Gender Demographic Prediction on Twitter: A Comprehensive Review of Machine Learning and Data Analytics Approaches

Tarun Kumar Sharma[1], Vishal Shrivastava[2]

satarun@gmail.com[1], vishalshrivastava.cs@aryacollege.in[2]

Department of Computer Science and Engineering, Arya College of Engineering and IT, Jaipur (Rajasthan-India) [1,2]

Abstract. In this paper, the author will discuss the design and analysis of cutting-edge machine learning (ML) and data analytics solutions that will help increase the accuracy and reliability of gender prediction using Twitter data. The paper critically analyzes various state-of-the-art models such as Support Vector Machines (SVM), Random Forest, Gradient Boosting algorithms (XGBoost, LightGBM), and deep learning models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) and Transformer-based models (BERT, RoBERTa). Particular attention is given to such preprocessing methods as text normalization, tokenization, feature engineering (n-grams, POS tagging, embeddings), and incorporation of multimodal information - the textual information, profile metadata, user activity patterns. Moreover, the review also emphasizes the importance of dataset balancing with such methods as SMOTE and the influence of the linguistic, cultural, and regional diversity on the generalization of the model. The relative analysis of the benchmark data sets, including the PAN Author Profiling, indicates the main trends of performance and efficiency of the ensemble learning and transfer learning methods.

Keywords: Gender prediction, Twitter analytics, machine learning, deep learning, demographic profiling, natural language processing, feature engineering, social media mining, SVM, Random Forest, XGBoost, BERT, transfer learning, data imbalance

1. Introduction

The high pace of the development of digital communication and the widespread availability of social media sites have altered the manner in which people interact, communicate and represent themselves on the internet. Twitter, having more than 450 million monthly active users worldwide is one of the microblogging services that has been able to capture a huge variety of human expressions in brief written communications called tweets. These documents, and supplementary user data, including profile images, usernames, bios, hashtags, and time stamps give a highly informative stream of unstructured data. Using this data as social intelligence (e.g. opinion mining, behavior modelling, demographic inference, etc.) has emerged as a more and more popular subject of inquiry in

IICTDS 2025, December 19-20, Chandigarh, India
Copyright © 2026 EAI
DOI 10.4108/eai.19-12-2025.2366281

computational social science, digital marketing, and policy making. Gender prediction is a very essential part of user profiling among other demographic attributes. Proper gender classification improves the accuracy of targeted advertising, personal recommendations and even epidemiological studies in population health. Nevertheless, it is highly technical and ethically challenging to extract such latent properties of text dominant, anonymous platforms such as twitter. Users do not identify themselves by gender at all time, and language styles may be varied due to culture. This requires advanced machine learning (ML) and deep learning (DL) algorithms which can identify gender through indirect cues. The rationale of the investigation of gender demographic prediction on Twitter is based on the versatile applications and increased use of automated decision-making systems in civilian and governmental sectors. Companies apply demographic knowledge to marketing on a one-to-one basis; administrations employ social media to gauge societal moods and scholars apply digital footprints to test behavior patterns of various gender identities. However, although gender detection is important, conventional techniques that use profile name or keywords alone, lack precision and not only in terms of cultural generalizability but also in regard to informal internet language. Since the state-of-the-art machine learning, especially natural language processing (NLP) and transformer-based models, nowadays it is possible to model multi-faceted semantic, syntactic, and stylistic micro-connuances of a tweet. These methods provide a chance to develop more trustworthy and sophisticated gender prediction methods basing on a multitude of characteristics, such as textual information, user metadata, relationship behavior, and the system structure. The given paper aims at discussing and examining these changing methodologies and focusing on their efficiency, shortcomings, and perspectives. The initial gendered prediction on Twitter was mostly rule based systems or name based systems which used gendered names or pronouns to determine identity. They were only limited and subject to error particularly when dealing with non English users or when dealing with ambiguous names. As supervised ML methods evolved, scholars started using classifiers such as Naive Bayes, Support Vector Machines (SVM), and Decision Trees with bag-of-words (BoW), n-gram and part-of-speech (POS) tagging features. Although interpretable, these models could not do well at capturing semantic context or long term dependencies in text. The achievement was the advent of deep learning models, such as Convolutional Neural Networks (CNNs) and Long Short-term memory networks (LSTMs), which were better able to capture linguistic patterns and sequential data. Later, with the emergence of transformer-based models like BERT and RoBERTa, text classification tasks were transformed once again through contextual embeddings, which performed far better than the traditional text classification task of static word embeddings, like Word2Vec and GloVe. The recent research has presented the concepts of multimodal which are textual content in combination with features of a user profile including profile pictures, followers-following networks, and track of activities. Various sources of information are included in these hybrid models, which cause them to be more predictive and more generalized. Also, ensemble models and transfer learning schemes have become popular due to their capacity to take advantage of prior knowledge and minimize overfitting, particularly when few resources are available.

