

User Profiling from Textual Data: A Systematic Literature Review

Jemal A. Jilo¹, Felipe Peixoto¹, Myriam Delgado¹, and Thiago H. Silva^{1,2}

¹Federal University of Technology - Parana, Curitiba, Brazil

²University of Toronto, Toronto, Canada

jemalabatejilo,felipefiaspeixoto@alunos.utfpr.edu.br

myriamdelg,thiagoh@utfpr.edu.br

ABSTRACT

This review examines current research on user profiling from textual data. The adopted methodology encompasses a systematic literature search, the selection of peer-reviewed studies based on predefined eligibility criteria, and a quality assessment of their research design, datasets, and machine learning methods. We conducted the search using multiple querying strategies across ACM DL, IEEE Xplore, Scopus, Web of Science, ScienceDirect, Springer-Link, and Google Scholar as a complementary source. The search retrieved 133 studies using different query formulations. After applying the predefined evaluation criteria, we selected 37 studies for detailed analysis, of which 3 are surveys. These studies investigate three main categories of user profiling dimensions: (1) demographic and sociodemographic factors, (2) personality and psychographic traits, and (3) behavioral patterns, preferences, and intent. The selected studies employ a wide range of methods, including classical machine learning, deep learning, BERT-based models, and large language models (LLMs). Regarding datasets, Twitter is the most frequently used platform and supports studies on gender, occupation, socioeconomic status, personality, and political profiling. Researchers also use data from Reddit, Amazon, and platform-specific corpora covering Russian, Arabic, Turkish, Persian, Korean, and Punjabi. However, English remains the predominant language in the datasets used by most studies. Evaluation practices are highly heterogeneous, with studies employing metrics such as accuracy, F1-score, macro-F1, AUC, correlation, error-based measures, and clustering quality indices. Despite recent advances, several challenges remain, including limited dataset diversity, demographic bias, feature sparsity, and computational constraints. In light of these challenges, and based on the review findings, we emphasize the importance of adopting multilingual and cross-cultural datasets, explainable AI frameworks, and standardized evaluation protocols to improve replicability and support the ethical deployment of user profiling systems. Overall, this review consolidates the current state of the art and identifies promising directions for developing more robust, transparent, and responsible approaches to user profiling from textual content.

KEYWORDS

User Profiling; User Attribute Inference; Textual Data; Natural Language Processing; Web-based Systems; Systematic Literature Review

1 INTRODUCTION

The growing use of digital platforms has led to the continuous production of large volumes of structured and unstructured data. Social media platforms, such as Twitter/X, Instagram, Facebook, and LinkedIn, are especially relevant in this context because they enable users to share opinions, preferences, activities, and social interactions through textual and multimedia content [9]. Among these data sources, textual content has become particularly valuable for computational approaches to user profiling, since language can reveal demographic attributes, interests, behavioral patterns, preferences, and psychological characteristics.

User profiling refers to the process of constructing representations of users based on their characteristics, behaviors, activities, and preferences. Traditionally, user profiling relies heavily on explicit information provided directly by users, such as questionnaires, registration forms, ratings, or self-declared preferences. This approach, often referred to as explicit, static, or factual profiling, has the advantage of relying on information intentionally provided by the user. However, it also presents important limitations. Users may be unwilling to disclose personal information due to privacy concerns, may provide incomplete or inaccurate answers, or may avoid long and repetitive forms. Moreover, static profiles can quickly become outdated as users' interests, behaviors, and contexts change over time [41]. These limitations have motivated the development of automated and dynamic profiling methods capable of inferring user attributes from behavioral traces and textual data.

Recent advances in Natural Language Processing (NLP), particularly contextual language models and transformer-based architectures, have improved the ability to capture semantic, syntactic, and contextual cues in user-generated content [14]. These advances have made automated user profiling an increasingly active research area at the intersection of NLP, machine learning, social media analysis, and human-centered computing. As social platforms generate content at scale, these methods support more dynamic and context-aware user representations by capturing not only explicit signals, such as stated preferences, but also implicit signals, including linguistic style, sentiment, interaction patterns, and topical interests.

User profiling from textual data is directly relevant to understanding users' characteristics, preferences, and behaviors, enabling more personalized and context-aware applications and services, such as recommender systems, personalized content delivery, targeted communication, and content moderation. Social media posts, comments, reviews, captions, search queries, forum discussions, and profile descriptions often contain signals about users' interests, preferences, opinions, intentions, and social characteristics. These signals

provide the basis for such personalized and adaptive Web applications. However, inferring user attributes from text also introduces important challenges for Web systems, especially when models rely on noisy, biased, platform-specific, or sensitive data. Issues such as privacy, fairness, transparency, reproducibility, and responsible data use therefore become central to the design and evaluation of user profiling methods. By systematically reviewing how textual data has been used for user profiling, this study helps organize current evidence and identify directions for more reliable, transparent, and responsible Web-based applications.

Despite the growing body of research on user profiling from textual data, the field remains fragmented. Existing studies vary widely in terms of profiling targets, data sources, languages, platforms, features, algorithms, and evaluation protocols. Some studies focus on demographic attributes, such as age, gender, occupation, income, or ethnicity; others investigate personality traits, interests, intentions, or behavioral patterns. In addition, important methodological and ethical challenges remain, including dataset bias, limited demographic and cultural diversity, lack of multilingual resources, privacy risks, explainability, reproducibility, and inconsistent evaluation practices.

This study aims to provide a systematic literature review of research on user profiling from textual data. Specifically, we analyze how previous studies define user profiling tasks and evaluate them, which textual data sources and computational methods they employ, which user attributes they attempt to infer, and what limitations and open challenges are reported in the literature. By organizing existing knowledge on this topic, this review seeks to support future research on more robust, transparent, and responsible approaches to user profiling from text.

This review builds on previous surveys on user profiling while addressing a different and broader scope. Wu et al. [50] focus mainly on user profiling in social networks, whereas Naz et al. [38] concentrate on personality trait detection. Eke et al. [16] provide a broad overview of user profiling methods and processes, but their review is not specifically centered on textual data as the primary profiling signal and does not cover recent transformer-based and LLM-based approaches. In contrast, the present review focuses on user profiling from textual data in Web-based environments and covers a wider range of profiling targets, including age, gender, occupation, income, race, ethnicity, personality traits, preferences, behavioral patterns, and user intent. We also examine recent methodological developments, including transformer-based models and LLM-based user profiling, as well as datasets, platforms, languages, evaluation practices, ethical concerns, and open research gaps. Table 1 summarizes the main differences between this review and previous surveys. The dataset list, the spreadsheet containing the scoring tables used in the study selection process, and links to identified source-code repositories are publicly available in a GitHub repository¹.

The main contributions of this paper are: (i) a systematic mapping of user profiling studies based on textual data, covering demographic, psychographic, and behavioral profiling; (ii) a comparative analysis of the textual sources, platforms, languages, datasets, and data availability addressed in the literature; (iii) a synthesis of the

computational methods adopted in the field, ranging from traditional feature-based approaches to transformer-based and large language model (LLM)-based methods; and (iv) the identification of methodological and ethical gaps particularly relevant to Web-based systems, including dataset bias, limited multilingual coverage, weak reproducibility, inconsistent evaluation protocols, and the insufficient treatment of sensitive attributes.

The rest of this systematic literature review is organized as follows. Section 2 presents the background and related work. Section 3 describes the review methodology used in the study. Section 4 presents literature analysis and findings. Section 5 discusses the synthesis of findings. Finally, Section 6 concludes the paper and provides final remarks.

2 BACKGROUND

This section provides the background necessary to understand user profiling from textual data. It introduces the fundamental concepts of the field, describes the main profiling categories and user attributes, reviews the textual sources and computational approaches commonly employed in the literature, and highlights the ethical and methodological challenges associated with the development and deployment of user profiling systems.

2.1 User Profiling from Textual Data

User profiling refers to the computational process of constructing representations of users based on their characteristics, preferences, behaviors, and interactions. In text-based environments, these representations are commonly inferred from user-generated content such as posts, comments, reviews, messages, search queries, profile descriptions, ratings, and captions. Textual data can reveal both relatively stable attributes, such as age, gender, occupation, income, or geographic location, and more dynamic aspects, such as interests, intentions, opinions, risk perception, and behavioral patterns.

In this review, we focus on studies that infer user attributes primarily from textual data. This scope includes demographic profiling, personality and psychographic profiling, behavioral profiling, and preference or interest modeling, as long as textual information plays a central role in the profiling process.

2.2 User Profiling Categories

Previous studies on user profiling have addressed a wide range of attributes. Demographic profiling typically focuses on attributes such as age, gender, location, education, occupation, income, race, ethnicity, or political affiliation [2, 11, 26, 27, 46, 52]. Personality and psychographic profiling attempt to infer psychological traits, preferences, emotional tendencies, values, or decision-making patterns, often using models such as the Big Five or Myers-Briggs Type Indicator (MBTI) [6, 8, 18, 25, 35]. Behavioral profiling focuses on users' actions, habits, interests, intentions, and engagement patterns.

These categories are not mutually exclusive. For instance, a recommender system may model user interests and preferences while also incorporating demographic or behavioral signals [31, 41]. Similarly, personality prediction studies may be treated as part of user profiling when they contribute to a broader representation of the

¹<https://github.com/felip-peixoto/Textual-User-Profiling-Survey>

Table 1: Comparison between this review and previous surveys on user profiling.

