavior in terms of searching, social sharing, and verification as an integrated model. Although some research utilizes diverse data formats, including textual and vector data, and applies advanced models like BERT or BART, challenges in accessing user-generated content due to privacy concerns have limited their applicability.

Following the methodology of prior work, user feedback is collected regarding their internet usage patterns, preferences, and social media involvement. The collected data is pre-processed and encoded. Three behavioral attributes—Search Openness, Online Extraversion, and Information Conscientiousness—are introduced. These attributes quantify users' tendencies in searching, sharing, and verifying information, respectively. K-means clustering is then applied to these attributes to identify user groups with similar behaviors. Clusters are labeled according to the characteristics of the users within them. These labeled clusters are merged with the original dataset and fed into various machine learning classifiers for user classification. Model performance is evaluated using standard metrics, while SHAP is employed to analyze the importance of each feature in classspecific predictions.

The trained model is tested using dynamic user interaction data, which is transformed into feature sets compatible with the model. User attributes derived from these interactions are analyzed to assess how the model interprets user intent. Labels from earlier profiling work are reused to annotate this data. Model predictions are further validated using Inter-Rater Reliability, involving assessments by two independent human raters. Tools such as PyCaret, pandas, and scikit-learn are utilized for feature engineering, modeling, and classification.

5. PROPOSED SYSTEM

The proposed system aims to effectively capture user intent during information retrieval and simplify the process of delivering relevant content to the user's screen. This research encompasses various aspects, including user navigation patterns, individual preferences, likes and dislikes, the relevance of search queries, and search outcomes. It also incorporates semantic analysis of websites to tailor search results more closely to user intent. To gain deeper insights into user behavior, the study considers user interactions and preferences on social media platforms. For instance, one study explores how users' physical characteristics influence their interactions during image searches, contributing to the development of an intent-based search system. Another research effort utilizes clickstream data and deep learning models to predict users' online shopping intentions. Additionally, personalization of search results based on users' search and click history has been investigated, while other work employs reinforcement learning to model user intent through data visualization techniques.

System Architecture

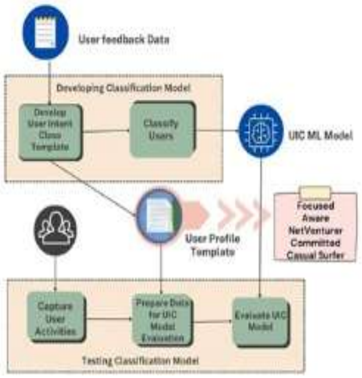

Fig:5.1 System Architecture

6. RESULT

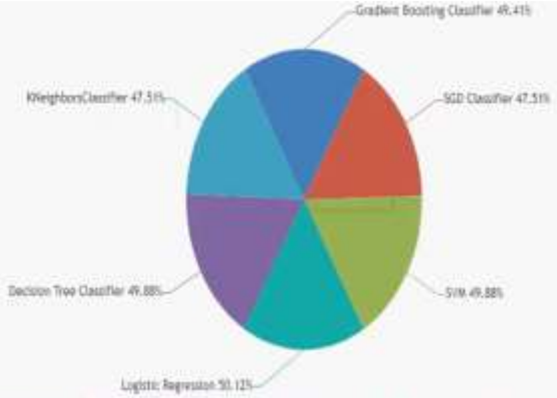

Fig: 6.1 Resultant graph

The **Service Provider** is responsible for all backend operations, including data training and behavior classification. It utilizes a training and testing dataset to develop machine learning models that classify user behavior. The system compares the performance of different algorithms, and the results—specifically the accuracy of each algorithm—are displayed using a pie chart for easy visualization. Additionally, the system can detect and display different types of user behaviours such as Searching, Sharing, and Verifying. The algorithms employed for this classification task include Gradient Boost Classifier, Support Vector Machine (SVM), Stochastic Gradient Descent (SGD) Classifier, Logistic Regression, Decision Tree Classifier, and K-Nearest Neighbors (KNN). These algorithms are used to evaluate which model best captures the patterns in user behaviour based on the collected data.

7. CONCLUSION

In conclusion, this research effectively bridges the gap in understanding online user behaviour by introducing a machine learning-based framework that captures the multifaceted aspects of information-seeking behaviour—namely searching, sharing, and verifying. Unlike previous models that focused narrowly on specific behaviour or required inaccessible data, this study leverages user feedback and interaction patterns to classify users into intent-based profiles using unsupervised and supervised learning methods. The system demonstrates promising accuracy and reliability in predicting user intent, which has valuable applications in targeted marketing, personalized content delivery, and search engine optimization. By addressing limitations in existing systems, such as data complexity and labelling issues, and incorporating dynamic user data for validation, the proposed model offers a comprehensive and practical solution for analysing and utilizing online behavioural patterns.

8. REFERENCES

[1] “A machine learning approach to user profiling for data annotation of online behavior,” CMC-Comput. Mater. Con., vol. 74, no. 1, pp. 123–135, Feb. 2024. M. Kanwal, N. A. Khan, and A. A. Khan.

[2] “Explainable AI for machine fault diagnosis: Understanding features’ contribution in machine learning models for industrial condition monitoring,” J. Ind. Inf. Integr., vol. 40, pp. 100–115, Feb. 2023. E. Brusa, L. Cibrario, C. Delprete, and L. G. Di Maggio.

[3] “Data clustering: Application and trends,” Artif. Intell. Rev., vol. 55, no. 9, pp. 7453–7479, Nov. 2022. G. J. Oyewole and G. A. Thopil.

[4] “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, Oct. 2022. X. Feng, X. Wang, and Y. Zhang.

[5] “Social media behaviour analysis in disaster-response messages of floods and heat waves via artificial intelligence,” AI Soc., vol. 38, no. 2, pp. 521–539, Jun. 2022. V. Ponce-López and C. Spataru.

[6] “A survey of human-in-the-loop for machine learning,” Future Gener. Comput. Syst., vol. 135, pp. 328–346, May 2022. X. Wu, L. Xiao, S. Yixuan, and L.

[7] “Fake news detection based on news content and social contexts: A transformer-based approach,” J. Inf. Technol. Polit., vol. 19, no. 1, pp. 42–59, Jan. 2022. S. Raza and C. Ding.

[8] “This is fake! Shared it by mistake: Assessing the intent of fake news spreaders,” in Proc. Int.

AAAI Conf. Web Social Media, Apr. 2022. X. Zhou, K. Shu, V. V. Phoha, and R. Zafarani.

[9] “A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects,” Eng. Appl. Artif. Intell., vol. 110, pp. 104743, Apr. 2022. A. E.-S. Ezugwu, A. M. Ikotun, O. Olaide, and A. Akinyelu.

[10] “Preserving integrity in online social networks,” Commun. ACM, vol. 65, no. 2, pp. 92–98, Feb. 2022. A. Halevy, C. Canton-Ferrer, H. Ma, and V. Stoyanov.

****