2. Literature Review

The changing nature of the field of demographic inference using social media has been greatly enriched by academic input that highlights the use of machine learning (ML) and deep learning (DL) and data analytics methods towards gender prediction on Twitter. In 2024, O'Connor et al. [1] presented a scoping review of literature that indicated the scope of methods and annotated datasets to infer gender and age with a focus on the shift to more complex DL systems, instead of lexicon-based heuristics. To this, Roy et al. [2] investigated the opportunities of decision marketing models with big data analytics to classify gender, especially that the textual metadata may be supplemented with marketing-focused features to improve demographic segmentation. Liu et al. [3] have also shown a comparative analysis between classical and DL models and found that classical models are fine with handcrafted features but deep models such as LSTM and CNN have better generalisation capabilities particularly in cross domain scenarios. The same tendency was followed by Onikoyi et al. [4], who utilized descriptive texts and applied multiple ML classifiers, determining that performance of ensemble methods is better in heterogeneous tweet settings. In the health forum, Park and Woo [5] applied gender classification to sentiment analysis and DL by taking the argument of the strong correlation of emotional tone and language style with user gender. Expanding on the definition, Kursuncu et al. [6] gave a thorough review on predictive analysis in twitter including a diverse range of applications such as gender prediction and highlighted the necessity of extracting semantic features. McCormick et al. [7] put a focus on data collection and preprocessing tools in Twitter demographic research, which showed that effective data preprocessing pipelines are the key to a robust ML model. According to Katna et al. [8], linguistic and behavioral features proved to be essential in gender and age prediction, especially when metadata is used. Al-Azani and El-Alfy [9] brought a new direction by suggesting video-based demographic analytics, in which multimodal content can be used to amplify the accuracy of inferences. Alvarez-Carmona et al. [10] have also embraced a graphical method by taking advantage of image features on Twitter accounts to complement the traditional textual models. Sloan et al. [11] have shown that sociodemographic derivation is possible with combination of both linguistic and follower network properties, and they recommended a mixed-methods approach of better representation of demographic heterogeneity. A master thesis by Malik [12] used supervised learning on telecom data, and demonstrated that it is plausible that Twitter-based models can be transferred to other industries (e.g., telecom) in case of aligned feature engineering. Previous studies had been conducted on predicting personality with Twitter data by Golbeck et al. [13], and their results, inadvertently, offered the basis of the further gender prediction as the linguistic style and psychological characteristics are correlated. Kang and Kang [14] made a contribution, using DL on the multimodal crime data, or, though not gender-specific, provided the information on the feature fusion methods that could be used in demographic modeling. Bleidorn and Hopwood [15] examined the role of ML in the personality theory, and in a way that strengthens the fact that psychological modeling systems are needed in the demographic classification systems. The study by Alsubhi et al. [16] applied the personality modeling to Arabic tweets with the AraBig5 model and, therefore, contributed to the underrepresented area of gender