Study	Main scope	Profiling attributes covered	Data sources and platforms	Main distinction of our work
Wu et al. [50]	User profiling in social networks	Mainly user attributes inferred from social network data	Primarily social networking platforms and relational data	Provides a broader view beyond social networks, including reviews, forums, search queries, conversational data, and other textual Web sources.
Naz et al. [38]	Personality trait detection	Primarily personality and psychographic traits	Mostly textual datasets used for personality prediction	Extends the scope beyond personality prediction to demographic, sociodemographic, behavioral, preference, and intent attributes.
Eke et al. [16]	General user profiling methods and processes	Broad but not specific to any profiling category	General, not focused on any particular platform or data type	Focuses specifically on textual data as the primary profiling signal and covers recent transformer-based and LLM-based approaches not addressed by Eke et al.
This review	User profiling from textual data in Web-based environments	Demographic, sociodemographic, personality, psychographic, behavioral, preference, and intent attributes	Social media posts, comments, reviews, captions, search queries, forum discussions, profile descriptions, and other user-generated textual data	Systematically analyzes profiling targets, datasets, platforms, languages, computational methods, evaluation practices, ethical concerns, and open research gaps, including recent transformer-based and LLM-based approaches.

user [24]. Therefore, this review uses these categories as analytical dimensions rather than rigid boundaries.

The computational formulation of user profiling depends on the type of attribute being inferred. Consequently, profiling from textual data may be modeled as a classification task (e.g., gender or political orientation prediction) [27, 43], a regression task (e.g., age estimation) [2], a clustering task (e.g., discovering user groups) [31], a recommendation problem (e.g., personalized content suggestion) [15], a representation learning task (e.g., learning user embeddings), or a profile generation task (e.g., automatically constructing textual user descriptions).

User profiling can also be understood through the distinction between content-based, collaborative, and hybrid methods. Content-based approaches infer users' current or future preferences from their own past content or behavior. Collaborative methods infer preferences by comparing users with similar users or items. Hybrid methods combine these perspectives to reduce the limitations of each individual approach [13]. In recommender-oriented profiling, related techniques include neighborhood-based methods, content-based filtering, collaborative filtering, matrix factorization, ontology-based profiling, and statistical modeling.

2.3 Textual Sources and Computational Approaches

User profiling from text has been studied using different data sources, including social media posts, tweets, Reddit comments, online reviews, profile descriptions, captions, conversational data, essays, and web communication corpora [7, 13, 24, 31, 41, 45]. These sources vary in length, formality, language, availability of labels, and ethical sensitivity. As a result, studies adopt different preprocessing strategies, feature representations, and modeling approaches.

To infer user attributes from textual data, traditional approaches typically rely on handcrafted features, Bag-of-Words representations, TF-IDF, psycholinguistic dictionaries, topic models, and word embeddings to encode textual information [2, 8, 22, 25]. More recent studies employ contextualized language models, transformer-based architectures, multimodal models, and large language models (LLMs) to learn richer semantic and contextual representations for user profiling tasks [7, 14, 30, 41, 43, 45]. While transformer-based models are commonly used to extract contextual representations for downstream prediction, recent LLM-based approaches support more flexible profile construction and interpretation. In particular, LLMs have been explored for extracting and generating natural-language profiles, modeling temporal changes in user preferences, and constructing profiles from conversational histories, reviews, and social media content. These approaches can provide richer and more interpretable user representations, but also raise challenges related to computational cost, model variability, reproducibility, privacy, and dependence on proprietary models.

Neighborhood-based approaches identify similar users or items by applying similarity measures, such as cosine similarity or the Jaccard index, to vectorized representations derived from textual data. These representations may include TF-IDF vectors, word embeddings, or features extracted from user reviews. Recent studies have employed KNN-based and collaborative recommendation models to leverage user reviews and item preferences for personalized recommendations [5, 23]. Ontology-based profiling adds semantic structure to user representations by mapping textual or behavioral signals to domain concepts, improving context-awareness and interpretability [10, 18, 51]. Filtering and recommendation-oriented profiling also include content-based methods for matching users and items [15], deep learning and self-supervised approaches for sparsity and cold-start scenarios [32], and hybrid models that combine sentiment classification with user profiling [48]. Statistical

and matrix-factorization approaches remain relevant for modeling user preferences and rating distributions, particularly when interpretability or data sparsity is a concern [29].

2.4 Ethical and Methodological Challenges

User profiling from textual data raises important methodological and ethical challenges. Methodologically, studies often face limitations related to dataset bias, class imbalance, small or platform-specific samples, lack of multilingual and cross-cultural data, inconsistent evaluation metrics, and limited reproducibility. Ethically, profiling may involve sensitive attributes such as race, ethnicity, income, education, political orientation, health-related behavior, or psychological traits. These attributes require careful consideration of privacy, consent, fairness, transparency, and potential misuse [11, 26].

Because of these challenges, this review analyzes not only the predictive methods and reported performance of user profiling studies, but also the datasets, evaluation protocols, limitations, and ethical issues discussed in the literature.

3 REVIEW METHODOLOGY

This section describes the methodology adopted to conduct the systematic literature review, ensuring a transparent, rigorous, and reproducible study selection and analysis process. It first presents the research questions that guided the review, followed by the search strategy used to identify relevant studies. Next, it describes the exclusion criteria and the study selection process employed to determine the final set of publications considered in this survey. The section then details the quality assessment procedure used to evaluate the methodological soundness of the selected studies and concludes by explaining the data extraction and synthesis process adopted to organize, compare, and analyze the evidence reported in the literature.

3.1 Research Questions

The following research questions define the scope of this systematic literature review and provide the framework for identifying, analyzing, and synthesizing the selected studies.

- **RQ1:** Which user attributes have been inferred from textual data in previous user profiling studies?
- **RQ2:** Which textual data sources, platforms, and datasets have been used for user profiling?
- **RQ3:** Which NLP, machine learning, and deep learning methods have been applied to user profiling from textual data?
- **RQ4:** How do existing studies evaluate user profiling models, and which metrics are commonly reported?
- **RQ5:** What are the main limitations, challenges, and future research directions identified in the literature?

3.2 Search Strategy

The search strategy was designed to retrieve studies on user profiling, author profiling, user modeling, user representation, and user attribute inference from textual data. We used the following search query:

("user profiling" OR "author profiling" OR "user modeling" OR "user representation" OR "demographic prediction" OR "trait inference") AND ("text" OR "textual data" OR "social media text" OR "user-generated content" OR "natural language processing" OR "NLP")

Using this query, we searched ACM Digital Library, IEEE Xplore, Scopus, Web of Science, ScienceDirect, and SpringerLink, with Google Scholar used as a complementary source to identify additional relevant studies.

The searches were executed between January 1, 2026, and June 1, 2026. The review considered studies published from 2012 onward, according to the inclusion criterion defined in Table 2. For each database, the search was applied to the available metadata fields, including title, abstract, and keywords; full-text search was used only when supported by the database interface. The review process was inspired by the systematic review framework of [4].

3.3 Exclusion Criteria

The review considered peer-reviewed journal, conference, and workshop papers on user profiling in which textual data played a central role. The selected articles were evaluated according to their relevance to user profiling techniques, data sources, methods, and reported results.

Venue quality indicators, such as impact factor (IF) and CAPES QUALIS classification in the field of Computer Science², were considered in the preliminary quality assessment. Table 2 shows the study selection process guided by predefined exclusion criteria.

Table 2: Exclusion criteria.

Category	Exclusion
Study type	Any study not published in journals or conferences/workshops (e.g. theses, preprints, etc...).
Topic	Studies with no user attribute inference or non-profiling NLP tasks.
Data modality	Studies relying exclusively on non-textual signals.
Language	None; studies in any language were eligible.
Period	Published before 2012.
Access	Full text inaccessible.
Quality	IF ≤ 1 AND QUALIS C or unclassified; QA score < 2.5 ; duplicated records.

The Impact Factor (IF) is a bibliometric indicator that reflects the scientific influence and citation impact of a journal [47]. QUALIS is a classification system developed by the Brazilian Coordination for the Improvement of Higher Education Personnel (CAPES) to support the evaluation of graduate programs. It classifies academic journals and conference proceedings into tiers as A1-A4, B1-B4, and C according to their perceived quality and relevance within specific research areas, providing a reference for assessing the scientific output of graduate students and faculty members. A1 represents the highest-quality and most influential journals and conferences, followed by A2, A3, and A4 as progressively lower but still highly

²For conference/workshop and journals: <https://sucupira-legado.capes.gov.br/sucupira>.

regarded tiers; B1–B4 indicate journals and conferences with moderate academic impact and visibility, while C denotes journals that receive no quality weighting in research evaluation [28].

3.4 Study Selection Process

Under the PRISMA 2020 framework [39], the study-selection flow diagram is conventionally organized into four main phases: identification, screening, eligibility, and included. Figure 1 summarizes the customized process adopted for this review. The quality assessment was performed independently by three authors using the QA criteria described in Section 3.4.3. Disagreements in scores were resolved through discussion and consensus before defining the final set of included studies.

3.4.1 Identification. We conducted an automated search to identify relevant studies across ACM Digital Library, IEEE Xplore, Scopus, Web of Science, ScienceDirect, and SpringerLink, while Google Scholar was used as a complementary source to ensure broader coverage of the literature. By applying the search query described in Section 3.2, we retrieved a total of 133 studies published in a variety of journals, conferences, and other academic venues.

3.4.2 Screening. This phase (2) encompasses four tasks: (2.1) title and abstract screening, (2.2) duplicate removal, (2.3) study eligibility, and (2.4) access verification.

