Digital Linguistic Markers for Age and Gender Prediction from Chat Data

- A Literature Review

[1]Rashmi S C, [2]Dr. Rachana P G

[1]Computer Programmer, [2]Assistant Professor
[1]Computer Science and Engineering,
[1]University B.D.T College of Engineering, Davanagere, India
[1]scrashmi@gmail.com, [2]rachanapg@gmit.ac.in

***Abstract*—This systematic literature review examines age and gender detection from instant messaging (IM) data, with particular emphasis on cross-lingual and code-switched contexts. Drawing on 49 peer-reviewed studies (2000–2024) from six major databases, it identifies linguistic, behavioral, and paralinguistic features (e.g., syntax, emoji use, word choice, and message patterns) as key demographic markers. The field has shifted from traditional machine learning (SVMs, Naïve Bayes) to deep learning and transformer-based models (e.g., BERT, Text2Gender), which perform better with informal and multilingual data. Major challenges include scarcity of annotated IM datasets, lack of support for low-resource/code-switched languages, and ethical concerns around privacy, consent, and gender inclusivity. Current approaches often assume binary gender categories, limiting fairness. The review calls for inclusive, privacy-preserving, and culturally adaptable NLP frameworks, emphasizing future research on scalable multilingual solutions, diverse gender/linguistic representation, and ethically responsible design.**

***Index Terms*—**Instant Messaging, Age and Gender Detection, Cross-Lingual Language, Emoji Usage, Code-Switching, Machine Learning, Author Profiling.

I. INTRODUCTION

With the rapid spread of smartphones and mobile broadband, instant messaging (IM) has emerged as the most widely used mode of daily communication, overtaking calls, emails, and social media. Popular apps such as WhatsApp (2B users), Messenger (1.3B), and WeChat (1.2B) dominate global use. IM platforms offer diverse features like group chats, multimedia sharing, and emoji/sticker exchange, which has fueled large-scale multilingual data generation. Research highlights that this multilingual nature creates opportunities for demographic profiling, where cross-lingual approaches extend inference to low-resource and code-switched languages increasingly prevalent in global IM communication.

Advances in deep learning and NLP have enhanced age and gender detection in instant messaging, where traditional approaches such as SVMs and lexicon-based methods often fall short with short, informal, and emotional text. Transformer-based models like BERT and its variants provide stronger contextualization for informal and code-switched conversations, showing superior performance in gender and emotion detection tasks [8][1]. Despite these gains, challenges persist in building inclusive and fair systems, as many still reinforce binary gender assumptions and overlook non-binary users [7]. Future work should advance context-aware, multilingual, and ethically inclusive models for demographic profiling in instant messaging.

Research has frequently found the linguistic differences over age and gender, indicates that women's language highlights emotional, while men more often reference professional. Age-related patterns include the youngsters use lesser negative emotions, more positive emotions, and older use greater future-tense. The emoji adds a contextual meaning, and their usage varies by demographics, women typically use more emoji, while men may employ a wider variety. Ages also effects the inconsistent, with some studies found higher usage among youngsters. The Demographic prediction success in social media, this review address the gap in demographic prediction from instant messages, focusing on Multi-lingual approaches to increase the applicability across variety of languages.

II. FINDINGS OF THE STUDY

Emoji Usage as Demographic Indicators in Instant Messaging

Research shows that emojis are prevalent in instant messaging (IM) and social networking platforms, with billions of users sent daily. Emoji usage increases the happiness and become more informative. In spite of the large variety of emojis are available, only a small subsets for nearly used on Twitter 3.5% of emojis represent 99% of usage, while on Facebook 15 emojis account for 99.6% of usage. This suggests that familiarity influences user choice.

In Gender-based women tends to use more emojis and more frequently than men, though men use a wider range of emojis. In business the IM contexts, gender differences were less pronounced, though women tends to use more emojis. Women emojis more negatively, with female participants showing stronger emotional responses to negative smileys.

Age-related trends shows that the overall emoji usage does not differ by age, younger users (18–29) prefer more playful and cheeky emojis compared to older users (40+). This suggests that emoji choice and expressiveness of usage can act as demographic markers reflecting generational communication styles.

Cross-Lingual and Code-Switching Approaches in Age and Gender Profiling from Instant Messaging Data

The rise of multilingual instant messaging (IM) platforms has increased the importance of cross-lingual and code-switching strategies for demographic inference. Code-switching and translanguaging serve as identity markers shaped by age, gender, and cultural context, rather than being purely functional. Studies highlight their social role—such as German students switching between German and English on WhatsApp to build group cohesion and bilingual adolescents maintaining peer connections across cultures through IM. Computational advances show feasibility in profiling code-mixed data: achieved strong gender classification (F1 = 0.947) on Roman Urdu–English SMS, while demonstrated gender prediction from English-Hindi tweets despite irregular grammar and transliteration challenges.