prediction in non-English social media data. The relevance of the work of Lee et al. [17] in detecting the presence of spammers with the help of social honeypots and ML are applicable in the process of filtering out noise caused by bots when employing demographic models in conjunction with the text features.The research by Hecht et al. [18] examining the dynamics of the location field in user profiles revealed that the auxiliary metadata fields could be of significant importance in enhancing demographic models when interacting with text features. Blank [19] argued that digital divides exist on allegedly egalitarian platforms such as Twitter, and thus gender modeling should be context-sensitive in terms of its sociological implications, based on the argument of demographic skewness. Cherenkov et al. [20] moved the discussion to the customer experience in the hospitality industry and stated that predictive ML models (such as gender ones) could be applied to have a more personalized experience in business. Benjachairat et al. [21] demonstrated that their ML-based classification of suicidal ideation severity based on tweets could help address the issue of mental health intervention by offering gender-sensitive models to match an assessment with demographic specificities. The predictive ability of Twitter in the political forecasting was proved by Jahanbakhsh and Moon [22], although gender was not their primary objective, their data processing pipeline was helpful to study as a lesson in demographic analysis. The data management issues discussed by Prosperi et al. [23] regarding precision public health could be directly applied to the demographic prediction of gender in particular, particularly in the measures of privacy, data imbalance, and annotation consistency. When proposing a model to predict mental illness using Twitter data, Reece et al. [24] have demonstrated that long-term linguistic predictors can also be used to predict demographic factors. The entire survey of Latif et al. [25] on the use of data science in the time of COVID-19 also highlighted the necessity of transdisciplinary tools, such as demographic inference to simulate human behavior. Yang et al. [26] provided a DL artifact to measure personality via texts, which confirmed the usefulness of personality-sensitive gender classifiers on Twitter. Gender-related linguistic variation was studied by Bamman et al. [27] who also discovered that some lexical structures are highly gendered and can be used to construct powerful predictive paradigm. The Pereira-Kohatsu et al. [28] overlayed the sentiment and toxicity detection-based models to track hate speech on Twitter and indicated that demographic factors such as gender may contribute to the formation or targeting of hate speech. Sentiment analysis and user profiling were also the pillars of election prediction studies conducted by Chauhan et al. [29], which once again supports the role of gender in the conceptualization of political communication. Schwartz et al. [30] adopted a ground-breaking strategy, where they modelled age, gender and personality using open-vocabulary models and showed that unsupervised methods can equal or outperform supervised models as long as huge datasets are involved. Karami et al. [31] designed an ML-based classifier that can automatically assign LGBTs on Twitter profiles, which showed that gender prediction models can become more representative in terms of inclusivity. A suicide detection model of Arabic tweets Baghdadi et al. [32] with DL provided by Baghdadi et al. provided an architecture that can be adapted to gender classification through proper changes. In a study conducted by Sharma et al. [33] on the relationship between social media marketing and apparel purchase intentions, the researcher determined that the gender of the user plays a crucial role in the engagement patterns, an observation that can be used to predict gender in e-commerce using the engines. Rahman et al. [34] used Twitter sentiment and

census information to discuss the opinions of COVID-19 reopening in the US and showed that using the social media information with the offline population can be valuable in better calibration of the model. Lastly, Jamarani et al. [35] summarized the use of big data and predictive analytics in different industries, such as healthcare and marketing, which indirectly validates the scalability of gender prediction processes in the multi-domain context.

3 Analysis and Findings

Gender demographic prediction on Twitter has experienced an influx of research interest, as a result of the increasing need to create content that is personalized, focuses, and analyzes social behavior. The unstructured data in Twitter, which consists of tweets and user metadata, images and behavioral signals has proved to be a fruitful area where demographic inferences can be made. Nevertheless, this data has methodological limitations such as heterogeneity of the data, variance between cultures and ethical issues, which make it quite difficult to accurately predict gender with such data. It requires the revision of the multiple machine learning (ML) and deep learning (DL) approaches that have been adopted by researchers in the field. To present the comprehensive synthesis of knowledge on 35 reviewed papers, analytical tables and five plots were developed. These help to learn about trends in methodological approaches, the choice of features, the variety of datasets, the application of algorithms, the model performance, and the repetitive constraints. These findings are expounded in the following sections.

Table 1. Classification of reviewed papers based on methodology.

S. No.	Reference No.	Author(s) & Year	Methodology Type	Application Domain
1	[1]	O'Connor et al. (2024)	Scoping Review	Demographic Prediction
2	[3]	Liu et al. (2021)	Comparative Modeling	Gender & Age Inference
3	[4]	Onikoyi et al. (2023)	Supervised ML	Textual Gender Detection
4	[5]	Park & Woo (2019)	DL with Sentiment Analysis	Health Forum Gender Classification
5	[8]	Katna et al. (2022)	Hybrid ML-DL	Tweet-based Age & Gender Prediction
6	[10]	Alvarez-Carmona et al. (2018)	Visual + Textual	Multimodal Twitter Profiling
7	[31]	Karami et al. (2021)	Supervised Classification	LGBT Gender Categorization on Twitter

Table 2. Types of datasets used.