Title and abstract screening: After retrieving the search results, we analyzed the titles and abstracts of the identified studies to ensure their relevance to the objectives of this review. Based on this initial screening, 65 studies were excluded because they did not address the scope of our research, while the remaining 68 studies were retained for further evaluation in the subsequent stages of the review process.

Duplicate removal: In this task, we conducted a thorough review to identify and eliminate duplicate articles from the collection. All references were managed in BibTeX within the Overleaf environment and formatted according to the WebMedia 2026 SBC LaTeX template. Two duplicate records were then identified and removed using a structured spreadsheet, resulting in a set of 66 unique studies.

Preliminary Quality Assessment: In this task, we assessed the quality of the selected papers based on the IF of their publication venues and their QUALIS classifications. We retained only studies published in venues with an IF greater than one or a QUALIS classification of A1–A4 or B1–B4. Based on these criteria, 20 studies (19 using IF and 1 using QUALIS) were excluded because they did not satisfy the predefined quality requirements, leaving 46 studies for the subsequent stages of the review.

Access verification: Then, we verified the accessibility of the remaining studies by checking whether their full texts were available for analysis. This step ensured that only studies with complete access could be included in the review. Among the 46 remaining studies, one was excluded because its full text could not be obtained.

After screening, the selected articles were evaluated for methodological and scientific quality in the next phase.

3.4.3 Eligibility. At this phase, we examined the full texts of the remaining 45 studies to verify whether they contained the information required to answer the research questions. The quality assessment considered each paper’s methodology, research design, data collection procedures, experimental setup, and contribution to user profiling. Then, after screening, all studies were evaluated using five quality assessment (QA) questions: QAa) Is the profiling task clearly defined? QAb) Is the dataset adequately described? QAc) Are the methods and experimental procedures sufficiently explained? QAd) Are evaluation metrics and results clearly reported? QAE) Are limitations or challenges discussed? Each question was scored as 1 (fully satisfied), 0.5 (partially satisfied), or 0 (not satisfied), resulting in a maximum score of 5 points. Studies scoring 2.5 or higher were included in the final review to ensure adequate methodological quality and reporting transparency. The quality assessment of the 45 eligible studies excluded eight that did not provide sufficient information to address the objectives of this review, resulting in a set of 37 studies.

3.4.4 Included. In the final phase, we systematically extracted relevant information from each selected study. The final corpus comprised 37 studies, including 34 primary studies and 3 survey papers. The survey papers, summarized in Table 1, were retained to contextualize the present review and position its contributions relative to the existing survey literature. For each of the 34 primary studies, we extracted information on the study objective, profiling target, dataset and data source, language, platform, computational methods and models, experimental design, evaluation metrics, best reported results, and reported limitations. In addition, we coded whether each study explicitly discussed ethical issues related to user profiling, including privacy and consent, bias and fairness, sensitive attributes, and reproducibility. These categories were defined based on recurring concerns reported in the reviewed literature and were used to synthesize the ethical gaps presented in Section 5. We also identified the tools, frameworks, and machine learning algorithms used in training and evaluating the user profiling models. Table 3 provides a summary of all the data extraction fields we used to address the research questions.

Table 3: Data extraction fields aligned with the research questions.

Extraction Field	Description	Maps to
Profiling target	User attribute(s) inferred	RQ1
Dataset & platform	Dataset name and source platform (Twitter, Reddit, Amazon)	RQ2
Language & size	Language(s) of the data and number of users/documents	RQ2
Methods & models	ML/DL models used (TF-IDF + SVM, BERT, GPT-4o)	RQ3
Metric & best result	Primary evaluation metric and best reported value	RQ4
Limitations	Dataset bias, class imbalance, generalizability, ethical, or reproducibility issues	RQ5
Ethical issues	Privacy and consent, bias and fairness, sensitive attributes, and reproducibility concerns	RQ5

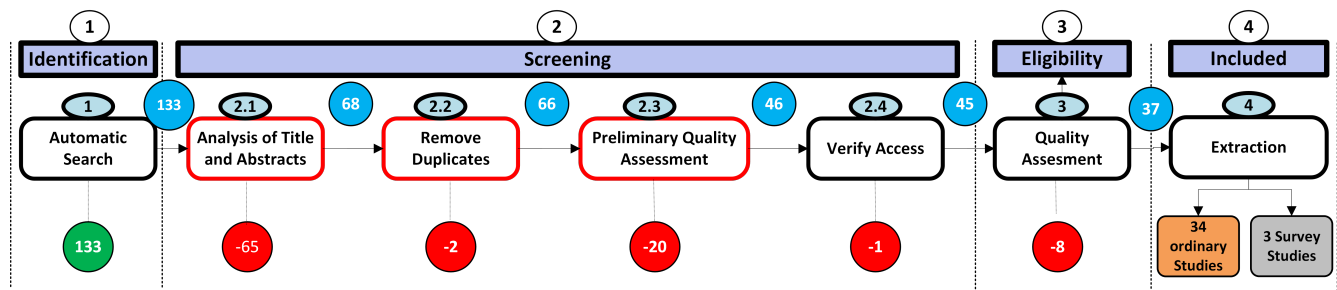


Figure 1: Flow diagram illustrating our study selection process, adapted from [4] under PRISMA 2020 perspective.

4 LITERATURE ANALYSIS AND FINDINGS

This section presents the analysis and synthesis of the studies selected through the systematic review process. It begins with an overview of the final filtered subset. Subsequently, it analyzes the user attributes inferred from textual data, the textual sources, platforms and datasets employed, the computational methods and models adopted for user profiling. The section also examines the evaluation metrics and reported performance used to assess profiling models and concludes by discussing the current limitations and challenges identified in the literature. Together, these analyses provide a comprehensive overview of existing research, reveal prevailing trends and methodological practices, and establish the foundation for the discussion of research gaps and future directions presented in the next section.

4.1 Overview of Selected Studies

The final corpus comprises 37 studies, published from 2012 up to the search cutoff date, June 1, 2026, of which 3 are survey studies related to user profiling. Most of the work is relatively recent - the majority of studies appeared after 2018, with a particularly dense cluster in 2024 and 2025.

Figure 2 presents a grouped bar chart illustrating the distribution of selected studies across three profiling dimensions or categories (1) demographic and sociodemographic factors, (2) personality and psychographic traits, and (3) behavioral patterns, preference, and intent, organized into three publication periods. Because some studies address more than one profiling target, the category counts in Figure 2 are not mutually exclusive.

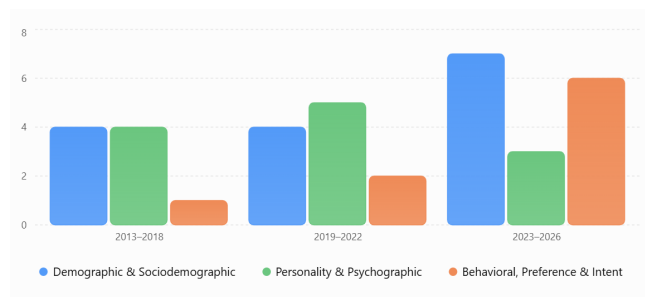


Figure 2: Distribution of Studies across profiling categories not mutually exclusive and time periods.

The selected studies cover a diverse set of user profiling tasks, ranging from demographic and socioeconomic attribute inference (e.g., age, gender, occupation, income, race, and ethnicity) to personality profiling, user intent clustering, recommendation-oriented profiling, and LLM-based profile generation.

The final corpus includes studies using traditional machine learning, deep learning, word embeddings, graph embeddings, ontology-based methods, multimodal fusion, and LLM-based methods. The earlier studies, from 2013 to 2018, tend to rely on traditional machine learning and early deep learning, and they provide much of the conceptual and methodological foundation that later work builds on. In terms of where the studies were published, the corpus spans journal articles, conference papers (including those published in workshops of high-quality conferences).

Most studies frame their task as classification or prediction, inferring a specific user attribute from text, though a few take different approaches, such as clustering, behavioral inference, or LLM-based profile generation.

Finally, English dominates as the language of the underlying data, accounting for 20 studies, while the remaining studies focus on Russian, Arabic, Turkish, Punjabi, Persian, Korean, Spanish, Indonesian, or multilingual corpora, indicating that non-English user profiling research is growing but remains scattered rather than forming a coherent cross-lingual research effort.

The findings are summarized in three tables according to the user profiling categories identified in this review. Table 4 presents demographic and sociodemographic profiling, Table 5 summarizes personality and psychographic profiling, and Table 6 covers behavioral, preference, and intent profiling.

4.2 User Attributes Inferred from Textual Data

RQ1 investigates which user attributes have been inferred from textual data in previous user profiling studies. The reviewed literature shows that user profiling from text covers a broad set of attributes, ranging from relatively stable sociodemographic characteristics to more dynamic behavioral and preference-related dimensions. In general, the identified studies can be grouped into three main categories: demographic and sociodemographic attributes, personality and psychographic attributes, and behavioral, preference, and intent attributes.