However, significant limitations persist. Cross-lingual models often lose accuracy across diverse datasets and remain biased toward dominant languages. Platform-specific differences further complicate analysis, with Twitter and stagram favoring English-dominant discourse, while WhatsApp promotes native language use [12]; platform affordances like character limits also shape switching behavior [13]. Addressing these gaps requires adaptive models, culturally contextualized features, and multilingual datasets to strengthen the robustness and generalizability of demographic profiling in IM.

Machine Learning and NLP Models for Age and Gender Detection from IM

Research on age and gender recognition in instant messaging (IM) has evolved from traditional machine learning to deep learning and transformer-based methods. Early studies used handcrafted features—such as lexical n-grams, stylistic markers, and POS tags—combined with classifiers like SVMs and Naïve Bayes. Schler et al. (2006) showed that informal blog data could reveal demographic traits, while Peersman et al. (2011) achieved over 80% accuracy on Netlog chats using character-based SVMs, proving that shallow lexical cues capture demographic information despite platform bias. More recent work leverages deep learning for end-to-end modeling of IM text. CNNs and RNNs have been applied to capture both local and sequential dependencies [4], while transformer-based approaches demonstrate strong performance on noisy, code-mixed datasets. Compared to feature-engineered pipelines focused on emoticons, punctuation, or slang [9], deep learning reduces manual effort and generalizes better across platforms. State-of-the-art models such as Text2Gender [11] achieved 84% accuracy for age and 86% for gender prediction, highlighting the effectiveness of contextual embeddings in IM demographic profiling.

III. ETHICAL, TECHNICAL, AND DATA CHALLENGES IN IM-BASED DEMOGRAPHIC PROFILING

Demographic profiling in instant messaging (IM) raises major ethical, technical, and inclusivity concerns. The casual, private, and highly variable nature of IM data makes issues of privacy, consent, and fairness particularly significant. Using IM content for age and gender prediction risks reinforcing existing stereotypes if bias is not carefully mitigated, while users are often unaware that their conversational data could be analyzed for demographic inference.

From a technical perspective, access to IM datasets remains restricted due to proprietary controls and strict privacy regulations. Even when data is available, anonymization can be degrade model performance by stripping away valuable linguistic cues, creating a trade-off between privacy and accuracy. Inclusivity poses another challenge: most systems continue to enforce binary gender classifications, thereby excluding non-binary and gender-diverse individuals and over simplifying social identity. Legal compliance further complicates data use, with regulations such as the GDPR requiring strict safeguards for collection, storage, and sharing of IM data.

To address these issues, researchers are encouraged to prioritize informed consent, actively reduce algorithmic bias, and adopt inclusive gender labels. Transparency and reproducibility in methods are essential, alongside interdisciplinary oversight to ensure ethical standards are upheld. Privacy-preserving datasets, fairness-aware algorithms, and strong security protocols offer practical pathways toward more responsible and socially conscious demographic profiling in IM.

IV. Comparative Analysis of Accuracy of model for age and gender detection in chat Data

To provide a clear understanding of which models perform best in different contexts, Table 1 summarizes the comparative accuracy of ML(Machine Learning), DL(Deep Learning), hybrid, and transformer-based models applied to IM data.

Table 1

Model Type	Example Studies / Models	Age Detection Accuracy	Gender Detection Accuracy	Code-Mixed / Cross-Lingual Performance

Traditional ML (SVM, Naïve Bayes, RF)	(Schler et al., 2006);
	(Peersman et al., 2011);
	(Imran & Iqbal, 2018)
		50–60%	73–80%	Weak in code-mixed settings; accuracy drops significantly

Deep Learning (CNN, RNN, LSTM, GRU)	(Basile et al., 2017); (Escobar-Grisales et al., 2021)	65–75%	80–85%	Moderate, struggles with transliteration and noisy IM data
Hybrid ML + Stylometry	(Abdallah et al., 2020); (Koch et al., 2022)	~70%	85.7% (WhatsApp dataset)	Limited, mainly tested in monolingual IM datasets

Transformer based (BERT, Text2Gender, Multilingual Embeddings)	(Thakur & Tickoo, 2023);
	(Kalra & Zubiaga, 2021);
	(Borquez et al., 2024);
		84%	86% (up to >90% in some corpora)	Best performance; robust to multilingual and code-switched contexts

Model Type	Example Studies / Models	Age Detection Accuracy	Gender Detection Accuracy	Code-Mixed / Cross-Lingual Performance
	(Younkin et al., 2024).			
Low-Resource / Code-Mixed ML & DL	(Arshad et al., 2024)– Roman Urdu; Devi & (Devi & Kannimuthu, 2023) – Tamil WhatsApp	54.28% (Roman Urdu)	71.14% (Roman Urdu); 70–75% (Tamil WhatsApp)	Shows feasibility; lower accuracy than high-resource datasets

V. Methodology

Datasets

The publicly available datasets derived from open chat corpora and synthetic IM data containing user demographic annotations (age and gender). Preprocessing steps includs tokenization, removal of Personally Identifiable Information (PII), conversion of abbreviations and emoji normalization to standardized forms.