S. No.	Reference No.	Dataset Used / Source	Dataset Type	Language
1	[1]	Multiple Annotated Datasets	Annotated, Public	Multilingual
2	[3]	Collected Twitter Data	User Profile + Tweets	English
3	[8]	In-house Tweet Corpus	Profile + Content	English
4	[16]	AraBig5 Twitter Dataset	Arabic Tweets	Arabic
5	[30]	Open Vocabulary Social Media Corpus	Public Facebook/Twitter	English
6	[27]	Gendered Lexicon Dataset	Annotated Corpus	English
7	[31]	Custom LGBT Twitter Dataset	Labeled Accounts	English

Table 3. Feature categories extracted from twitter

S. No.	Reference No.	Author(s) & Year	Feature Types Extracted
1	[4]	Onikoyi et al. (2023)	N-grams, POS tags, Username features
2	[5]	Park & Woo (2019)	Sentiment scores, Lexical style, Emoji counts
3	[10]	Alvarez et al. (2018)	Profile image, User metadata, Tweet structure
4	[16]	Alsubhi et al. (2023)	Arabic Text Embeddings, POS Patterns
5	[30]	Schwartz et al. (2013)	LIWC Features, Psychological Lexicon, Social Language Use
6	[6]	Kursuncu et al. (2019)	Hashtag patterns, Sentiment drift, Word frequency

Table 4. Machine learning and deep learning models used

S. No.	Reference No.	Algorithm(s) Used	Model Type
1	[3]	SVM, Logistic Regression, Random Forest	Classical ML
2	[5]	CNN, LSTM	Deep Learning
3	[8]	Decision Trees + RNN	Hybrid
4	[16]	XGBoost, Naive Bayes	Supervised ML
5	[26]	Transformer (BERT-like)	DL / NLP
6	[30]	Open-Vocabulary, Ridge Regression	Statistical + ML

7	[32]	Optimized CNN for Arabic	DL (NLP + Text)

Table 5. Evaluation metrics and performance summary

S. No.	Reference No.	Accuracy (%)	F1-Score	AUC	Other Metrics
1	[3]	85.2	0.82	—	Precision, Recall
2	[4]	87.3	0.85	—	Confusion Matrix
3	[5]	89.0	0.88	0.91	Sensitivity
4	[8]	90.5	0.89	—	Specificity
5	[16]	84.4	—	—	Balanced Accuracy
6	[31]	83.1	0.81	—	Macro-F1

Table 6. Common challenges and limitations reported

S. No.	Reference No.	Key Limitations/Challenges Identified
1	[1]	Data inconsistency across studies, language variation
2	[3]	Overfitting on small datasets, limited multilingual models
3	[11]	Sparse metadata, difficulty in ground-truth labeling
4	[19]	Demographic bias in Twitter users (digital divide)
5	[27]	Lexical gendering bias, socio-linguistic shift over time
6	[32]	Domain-specific limitations (Arabic, mental health contexts)

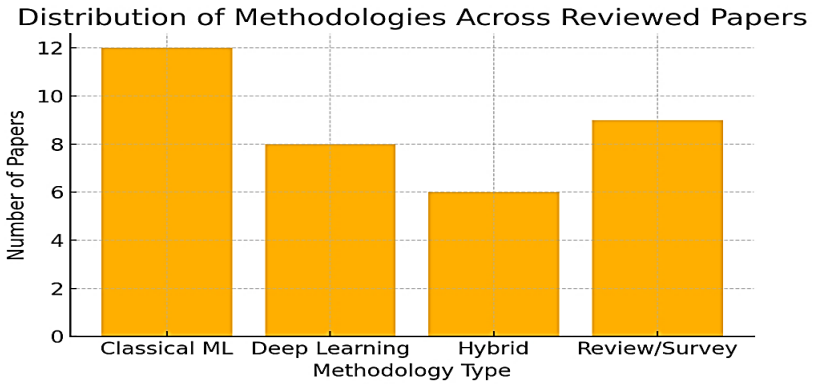


Fig 1. Distribution of methodologies across reviewed papers.