These categories are not mutually exclusive. Some studies focus on a single attribute, such as gender, age, occupation, or personality, while others combine multiple signals to construct broader user

Table 4: Summary of related studies on demographic and sociodemographic profiling

Author	Goal	Dataset	Techniques	Gap
[2]	Word embeddings for user profiling (age, gender) in social networks	Odnoklassniki (868K users): text posts, likes, demographics Russian	Word2Vec, clustering, logistic regression	Limited to a single Russian social network; low cross-platform generalizability
[27]	Gender identification on Twitter across multiple languages	Seven CLEF-PAN corpora (15K users) English, Spanish, Portuguese, Arabic	Logistic regression, decision tree, KNN, SVM, Naive Bayes, neural networks, random forest	Binary gender assumption; English-dominant data; Twitter-specific scope (2014–2019)
[46]	Gender identification for authorship analysis	1,944 samples — Punjabi (Twitter)	LightGBM; AGI-P fine-tuned multilingual model	Small dataset; narrowly focused on gender; limited language coverage
[26]	Author gender identification in Modern Standard Arabic	8,000 news articles labeled by author gender	CNN, LSTM, BiLSTM, BERT	Domain-specific (news); limited to Modern Standard Arabic; does not address dialectal variation
[11]	Predict author age and gender in resource-constrained digital forensics settings	NIKL Korean Dialogue Corpus (27K)	LoRA/QLoRA fine-tuning, EEVE, Bllossom (LLM-based)	Only NIKL corpus used; gender classification signals may introduce bias; single-language scope
[3]	Predict Twitter users' occupational class and income from network and text	Twitter (5.1K users) English	Graph embeddings, TF-IDF, logistic regression; network + language fusion	Small dataset; income labels are noisy proxies; platform-specific to Twitter
[52]	Occupation prediction from Twitter data	25K Turkish tweets from 10 occupations	Logistic regression, Naive Bayes, Random Forest, SVM, MLP, CNN, LSTM, BiLSTM, GRU, BERT	Few users per occupation; Turkish-only scope; limited generalization to other languages/platforms
[40]	User-level race and ethnicity prediction from Twitter text	Twitter survey (5.4M tweets) English	Logistic regression with n-gram and topic features	Class imbalance; proxy labels; limited cross-cultural generalizability; ethical risks
[49]	Predict Twitter user demographics using noisy self-reports	Twitter survey (2.67M users) English	BERT-based classifiers; noisy self-reported labels	Noisy labels; selection bias; Twitter-specific; underrepresentation of minority groups
[12]	Uncover sociodemographic attributes from Reddit participation signals	Reddit (850K users) English	statistical modeling	Platform-specific; limited to Reddit communities; potential privacy concerns
[19]	Relational user attribute inference (age, gender, location) in social media	Google+ (846K users) English	Probabilistic graphical models; relational attribute inference	Platform-specific (Google+, now defunct); limited to social graph signals
[20]	Multilingual author profiling (age, gender) on Facebook	Facebook (7.7K users) Roman Urdu and English	SVM, Naive Bayes, TF-IDF, n-grams, stylistic features	Small dataset; limited to two languages; platform-specific
[33]	Author profiling using distance measures (age, gender)	CLEF evaluation campaigns (English, Spanish, Dutch, Italian)	Distance-based classifiers; n-gram and character features	Relies on CLEF benchmarks; stylistometric approach may not generalize to informal text
[1]	Gender Inference from Place Reviews Applications to Urban Representativeness	Google Places review (33+M reviews) English	BERT	English only dataset
[36]	Author profiling for abuse detection	16K tweets, English	NLP-based author profiling, ML classifiers	Domain-specific (abuse); Twitter-only; limited demographic scope

representations. Moreover, several works complement textual data with additional sources, such as social network structure, profile metadata, interaction logs, ratings, or visual content. This indicates that user profiling is increasingly moving from isolated attribute prediction toward richer, more contextual user modeling.

4.2.1 Demographic and Sociodemographic Attributes. Demographic and sociodemographic attributes are among the most frequently investigated targets in user profiling from textual data. These attributes include age, gender, geographic location, education level, occupation, income, race, ethnicity, and political or partisan affiliation. In many studies, these characteristics are inferred from linguistic patterns, such as vocabulary choices, writing style, text complexity, topics of interest, and semantic representations of user-generated content.

Among the studies summarized in Table 4, age and gender are the most frequent demographic profiling targets, appearing either individually or jointly across several works. Other studies address occupation, income, race, ethnicity, location, and broader sociodemographic attributes. In contrast, education and unemployment do not appear as explicit profiling targets in the final selected corpus, representing a clear gap in the reviewed literature, particularly given their relevance to socioeconomic user modeling and the growing availability of platform data that could support such inference.

These age- and gender-related studies model user attributes using social media posts, profile descriptions, comments, or other forms of online communication. Alekseev and Nikolenko [2] used Word2Vec representations for user profiling in Odnoklassniki, a Russian social network, while Ikae and Savoy [27], Sarwar et al. [46], and Himdi and Shaalan [26] investigated gender identification using stylistic markers, machine learning classifiers, and deep learning models. Additional resources, such as PANDORA, combine personality and demographic labels in Reddit-based data [24]. Traditional approaches often rely on stylistic markers, n-grams, Bag-of-Words, TF-IDF, and supervised machine learning classifiers, whereas more recent studies employ word embeddings, contextualized language models, and transformer-based architectures.

Beyond age and gender, some studies address socioeconomic attributes such as occupation and income. Aletas and Chamberlain [3] predict Twitter users' occupational class and income by combining graph embeddings and textual features. Occupation prediction has also been studied in Turkish tweets using machine learning, deep learning, and BERT-based models [52]. These studies suggest that language and network structure can provide complementary signals for socioeconomic profiling. However, socioeconomic attributes remain less explored than age and gender and often depend on smaller or more context-specific datasets.

Race and ethnicity prediction also appear in the reviewed literature. Preotiuc-Pietro and Ungar [40] introduce models for user-level race and ethnicity prediction from Twitter text, while Wood-Doughty et al. [49] use noisy self-reports to predict Twitter user demographics. Cinus et al. [12] infer sociodemographic attributes from Reddit participation signals. Although some works report promising predictive performance, they also highlight important limitations, including class imbalance, selection bias, underrepresentation of minority groups, and limited generalizability across countries, languages, and cultural contexts. Because race and ethnicity are sensitive attributes, these studies should be interpreted not only as technical contributions, but also in relation to fairness, privacy, consent, and potential misuse.

Overall, demographic and sociodemographic profiling studies show that textual data can encode signals associated with users' social characteristics. However, the evidence is uneven across attributes. Age and gender are relatively well represented, while education, unemployment, income, race, ethnicity, and political affiliation are less consistently studied. In addition, many datasets are platform-specific, language-specific, or geographically restricted, which limits the generalization of models to broader and more diverse populations. Table 4 summarizes existing studies on Demographic and Sociodemographic Profiling, including their goals, datasets, techniques, and limitations.

4.2.2 Personality and Psychographic Attributes. Personality is a stable set of individual characteristics that explains behavioral differences between people [21]. Personality and psychographic attributes form a second major group of profiling targets. Associated studies aim to infer psychological traits, emotional tendencies, values, interests, beliefs, or decision-making patterns from textual data. The most common approach is personality prediction, especially using the Big Five model or MBTI. In these cases, user-generated text is treated as evidence of stable or semi-stable psychological tendencies.

Several studies use social media posts, tweets, Reddit comments, essays, or profile texts to predict personality traits. Traditional models usually transform text into numerical representations using Bag-of-Words, TF-IDF, n-grams, LIWC features, part-of-speech patterns, or word embeddings. Examples include MBTI prediction from Reddit [25], MBTI/persona trait identification using myPersonality and MBTI corpora [8], personality prediction from tweets [6], and Big Five prediction from tweet content [22]. Other studies apply deep learning or fusion strategies for personality prediction [17], including models focused on specific traits such as Extraversion [37].

The Big Five model is commonly used because it provides a well-established psychological framework, usually involving Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. MBTI-based studies are also frequent in computational personality prediction, particularly in social media datasets, although MBTI and Big Five should not be treated as equivalent models. Some studies focus on all dimensions of personality, while others target specific traits.

Psychographic profiling also extends beyond personality traits. Some works infer preferences, values, risk perception, political orientation, or decision-making tendencies. Kosinski et al. [34] show that private traits and attributes can be predicted from digital

behavioral records, while Ferwerda and Tkalcic [21] investigate personality prediction from Instagram images and captions, showing the potential of multimodal profiling. In recommender systems, user interests and preferences may be represented through natural-language profiles or embeddings derived from rating histories and textual content. Recent LLM-based approaches suggest that natural-language user profiles can improve interpretability, but they also raise concerns related to reproducibility, cost, hallucination, privacy, and dependence on proprietary models.

Overall, personality and psychographic profiling studies demonstrate that textual data can provide useful signals about how users express themselves, make decisions, and reveal preferences. Nevertheless, this category presents important limitations. Many studies rely on small or imbalanced datasets, platform-specific samples, or self-reported personality labels. In addition, psychological attributes are sensitive and may be used in persuasive, manipulative, or discriminatory systems if not handled responsibly. Table 5 presents the summary of related works on personality and psychographic profiling.

4.2.3 Behavioral, Preference, and Intent Attributes. A third group of studies focuses on behavioral, preference-related, and intent-oriented attributes. Unlike demographic attributes, which are often treated as stable, behavioral attributes are dynamic and context-dependent. They may include user interests, content preferences, purchasing intentions, information-seeking behavior, engagement patterns, verification practices, risk perception, or future actions. In textual environments, these attributes are usually inferred from posts, comments, reviews, ratings, queries, or other forms of user-generated content.

Some studies model behavior directly from text, while others combine textual information with interaction logs, social network structure, item ratings, or browsing behavior. Kanwal et al. [31] clustered users according to online behaviors related to searching, sharing, and verifying information. The clustering produces User Intent Classes that characterize distinct behavioral patterns, which are then generalized into a user-intent template for characterizing and labeling users. Although the resulting profiles may support downstream recommendation or personalization, in our taxonomy, this work is classified as behavioral and intent profiling.