Feature-Extraction

Feature engineering combined linguistic, stylometric, and behavioral markers:
Linguistic Features: word n-grams, part-of-speech tags, emoji frequency.
Stylometric Features: average message length, punctuation frequency, use of capital letters, typing delay.
Behavioral Features: message timing patterns, response delay, keystroke intervals (where available).

Machine Learning Models

We can use the following algorithms:
Support Vector Machine (SVM): used for high-dimensional stylometric feature classification.
Random Forest (RF): to handle nonlinear relationships among features.
LSTM(Long Short Term Memory) : for sequential modeling of text messages.
BERT(Bidirectional Encoder Representation Transformer)-based Model: fine-tuned on chat data for contextual understanding.
Hybrid Model: combining BERT embeddings with stylometric and behavioral features for increased prediction.

Evaluation Metrics

The models were evaluated using standard classification metrics: accuracy, precision, recall, and F1-score. Cross-validation was applied to ensure model robustness. Data imbalance was addressed through SMOTE (Synthetic Minority Over-sampling Technique).

VI. Conclusion

This study focused on cross-lingual and behavioral-linguistic techniques to examine the state of age and gender detection using instant messaging data. A wide range of linguistic, behavioral, and paralinguistic characteristics were found to be useful markers for drawing conclusions about demographics, including code-switching, emoji patterns, message timing, and word use. The transition from conventional machine learning models to deep learning and transformer-based architectures has greatly increased the accuracy of predictions in informal and multilingual communication contexts. However, the discipline continues to confront several obstacles, such as the lack of annotated datasets, particularly for low-resource and code-switched languages, and privacy, prejudice, and non-binary gender identity underrepresentation issues. Notwithstanding these problems, research in this field is still moving in the direction of more morally sound, inclusive, and situation-specific solutions. The research comes to the conclusion that further development of multilingual natural language processing, privacy-preserving frameworks, and socially acceptable and technically sound demographic profiling technologies are necessary for future advancement. These advancements will be essential for facilitating fair and accurate age and gender inference in the quickly growing field of international instant messaging.

References

[1] Burghoorn, M., de Boer, M. H. T., & Raaijmakers, S. (2020). Gender prediction using limited twitter data. ArXiv Preprint ArXiv:2010.02005.

[2] Dhir, A., Kaur, P., & Rajala, R. (2020). Continued use of mobile instant messaging apps: A new perspective on theories of consumption, flow, and planned behavior. Social Science Computer Review, 38(2), 147–169.

[3] Evans, D. (2020). Why the US government is questioning WhatsApp’s encryption. Erişim Adresi: Https://T. Ly/047W Filerskeepers.(2021). Solving Records Retention Once and for All: Confidently Decide How Long You Keep Your Data with Our Records Retention Schedules. Erişim Adresi: Https://T. Ly/OH53.

[4] Escobar-Grisales, D., Vásquez-Correa, J. C., & Orozco-Arroyave, J. R. (2021). Gender recognition in informal and formal language scenarios via transfer learning. Workshop on Engineering Applications, 171–179.

[5] Fullwood, C., Orchard, L. J., & Floyd, S. A. (2013). Emoticon convergence in Internet chat rooms. Social Semiotics, 23(5), 648–662.

[6] Goodin, D. (2021). WhatsApp gives users an ultimatum: share data with Facebook or stop using the app. Ars Technica, Available at: Https://Arstechnica. Com/Tech-Policy/2021/01/Whatsapp-Usersmust-Share-Their-Data-with-Facebook-or-Stop-Using-the-App.

[7] Jazi, S. Y., Mirzaeinia, A., & Jazi, S. Y. (2024). Analyzing gender polarity in short social media texts with BERT: the role of emojis and emoticons. ArXiv Preprint ArXiv:2406.09573.

[8] Kalra, A., & Zubiaga, A. (2021). Sexism identification in tweets and gabs using deep neural networks. ArXiv Preprint ArXiv:2111.03612.

[9] Koch, T. K., Romero, P., & Stachl, C. (2022). Age and gender in language, emoji, and emoticon usage in instant messages. Computers in Human Behavior, 126, 106990.

[10] Statista. (2021). Most popular global mobile messenger apps as of February 2025, based on number of monthly active users. https://www.statista.com/statistics/258749/most-popularglobal-mobile-messenger-apps/.

[11] Thakur, V., & Tickoo, A. (2023). Text2Gender: A Deep Learning Architecture for Analysis of Blogger's Age and Gender. ArXiv Preprint ArXiv:2305.08633.

[12] Wulandari, C., Hadianti, A., & Fhadilathusy, S. (2024). Code-Switching in Digital Communication: A Study of Bilingual Language Use in Popular Platforms. International Journal of Language and Culture, 3, 20–30. https://doi.org/10.63762/ijolac.v3i1.28.

[13] Yousif, A. S. A. (2024). Multilingualism in the Digital Age: Code-Switching and Translanguaging Online. Theory and Practice in Language Studies, 15(4), 1217–1225.