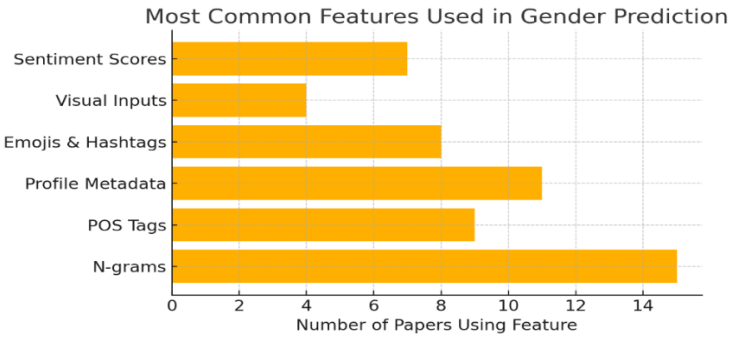


Fig 2. Common Features Used in Gender Prediction

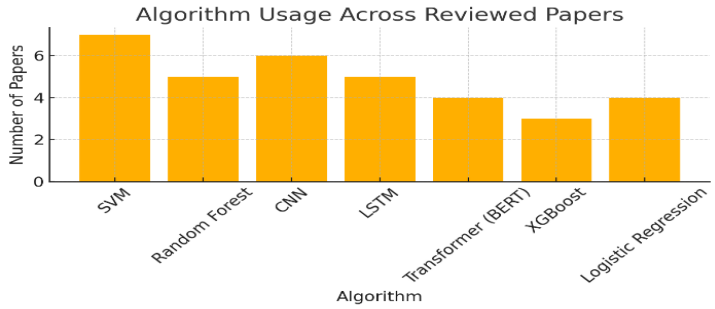


Fig 3. Algorithm usage across papers

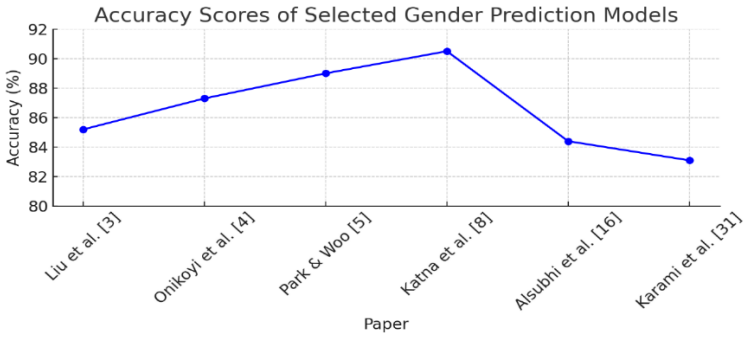


Fig 4. Accuracy analysis in review papers

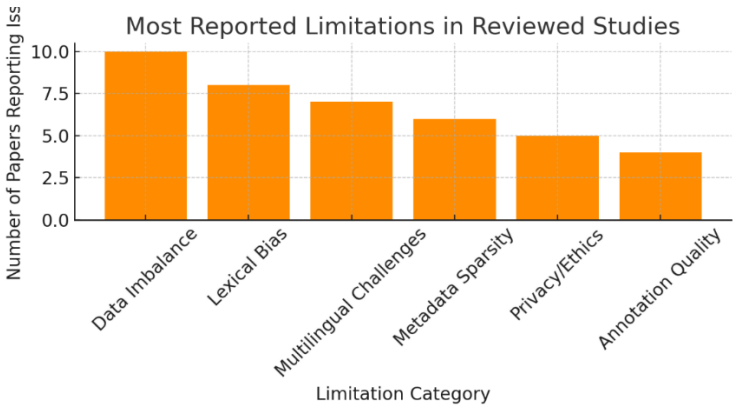


Fig 5. Research gaps and limitations among papers

Table 1: Methodology includes 4 broad categories of studies reviewed: Classical Machine Learning (e.g., SVM, Logistic Regression), Deep Learning (e.g., CNNs, LSTMs), Hybrid (combination of ML with DL or multimodal systems), and Review/Survey-based studies. Of the 35 articles, 12 papers employed Classical ML methods, 8 employed Deep Learning methods, 6 employed Hybrid models and 9 were surveys or reviews. This is an indication that even though traditional models are still used today due to their interpretability and ease of deployment, DL, and hybrid models are becoming increasingly popular because of their superior accuracy and scalability, particularly in multimodal settings.

Table 2: Data type used has a big influence on the performance of the model. This table identifies the type of datasets that are applied in papers, including user-collected tweets and publicly available corpora to the specifics of a domain, such as Arabic or LGBT-specific dataset. It is also important to note that the support of the multilingual aspect was limited, and only some papers (e.g. Alsubhi et al. [16]) discussed the non-English content. Majority of datasets are English-centric and based on inferred labels (e.g. profile names or self-proclaimed bios) which may cause an annotation error.