Rahimzadeh et al. [43] used LLM-based extractive and abstractive methods to generate user profiles from political tweets. Sabouri et al. [45] explored explainable temporal user profiling with LLMs using Amazon reviews, and some emerging benchmarks study profile reconstruction from conversational histories [11].

Preference and interest profiling is particularly relevant for recommender systems and personalized services. In these studies, textual data can be used to represent user interests, item descriptions, review content, or natural-language explanations of user preferences. Some approaches rely on content-based filtering, collaborative filtering, hybrid recommender models, or matrix factorization, while more recent work explores embeddings and LLM-generated user profiles.

However, behavioral and intent profiling present specific challenges. User behavior changes over time, varies across platforms, and may depend strongly on context. Therefore, models trained on static datasets may not capture temporal changes in user interests

Table 5: Summary of related studies on personality and psychographic profiling.

Author	Goal	Dataset	Techniques	Gap
[25]	Personality prediction (MBTI) from Reddit posts	Reddit MBTI dataset (8,600 rows) English	Logistic regression, MLP, n-grams, POS tagging	Small dataset; limited to Reddit; self-reported MBTI labels; potential label noise
[8]	MBTI personality trait identification from social media text	myPersonality and MBTI corpus (Twitter) English	Naive Bayes, SVM, TF-IDF, LIWC features	Imbalanced and Twitter-only dataset; MBTI self-report bias
[6]	Personality prediction of users based on tweets (Big Five)	myPersonality dataset (260 users, 9,920 tweets) English	Logistic regression	Very small dataset; results may not generalize; reliance on deprecated myPersonality data
[22]	Compare ML algorithms for content-based personality prediction from tweets	Two datasets (8,675 and 8,789 users, 50 tweets per user) English	Supervised ML with BoW, TF-IDF, GloVe (SVM, Naive Bayes, Random Forest)	Self-reported MBTI mapped to only four Big Five traits; Twitter-only dataset
[35]	Big Five personality prediction on Twitter with LSTM	240K tweets (269 users) Indonesian	LSTM	Only 269 users; Indonesian-only; class imbalance (Extraversion label); limited cross-lingual generalizability
[37]	Prediction of Extraversion personality trait (Big Five)	MBTI dataset English	Bi-LSTM	Focused on a single trait (Extraversion); limited scope for full personality modeling
[17]	Deep learning fusion strategies for Big Five personality prediction	Student Essays and myPersonality English	ELMo, ULMFiT, BERT fusion	Performance below state-of-the-art; focus only on Big Five; dataset size limitations
[24]	Personality and demographic profiling on Reddit	PANDORA Reddit dataset (10K users) English	Multiple classifiers; combined Big Five and demographic labels	Self-reported labels; Reddit-specific;
[34]	Predict private traits and attributes from digital behavioral records	Facebook (58K users) English	Logistic regression, SVM; Like-based feature vectors	Privacy and consent concerns; Facebook-specific behavioral signals; ethical implications of sensitive inference
[44]	Impact of emotions on author profiling (age, gender, personality)	Blogs and social media (84K posts) Spanish	SVM, emotion features, stylometric features	Spanish-only; emotion features may not transfer across languages; limited reproducibility
[45]	Explainable temporal user profiling with LLMs from Amazon reviews	Amazon Movies/TV and Video Games reviews (18K) English	GPT-4o-mini; extractive and abstractive LLM-based profiling	Below benchmark performance; English-only; cost and reproducibility concerns with proprietary LLMs
[43]	LLM-based user profiling from political tweets (Persian presidential election)	6M tweets from 3,000 users Persian	GPT-4o	Political-domain focus; high cost; reproducibility concerns; interpretability issue

or intentions. In addition, behavioral traces can reveal sensitive information about routines, vulnerabilities, beliefs, and personal circumstances, which makes privacy and responsible data use especially important. Table 6 summarizes existing studies on Behavioral, Preference, and Intent Profiling, including their goals, datasets, techniques, and limitations.

4.3 Data Sources, Platforms, and Datasets

The reviewed studies use a wide variety of textual sources and platforms. Twitter/X is one of the most common sources, supporting studies on socioeconomic attributes [3], gender [27], occupation [52], personality [6, 22], political profiling [43], and race or ethnicity [40, 49]. Reddit is used for personality and demographic studies [12, 24, 25]. Other datasets include Odnoklassniki [2], Facebook-like behavioral records [34], CLEF-PAN web communication corpora [27], Arabic news articles [26], Persian political tweets [43], Korean dialogue data [11], Punjabi news data [46], Amazon reviews [45], and synthetic smart-home conversational sessions [7]. Although English-language datasets dominate the literature, some studies have explored user profiling in Arabic, Turkish, Persian, Korean, and Punjabi contexts [26, 43, 46, 52]. Table 7 presents an overview of the datasets used in the selected studies, detailing their source platforms, languages, corpus sizes, and public availability. Table 7 reports only those studies for which the underlying dataset is publicly available.

4.4 Methods and Models

The reviewed studies include traditional machine learning models, deep learning architectures, embeddings, graph-based models, ontology-based methods, recommender-system models, and LLM-based approaches. Traditional approaches include Logistic Regression, Decision Trees, KNN, SVM, Naive Bayes, Random Forests, MLPs, and XGBoost, usually combined with BoW, TF-IDF, n-grams, POS tags, LIWC features, or stylistic markers [8, 22, 25, 27]. Embedding-based approaches include Word2Vec [2], GloVe [22], graph embeddings for socioeconomic prediction [3], and transformer-based models such as BERT and DistilBERT [14, 26, 49, 52].

Deep learning methods include CNN, LSTM, BiLSTM, GRU, and fusion strategies for personality or author profiling [17, 26, 37]. Ontology-based and semantic user modeling approaches use domain knowledge and tagging histories to enrich user profiles [10, 18]. Recommender-oriented studies include content-based filtering, collaborative filtering, hybrid filtering, matrix factorization, and deep learning models for sparsity and cold-start challenges [15, 29, 32, 48]. LLM-based studies include GPT-4o, GPT-4o-mini, DeepSeek-V3, LoRA/QLoRA fine-tuning, and profile generation from conversational or review histories [30, 43, 45].

The selected studies reflect a broad methodological range, from classical techniques to the most recent LLM-based approaches. Fourteen studies built their profiling models on traditional machine learning methods such as logistic regression, SVM, Naive Bayes, and random forests [2, 6, 8, 12, 20, 22, 24, 25, 27, 33, 34, 36, 40, 44]. Four studies applied deep learning models such as CNN, LSTM, and

Table 6: Summary of related studies on behavioral, preference, and intent profiling.

Author	Goal	Dataset	Techniques	Gap
[31]	Machine learning approach to user profiling for data annotation of online behavior	Online behavioral dataset English	K-Means clustering, distortion score, silhouette score; ML classifiers	Static behavioral snapshots; difficult to capture temporal evolution; annotation subjectivity
[7]	On-device user profiling from natural language interactions	EdgeWisePersona: synthetic smart-home conversational sessions English	LLM inference	Synthetic dataset; limited real-world diversity; on-device constraints restrict model capacity
[30]	Benchmark LLMs for dynamic user profiling and personalized response generation	Multi-domain benchmark (conversational histories) English	GPT-4o, Llama	Benchmark scope may not reflect real deployment; LLM variability; cost
[5]	LLM-based cross-modality retrieval to improve recommendation performance	Cross-modal recommendation benchmarks English	LLM-based retrieval; cross-modal feature fusion	Multimodal complexity; reproducibility depends on proprietary LLM access; limited user behavioral signals
[48]	Enhance recommendations through hybrid sentiment classification and user profiling	E-commerce reviews English	Hybrid model: sentiment analysis + collaborative filtering + user profiling	Sentiment labels may be noisy; domain-specific; limited evaluation across diverse platforms
[43]	Generate user profiles from political tweets to capture political intent and opinion	6M tweets from 3,000 users (2024 Persian Presidential Election)	GPT-4o; extractive and abstractive profiling; macro-F1	Narrow political domain; Persian-only; high inference cost; closed-model reproducibility issues
[45]	Explainable temporal user profiling from Amazon reviews (preference/interest modeling)	Amazon Movies/TV and Video Games (18K reviews) English	GPT-4o-mini; temporal LLM-based profile construction	English-only; underperforms baselines; cost and reproducibility concerns; temporal drift not fully addressed
[15]	Content-based filtering for moisturizer recommendation based on skin type	Moisturizer product dataset	Content-based filtering	Narrow domain (cosmetics); limited user behavioral signals
[18]	Learn user profiles from ontology-based representations	Web-based user interaction data	Ontology-based user profile learning	limited to structured domain knowledge; scalability concerns

Table 7: Summary of the dataset, platform, languages, size, and availability (studies with explicitly stated public data availability).

Author	Platform	Language	Size	Availability
[3]	Twitter	English	5.1K	GitHub
[27]	Twitter	English, Spanish, Portuguese, and Arabic	15K	Zenodo
[52]	Twitter	Turkish	30K	GitHub
[6]	myPersonality	English	9.9K	Kaggle
[25]	Reddit	English	9K	Kaggle
[24]	Reddit	English	10K	Kaggle
[12]	Reddit	English	850K	GitHub
[26]	News platforms	Arabic	8K	GitHub
[45]	Amazon	English	18K	GitHub
[22]	tweets	English	17K	Kaggle
[36]	tweets	English	16K	GitHub
[44]	Blogs and social media	Spanish	84K	PAN
[46]	Twitter	Punjabi	1.9K	GitHub
[11]	NIKL	Korean	27K	KLI
[35]	Twitter	Indonesian	240K	GitHub
[1]	Google Places Review	English	33+M	cseweb
[37]	PersonalityCafe forum	English	8.6K	Kaggle
[8]	Twitter	English	8.6K	Kaggle and myPersonality
[17]	Facebook	English	12K	myPersonality
[7]	Smart-home NLI	English	10K	huggingface
[30]	Multi-domain conversational	English	10K	GitHub
[5]	MovieLens	English	100K	MovieLens

BiLSTM [17, 26, 35, 37], while four others turned to transformer-based models, primarily BERT, to capture richer semantic representations [1, 46, 49, 52]. At the more recent end of the landscape, six studies adopted LLM-based approaches [5, 7, 11, 30, 43, 45], and six combined multiple methods or data sources in hybrid or multimodal designs [3, 10, 15, 18, 19, 31]. For these aggregate counts,

each study was assigned to its predominant methodological family, although some studies evaluate methods from multiple families.

4.5 Evaluation Metrics and Reported Performance

The studies report heterogeneous evaluation metrics, including accuracy, F1-score, macro-F1, AUC, Pearson correlation, mean absolute error, distortion score, and silhouette score. For example, gender identification studies report accuracy values around 0.84–0.92 in specific datasets [26, 27, 46]; personality prediction studies report accuracy values that vary widely depending on the dataset and personality framework [6, 8, 22, 25, 37]; socioeconomic prediction uses accuracy for occupational class and correlation or error metrics for income [3]; and clustering-based behavioral profiling reports distortion and silhouette scores [31]. LLM-based political profiling reports macro-F1 [43].

Because tasks, datasets, labels, and class distributions differ substantially, reported values should not be compared as if they came from a single benchmark. Results are often affected by class imbalance, dataset size, platform specificity, label quality, and whether the task is binary, multiclass, regression-based, or clustering-based.

4.6 Limitations and Challenges

The reviewed literature reveals several recurring limitations. First, many studies rely on narrow, platform-specific datasets from Twitter, Reddit, Facebook, or single-language corpora, which limits generalizability to multilingual and cross-cultural settings [2, 24, 34]. Second, several models depend on surface-level textual features such as BoW and TF-IDF [8], while systematic comparisons with contextualized models and LLMs remain limited. Third, demographic profiling research focuses heavily on age and gender, while attributes such as race, ethnicity, occupation, income, education, unemployment, and partisan affiliation remain less cohesive as a

research agenda [12, 40, 49, 52]. Fourth, multimodal profiling shows promise [21], but integration costs and evaluation complexity remain barriers. Fifth, class imbalance, data sparsity, and cold-start scenarios degrade performance, especially for minority groups and underrepresented personality types [24, 25, 37]. Finally, ethical concerns regarding privacy, informed consent, transparency, and fairness are frequently acknowledged but less often addressed through concrete technical or procedural mechanisms [16, 34, 41].

5 SYNTHESIS OF FINDINGS

Figure 3 summarizes the field based on the findings of the systematic review. This section discusses the findings of the review and their implications for user profiling from textual data. The section first examines emerging trends in the literature, identifies research gaps and future directions, and discusses their relevance to Web and Multimedia Systems. Finally, it outlines potential threats to validity associated with the review process and its findings.

Main Trends. The reviewed studies suggest a shift from hand-crafted textual features and classical machine learning toward embeddings, transformer-based models, multimodal fusion, and LLM-based profiling. Word embeddings and graph embeddings have improved over purely surface-level representations in some tasks [2, 3], while multimodal work suggests that combining textual and behavioral data can enrich user representations [21]. More recent work explores LLMs for user profile extraction, temporal profiling, and natural-language profile generation [42, 43, 45].

Research Gaps. Several gaps emerge from the selected literature. There is still a lack of diverse and representative datasets, particularly for multilingual and cross-cultural user profiling. Non-English studies exist for Arabic, Turkish, Persian, Korean, and Punjabi contexts [26, 43, 46, 52], but these efforts remain fragmented. Sensitive attributes such as race, ethnicity, income, education, and unemployment are less studied than age and gender, and when they are studied, they often rely on specific populations, noisy labels, or proxy variables [3, 40, 49]. There is also limited standardization of metrics and benchmarks, making it difficult to compare results across studies.

Most of the research gaps identified in this review reflect recurring patterns in the literature, including a strong reliance on English-language datasets, a predominant focus on age, gender, and personality traits, and the limited attention given to attributes such as education level and unemployment. Many studies faced recurring practical limitations, including small, platform-specific, or demographically skewed datasets, which restrict the generalizability of their findings.

Implications for Web and Multimedia Systems. For Web and multimedia systems, user profiling from text can support personalization, recommendation, adaptive interfaces, moderation, and accessibility. However, WebMedia systems increasingly combine multiple modalities and interaction signals, so future work may also investigate how textual profiles interact with visual content, user interactions, temporal dynamics, and social context. This is especially relevant for social platforms, recommender systems, and user-centered multimedia applications.

Ethical Considerations. Ethical concerns are central to user profiling because inferred attributes may include sensitive demographic,

socioeconomic, psychological, political, or behavioral information. Studies predicting race, ethnicity, income, political affiliation, personality, or health-related behavior require particular caution. Future work should incorporate informed consent, privacy-preserving methods, fairness evaluation, transparent reporting, and explainable AI techniques. The review should avoid presenting sensitive-attribute prediction as a purely technical task, since such predictions can reinforce discrimination or enable manipulative personalization. However, the majority of the reviewed studies do not address ethical considerations in their research. This omission may lead to discriminatory outcomes and facilitate manipulative personalization practices. Based on the ethical coding performed during data extraction, Table 8 summarizes the main ethical issues identified in the reviewed studies, how they were typically addressed, and the remaining gaps.

Threats to Validity. Different factors may affect the validity of this review, including limitations in search coverage, selection bias, terminology variation, publication bias, and coding subjectivity. User profiling studies use diverse terminology, including user modeling, author profiling, demographic prediction, trait inference, and user representation, which may cause relevant studies to be missed. Although Google Scholar was used only as a complementary source, its search results and ranking may vary over time, which can affect reproducibility. To mitigate this limitation, we reported the search queries, search period, databases consulted, eligibility criteria, and study selection workflow. In addition, performance values were extracted from heterogeneous studies with different datasets and metrics, limiting direct comparability.

The search for relevant studies was conducted between January and June 2026. The retrieved studies were screened according to predefined quality assessment criteria, including Impact Factor and QUALIS classifications. To reduce selection bias, three authors independently reviewed the studies and evaluated their relevance to the objectives of this review. Any disagreements between the reviewers, regarding study selection or quality assessment, were resolved through discussion and consensus. Nevertheless, the inclusion and exclusion criteria involve some degree of subjective judgment, and relevant studies may have been inadvertently excluded during the title, abstract, or full-text screening stages.

6 CONCLUSION

This systematic literature review examined research on user profiling from textual data from multiple perspectives, including the user attributes investigated in the literature, the textual sources and datasets employed, the computational methods adopted, and the evaluation strategies used to assess profiling models. The reviewed studies show that textual data has been used to infer a broad range of user attributes, including demographic characteristics, personality traits, preferences, behaviors, intentions, and socioeconomic indicators.

The reviewed studies address three broad profiling targets: demographic attributes, mainly age and gender across 15 studies, personality traits using Big Five or MBTI frameworks across 12 studies, and behavioral or intent attributes across 9 studies. Because some studies address more than one profiling target, category counts are not mutually exclusive. Education and unemployment

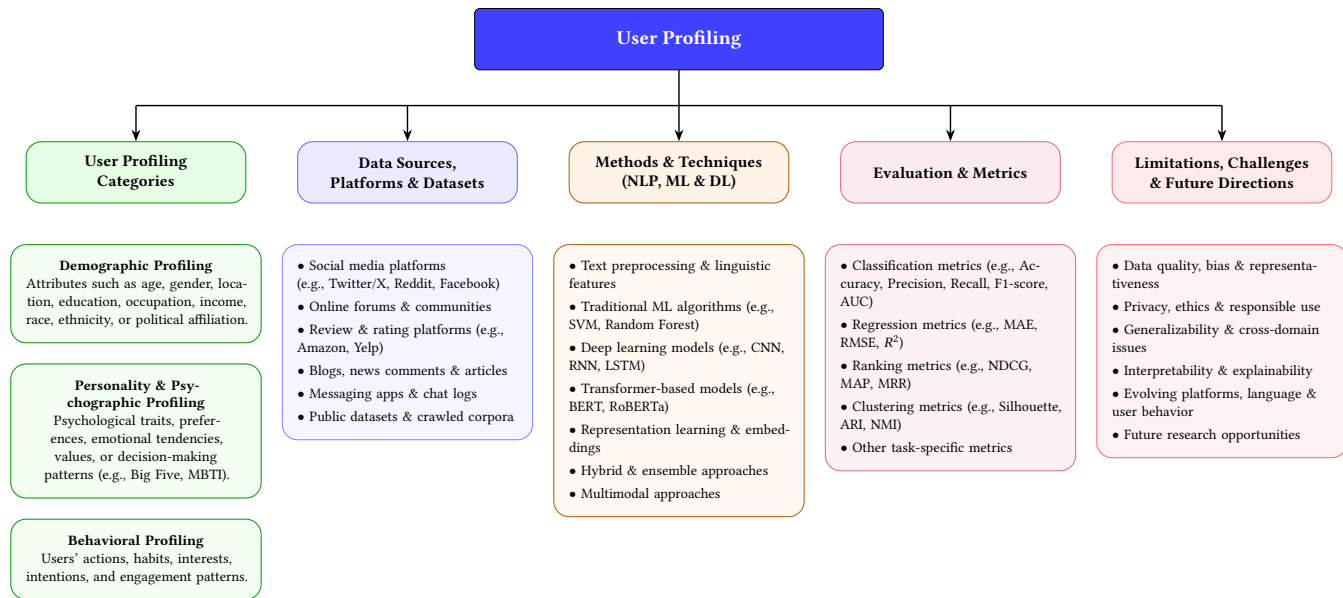


Figure 3: User profiling overview.