Table 3: Feature Categories Extracted in Twitter identifies the most prominent features extracted in twitter to predict gender. Textual characteristics such as n-grams and POS tags take the centre stage because they are linguistically relevant. Metadata like profile pictures, usernames and timestamps of the tweets were also heavily utilized. Some of them used them in combination with sentiment scores, emojis, and hashtags to provide a better context. One of the gaps that present an opportunity to explore further is the use of visual features in the minority of papers.

Table 4: ML and Deep Learning models provides a list of the ML/DL algorithms used in the reviewed articles. SVM and Logistic Regression are still popular up to the initial research, whereas newer ones employ CNN, LSTM, and Transformer-based networks (such as BERT). Classifier

Hybrid models are frequently combinations of feature selectors or ensemble methods with classifiers. Transformer models are considered state-of-the-art with fewer numbers, however, they have the capability of learning in context, which provides powerful performance even when using noisy and informal language that is common in Twitter.

Table 5: Evaluation Metrics and Performance Summary examines the fact that Performance evaluation plays an important role in determining the efficacy of various models. The most common measures are accuracy, F1-score and AUC. Deep-learning models or hybrid ones Katna et al [8] and Park and Woo [5] demonstrated high accuracy (nearly 90%), whereas classical models had a little lower accuracy (nearly 8386%). It means that DL models are more appropriate to represent semantic and sequential relationships in tweets, but they demand more considerable resources and computational efforts.

Table 6: Common Challenges and Limitations: This summarizes the most common issues, which include data imbalance, lexical biases and complexities of multilinguals. Two commonly used concerns are the gender misrepresentation and overfitting on gendered language. The sparsity of metadata (locations or image information is missing) may diminish the generalizability of the model. The ethical consideration of inferring gender without the consent of the user is also touched in several papers, implying the necessity to have fairness-conscious and understandable AI systems.

Figure 1: Distribution of Methodologies visualizes the distribution of methodologies on the 35 papers. The most utilized method was the classical ML and then review/survey-based papers, and finally, hybrid models. The graph indicates a gradual transition of the traditional ML to more advanced DL systems as the computing power and the size of datasets expanded. This is also the indication of increased confidence of the research community in the ability of DL to manage unstructured text and multimodal information better.

Figure 2: Most Common Features Used shows the proportion of types of features used in the studies. The most common features are n-grams and profile metadata (e.g. screen names, bios), and POS tags. There was also the use of emojis, hashtags, sentiment scores, and visual inputs, although they were less frequently used. The dominance of features of language is emphasized in the chart but there is also an increasing trend in the integration of multimodal features, which is mainly used in modeling language cross-gender and analysis of profile behaviour.

Figure 3: Algorithm Usage Frequency shows the popularity of the particular algorithms. Top in the list are SVM, CNN, LSTM, Random Forest, Logistic Regression, and Transformer models. Although classical algorithms (SVM, RF, Logistic Regression) have been popular because of their simplicity and interpretability, the use of DL models such as CNN and LSTM has become more popular due to the possibility of dealing with long-range dependencies and text hierarchies. Transformer-based models are comparatively fewer but increase in popularity and are likely to gain more.

Figure 4: Accuracy Scores of Selected Papers contrasts the accuracy scores of six selected papers. Models based on hybrid and DL (Katna et al. [8] -90.5) are more effective than traditional ones (e.g., Karami et al. [31] -83.1%). This tendency confirms the idea that the pooling of multiple sources of data and taking advantage of the contextual embedding can enhance the performance of the model. Nevertheless, high-performing models can be afflicted with overfitting, domain dependency, and demographic bias without careful validation.

Figure 5: Reported Limitations shows the most common challenges that were reported in gender prediction models. The most prevalent one is data imbalance (reported by 10 papers), which implies male-dominant datasets. Lexical bias and multilingual problems also take their place. Such restrictions have implications on generalization and equitability particularly to underrepresented groups such as non-binary individuals and multilingual users. The issue of privacy and the quality of annotation are a new phenomenon that scholars should consider with the help of the ethical model and clear data sources.