Table 8: Ethical issues identified in the reviewed studies.

Ethical issue	# Studies	Typical treatment	Main gap
Privacy and consent	8 studies: [1, 11, 12, 19, 24, 26, 40, 49]	Usually acknowledged as a concern when collecting or processing user-generated content.	Rarely operationalized through privacy-preserving methods, consent procedures, or data minimization strategies.
Bias and fairness	8 studies: [1, 11, 12, 19, 24, 26, 40, 49]	Sometimes discussed in relation to class imbalance, demographic skew, or platform-specific samples.	Few studies report fairness metrics, subgroup performance, or bias mitigation strategies.
Sensitive attributes	7 studies: [1, 12, 19, 24, 26, 40, 49]	Often modeled as prediction targets, especially for demographic, socioeconomic, political, or psychological profiling.	Limited discussion of potential harms, misuse, discrimination, or downstream effects on users.
Reproducibility	6 studies: [1, 12, 19, 24, 40, 49]	Dataset and code availability vary substantially across studies.	Weak reporting of preprocessing steps, hyperparameters, dataset access, and experimental protocols.

appeared in none of them, which is a notable gap. Twitter was the most common data source, followed by Reddit and Amazon, with additional corpora in Russian, Arabic, Turkish, Persian, Korean, Indonesian and Punjabi. Most of the studies used English data, and non-English efforts remain scattered with no shared benchmarks.

The literature has progressed from traditional feature engineering and classical machine learning toward embeddings, transformer-based models, multimodal approaches, and LLM-based user profiling. Methodologically, most of the studies used classical machine learning, some have applied deep learning and BERT-based models, while a few have adopted LLMs and hybrid or multimodal designs, reflecting a clear shift from handcrafted features toward contextualized and generative approaches. The analysis also revealed a growing interest in explainable AI techniques, reflecting an increasing demand for transparency and interpretability in user profiling systems. Evaluation metrics varied widely across studies, including accuracy, F1-score, AUC, and Pearson correlation.

However, important limitations remain, including dataset bias, class imbalance, lack of multilingual and cross-cultural datasets, limited reproducibility, inconsistent evaluation protocols, and insufficient treatment of ethical risks. The findings further highlight

the need for approaches capable of generalizing across domains, languages, platforms, and user populations.

Overall, this review consolidates the current body of knowledge on user profiling from textual data, identifies key research gaps, and outlines promising directions for future research toward more reliable, interpretable, and ethically responsible profiling systems. Specifically, further work should prioritize multilingual and cross-cultural datasets, standardized evaluation benchmarks, transparent reporting, privacy-preserving methods, explainable AI, and careful treatment of sensitive attributes. Integrating textual, visual, behavioral, and network-based data may provide richer user representations, but such integration must be guided by responsible data practices and clear user protections. Given the increasing relevance of user profiling for Web and Multimedia Systems, advances in these areas have the potential to enhance personalization, recommendation, user modeling, content adaptation, and human-computer interaction.

ACKNOWLEDGMENTS

This study is partially supported by the National Council for Scientific and Technological Development - CNPq (processes 314603/2023-9, 441444/2023-7, and 444724/2024-9) and INCT TILD-IAR (proc. 408490/2024-1).

The authors used generative artificial intelligence tools to improve the English language, grammar, and readability of the paper. All scientific content, study design, analyses, interpretations, and conclusions were conceived, verified, and approved by the authors, who assume full responsibility for the content of this work.

REFERENCES

- [1] Jemal Abate, Felipe Peixoto, João Bald, Myriam Delgado, and Thiago Silva. 2026. A BERT-Based Approach for Gender Inference from Place Reviews: Applications to Urban Representativeness. In *Anais do X Workshop de Computação Urbana (Praia do Forte/BA)*. SBC, Porto Alegre, RS, Brasil, 1–14. <https://doi.org/10.5753/courb.2026.24116>
- [2] Anton Alekseev and Sergey Nikolenko. 2017. Word embeddings for user profiling in online social networks. *Computacion y Sistemas* 21, 2 (2017), 203–226. <https://doi.org/10.13053/CyS-21-2-2734>
- [3] Nikolaos Aletras and Benjamin Paul Chamberlain. 2018. Predicting Twitter User Socioeconomic Attributes with Network and Language Information. In *Proceedings of the 29th on Hypertext and Social Media*. ACM, New York, NY, USA, 20–24. <https://doi.org/10.1145/3209542.3209577>
- [4] Rodrigo Ribeiro Alves, Victor Hugo da Silva Lembor, and Alessandra Alaniz Macedo. 2025. Systematic review: gamification on smartphones for children in educational and therapeutic contexts. In *Brazilian Symposium on Multimedia and the Web (WebMedia)*. SBC, 676–685.
- [5] Fahad Anwaar, Adil Mehmood Khan, Muhammad Khalid, Ahammad Khalil, and Muhammad Awais. 2024. LLM Based Cross Modality Retrieval to Improve Recommendation Performance. In *2024 29th International Conference on Automation and Computing (ICAC)*. 1–6. <https://doi.org/10.1109/ICAC61394.2024.10718796>
- [6] S. Aslam, M. Usman Javed, S. Maria Aslam, M. M. Iqbal, H. Ahmad, and A. Tariq. 2025. Personality Prediction of the Users Based on Tweets through Machine Learning Techniques. *Journal of Computing & Biomedical Informatics* 8, 02 (2025). <https://www.jcbi.org/index.php/Main/article/view/796>
- [7] Patryk Bartkowiak and Michal Podstawski. 2026. EdgeWisePersona: A Dataset for On-Device User Profiling from Natural Language Interactions. In *Proceedings of the 34th ACM Conference on User Modeling, Adaptation and Personalization (Gothenburg, Sweden) (UMAP '26)*. Association for Computing Machinery, New York, NY, USA, 361–365. <https://doi.org/10.1145/3774935.3806788>
- [8] Srilakshmi Bharadwaj, Srinidhi Sridhar, Rahul Choudhary, and Ramamoorthy Srinath. 2018. Persona Traits Identification based on Myers-Briggs Type Indicator(MBTI) - A Text Classification Approach. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 1076–1082. <https://doi.org/10.1109/ICACCI.2018.8554828>
- [9] Craig S. Brinkman, Shira Gabriel, and Elaine Paravati. 2020. Social achievement goals and social media. *Computers in Human Behavior* 111 (oct 2020), 106427. <https://doi.org/10.1016/j.chb.2020.106427>
- [10] Iván Cantador, Martin Szomszor, Harith Alani, Miriam Fernández, and Pablo Castells. 2008. Enriching ontological user profiles with tagging history for multi-domain recommendations. *CEUR Workshop Proceedings* 351 (2008), 5–19.
- [11] Sang-Hyun Cho, Dohyun Kim, Hyuk-Chul Kwon, and Minh Kim. 2024. Exploring the potential of large language models for author profiling tasks in digital text forensics. *Forensic Science International: Digital Investigation* 50 (2024), 301814. <https://doi.org/10.1016/j.fsidi.2024.301814> DFRWS APAC 2024 - Selected Papers from the 4th Annual Digital Forensics Research Conference APAC.
- [12] Federico Cinus, Corrado Monti, Paolo Bajardi, and Gianmarco De Francisci Morales. 2025. Uncovering the Sociodemographic Fabric of Reddit. *arXiv preprint arXiv:2502.05049* (2025).
- [13] Ayse Cufoglu. 2014. User Profiling - A Short Review. *International Journal of Computer Applications* 108, 3 (2014), 1–9. <https://doi.org/10.5120/18888-0179>
- [14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Jill Burstein, Christy Doran, and Tamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. <https://doi.org/10.18653/v1/N19-1243>
- [15] Muhammad Edi, Azzahra Putri, Andi Nur, and Genta Nazwar. 2025. Application of Content-Based Filtering for Moisturizer Recommendation System Based on Skin Type Suitability. 8 (2025), 25–29.
- [16] Christopher Ifeanyi Eke, Azah Anir Norman, Liyana Shuib, and Henry Friday Nweke. 2019. A Survey of User Profiling: State-of-the-Art, Challenges, and Solutions. *IEEE Access* 7 (2019), 144907–144924. <https://doi.org/10.1109/ACCESS.2019.2944243>
- [17] Kamal El-Demerdash, Reda A. El-Khoribi, Mahmoud A. Ismail Shoman, and Sherif Abdou. 2022. Deep learning based fusion strategies for personality prediction. *Egyptian Informatics Journal* 23, 1 (2022), 47–53. <https://doi.org/10.1016/j.eij.2021.05.004>
- [18] Victoria Eyharabide and Analía Amandi. 2012. Ontology-based user profile learning. *Applied Intelligence* 36, 4 (2012), 857–869. <https://doi.org/10.1007/s10489-011-0301-4>
- [19] Quan Fang, Jitao Sang, Changsheng Xu, and M. Shamim Hossain. 2015. Relational User Attribute Inference in Social Media. *Trans. Multi.* 17, 7 (July 2015), 1031–1044. <https://doi.org/10.1109/TMM.2015.2430819>
- [20] Mehwish Fatima, Komal Hasan, Saba Anwar, and Rao Muhammad Adeel Nawab. 2017. Multilingual author profiling on Facebook. *Information Processing Management* 53, 4 (2017), 886–904. <https://doi.org/10.1016/j.ipm.2017.03.005>
- [21] Bruce Ferwerda and Marko Tkalcic. 2018. Predicting Users' Personality from Instagram Pictures. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*. ACM, New York, NY, USA, 157–161. <https://doi.org/10.1145/3209219.3209248>
- [22] Shruti Garg and Ashwani Garg. 2021. Comparison of machine learning algorithms for content based personality resolution of tweets. *Social Sciences and Humanities Open* 4, 1 (2021), 100178. <https://doi.org/10.1016/j.ssho.2021.100178>
- [23] Teena George, T Sobha, Safa R H, Edrin Biju, Swathi Dinesh, and Akhil S Anil. 2025. Hybrid Recipe Recommendation System Using KNN and Collaborative Filtering Techniques. In *2025 Emerging Technologies for Intelligent Systems (ETIS)*. 1–6. <https://doi.org/10.1109/ETIS64005.2025.10961637>
- [24] Matej Gjurković, Vanja Mladen Karan, Iva Vukojević, Mihaela Bošnjak, and Jan Šnajder. 2021. PANDORA Talks: Personality and Demographics on Reddit. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, Lun-Wei Ku and Cheng-Te Li (Eds.). Association for Computational Linguistics, Online, 138–152. <https://doi.org/10.18653/v1/2021.socialnlp-1.12>
- [25] Matej Gjurković and Jan Šnajder. 2018. Reddit: A Gold Mine for Personality Prediction. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*. Association for Computational Linguistics, Stroudsburg, PA, USA, 87–97. <https://doi.org/10.18653/v1/W18-1112>
- [26] Hanen Himdi and Khaled Shaalan. 2024. Advancing Author Gender Identification in Modern Standard Arabic with Innovative Deep Learning and Textual Feature Techniques. *Information* 15, 12 (2024). <https://doi.org/10.3390/info15120779>
- [27] Catherine Ikae and Jacques Savoy. 2021. Gender identification on Twitter. *J. Assoc. Inf. Sci. Technol.* 73, 1 (Dec. 2021), 58–69. <https://doi.org/10.1002/asi.24541>
- [28] Rodolfo Jaffé. 2020. QUALIS: The Journal Ranking System Undermining the Impact of Brazilian Science. *bioRxiv* (2020). <https://doi.org/10.1101/2020.07.05.188425>
- [29] A. Jain, G. Jain, S. Nagar, et al. 2025. Rating Distribution-Aware Deep Cognitive Convolution Matrix Factorization for Recommendation Systems. *Arabian Journal for Science and Engineering* 50 (2025), 7441–7462. <https://doi.org/10.1007/s13369-024-09361-3>
- [30] Bowen Jiang, Zhuoqun Hao, Young Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo Jose Taylor, and Dan Roth. 2025. Know Me, Respond to Me: Benchmarking LLMs for Dynamic User Profiling and Personalized Responses at Scale. In *Second Conference on Language Modeling*. <https://openreview.net/forum?id=6ox8XZGOqP>
- [31] Moona Kanwal, Najeed A. Khan, and Aftab A. Khan. 2024. A Machine Learning Approach to User Profiling for Data Annotation of Online Behavior. *Computers, Materials and Continua* 78, 2 (2024), 2419–2440. <https://doi.org/10.32604/cmc.2024.047223>
- [32] Ikram Karabila, Nossayba Darraz, Anas El-Ansari, Nabil Alami, and Mostafa El Mallahi. 2024. Addressing data sparsity and cold-start challenges in recommender systems using advanced deep learning and self-supervised learning techniques. *Journal of Experimental & Theoretical Artificial Intelligence* 0, 0 (2024), 1–31. <https://doi.org/10.1080/0952813X.2024.2401364>
- [33] Mirco Kocher and Jacques Savoy. 2017. Distance measures in author profiling. *Information Processing Management* 53, 5 (2017), 1103–1119. <https://doi.org/10.1016/j.ipm.2017.04.004>
- [34] Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* 110, 15 (apr 2013), 5802–5805. <https://doi.org/10.1073/pnas.1218772110>
- [35] Dian Sa'adillah Maylawati, Zhafran Nouvaldi, Undang Syaripudin, Nur Lukman, Muhammad Insan Al-Amin, and Muhammad Ali Ramdhani. 2023. Big Five Personality Type Prediction on Twitter Users with the Long Short-Term Memory Algorithm. In *2023 9th International Conference on Wireless and Telematics (ICWT)*. 1–6. <https://doi.org/10.1109/ICWT58823.2023.10335257>

- [36] Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. 2018. Author Profiling for Abuse Detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, Emily M. Bender, Leon Derczynski, and Pierre Isabelle (Eds.). Association for Computational Linguistics, Santa Fe, New Mexico, USA, 1088–1098. <https://aclanthology.org/C18-1093/>
- [37] Anam Naz, Hikmat Ullah Khan, Sami Alesawi, Omar Ibrahim Abouola, Ali Daud, and Muhammad Ramzan. 2024. AI Knows You: Deep Learning Model for Prediction of Extroversion Personality Trait. *IEEE Access* 12 (2024), 159152–159175. <https://doi.org/10.1109/ACCESS.2024.3486578>
- [38] A. Naz, H. U. Khan, A. Bukhari, et al. 2025. Machine and deep learning for personality traits detection: a comprehensive survey and open research challenges. *Artificial Intelligence Review* 58 (2025), 239. <https://doi.org/10.1007/s10462-025-11245-3>
- [39] Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 372 (2021), n71. <https://doi.org/10.1136/bmj.n71>
- [40] Daniel Protiuc-Pietro and Lyle Ungar. 2018. User-level race and ethnicity predictors from twitter text. In *Proceedings of the 27th international conference on computational linguistics*. 1534–1545.
- [41] Erasmo Purificato, Ludovico Boratto, and Ernesto William De Luca. 2024. User Modeling and User Profiling: A Comprehensive Survey. (2024). [arXiv:2402.09660](https://arxiv.org/abs/2402.09660) <http://arxiv.org/abs/2402.09660>
- [42] Filip Radlinski, Krisztian Balog, Fernando Diaz, Lucas Dixon, and Ben Wedin. 2022. On Natural Language User Profiles for Transparent and Scrutable Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) (SIGIR '22). Association for Computing Machinery, New York, NY, USA, 2863–2874. <https://doi.org/10.1145/3477495.3531873>
- [43] Vahid Rahimzadeh, Ali Hamzehpour, Azadeh Shakery, and Masoud Asadpour. 2025. From Millions of Tweets to Actionable Insights: Leveraging LLMs for User Profiling. (2025). [arXiv:2505.06184v1](https://arxiv.org/abs/2505.06184v1) www.aaai.org
- [44] Francisco Rangel and Paolo Rosso. 2016. On the impact of emotions on author profiling. *Information Processing Management* 52, 1 (2016), 73–92. <https://doi.org/10.1016/j.ipm.2015.06.003>
- [45] Milad Sabouri, Masoud Mansoury, Kun Lin, and Bamshad Mobasher. 2025. Towards Explainable Temporal User Profiling with LLMs. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization (UMAP Adjunct '25)*. Association for Computing Machinery, New York, NY, USA, 219–227. <https://doi.org/10.1145/3708319.3733655>
- [46] Raheem Sarwar, Le An Ha, Pin Shen Teh, Fahad Sabah, Raheel Nawaz, Ibrahim A. Hameed, and Muhammad Umair Hassan. 2024. AGI-P: A Gender Identification Framework for Authorship Analysis Using Customized Fine-Tuning of Multilingual Language Model. *IEEE Access* 12 (2024), 15399–15409. <https://doi.org/10.1109/ACCESS.2024.3358199>
- [47] M. Sharma, A. Sarin, P. Gupta, S. Sachdeva, and A. V. Desai. 2014. Journal Impact Factor: Its Use, Significance and Limitations. *World Journal of Nuclear Medicine* 13, 2 (2014), 146. <https://doi.org/10.4103/1450-1147.139151>
- [48] T. Sumanth, gudapati Lakshmi Anasuya, Ganjinaboyina Tatwik Abhinav, Daggula Vasavi, and Dasari Rajesh. 2025. Enhancing Recommendations Through Hybrid Sentiment Classification and User Profiling. *International Journal on Advanced Computer Engineering and Communication Technology* 14, 1 (Apr. 2025), 69–79. <https://journals.mriindia.com/index.php/ijacect/article/view/172>
- [49] Zach Wood-Doughty, Paiheng Xu, Xiao Liu, and Mark Dredze. 2021. Using noisy self-reports to predict twitter user demographics. In *Proceedings of the ninth international workshop on natural language processing for social media*. 123–137.
- [50] Wenbo Wu, Masitah Ghazali, and Sharin Hazlin Huspi. 2024. A Review of User Profiling Based on Social Networks. *IEEE Access* 12 (2024), 122642–122670. <https://doi.org/10.1109/ACCESS.2024.3430987>
- [51] Tan Yigitcanlar. 2014. Empirical approaches in knowledge city research. *Expert Systems with Applications* 41, 12 (2014), 5547–5548. <https://doi.org/10.1016/j.eswa.2014.02.005>
- [52] Tolga İzdaş, Hikmet İskifoğlu, and Banu Diri. 2025. Occupation Prediction from Twitter Data. *Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi* 27, 80 (2025), 267–271. <https://doi.org/10.21205/deufmd.2025278013>