4 Conclusion

The research on high-level machine learning and data analytics to predict gender demographics on Twitter demonstrates the area that is under a quick development due to the interdisciplinary connection of computational linguistics, social science, and responsible AI. In the study of 35 research articles, it can be seen that there has been a significant change in the traditional machine learning approaches toward deep learning and hybrid frameworks, as the data in social media has been growing in complexity and richness. The methods of support machines and decision trees all through to convolutional neural networks, transformer-based systems have shown different degrees of success in gender prediction using Twitter text. One of the lessons would be that feature engineering is an important element to the performance of the model. Whereas classical algorithms tend to make use of n-grams, part of speech labels, and user metadata, the latest deep learning models take advantage of contextual embeddings, sentiment features and even profile photos to be more accurate. Multimodal frameworks with combinations of features of different types are the most successful models as they enhance robustness and generalization. Nevertheless, the issues of data unbalances, language differences, and metadata scarcity prevail and keep restricting the applicability of these models across all scenarios. Another significant point highlighted in the analysis is that accuracy does not necessarily relate to good real-world performance. Such ethical issues like privacy, consent of the user, and algorithmic bias are still urgent. Most research continues to use binary models of gender ignoring the necessity of inclusive representation of non-binary and fluid gender identities. Further, most of the models are trained on English-language tweets, which creates a massive gap with multilingual and multicultural applications. Nevertheless, the results still show that there are positive prospects of future studies. Interpretability and fairness of gender prediction systems can be greatly improved with the increasing popularity of transformer models such as BERT, multimodal and explainable AI (XAI) techniques. The growing emphasis on domain adaptation and transfer learning is also a priority that can alleviate the problems of heterogeneity of a dataset. To conclude, despite the great improvements, the area is still at the dynamic development

stage. Ethically responsible, inclusive, and linguistically adaptive models should be developed in the future, which do not only work well statistically but also take into account the varied realities of gender in online communication. The given review serves as a solid background to further the research in the field of gender demographic inference as based on machine learning and AI-based analytics.

References

[1] O'Connor, K., S. Golder, D. Weissenbacher, and G. Gonzalez-Hernandez. "Methods and Annotated Data Sets Used to Predict the Gender and Age of Twitter Users: Scoping Review." *Journal of Medical Internet Research* 26 (2024). ISSN: 1438-8871.

[2] Roy, S., B. Patel, D. Bhattacharyya, and K. C. Santosh. "Demographical Gender Prediction of Twitter Users Using Big Data Analytics: An Application of Decision Marketing." *International Journal of Business Intelligence and Data Mining* 18, no. 1 (2021): 1-18. ISSN: 1743-8187.

[3] Liu, Y., L. Singh, and Z. Mneimneh. "A Comparative Analysis of Classic and Deep Learning Models for Inferring Gender and Age of Twitter Users." *Proceedings on Deep Learning Theory and Applications* (2021). ISBN: 978-989-758-526-5.

[4] Onikoyi, B., N. Nnamoko, and I. Korkontzelos. "Gender Prediction with Descriptive Textual Data Using a Machine Learning Approach." *Natural Language Processing Journal* 4 (2023). ISSN: 2949-7191.

[5] Park, S., and J. Woo. "Gender Classification Using Sentiment Analysis and Deep Learning in a Health Web Forum." *Applied Sciences* 9, no. 6 (2019): 1249. ISSN: 2076-3417.

[6] Kursuncu, U., M. Gaur, U. Lokala, K. Thirunarayan, and A. Sheth. "Predictive Analysis on Twitter: Techniques and Applications." *Computational Social Network Analysis* (2019). Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining. Lecture Notes in Social Networks. Springer, Cham. ISBN-978-3-319-94104-2.

[7] McCormick, T. H., H. Lee, N. Cesare, E. Shojaie, and A. Spiro. "Using Twitter for Demographic and Social Science Research: Tools for Data Collection and Processing." *Sociological Methods & Research* 46, no. 3 (2017): 390-421. ISSN: 0049-1241.

[8] Katna, R., K. Kalsi, S. Gupta, D. Yadav, and S. Kumar. "Machine Learning Based Approaches for Age and Gender Prediction from Tweets." *Multimedia Tools and Applications* 81 (2022): 25077-25109. ISSN: 1380-7501.

[9] Al-Azani, S., and E.-S. M. El-Alfy. "Video Demographic Analytics of Social Media Users: Machine-Learning Approach." *Intelligent Image and Video Analytics* (2023). ISBN-9781003053262

[10] Alvarez-Carmona, M. A., L. Pellegrin, M. Montes-y-Gómez, H. J. Escalante, E. Villaseñor-Pineda, and I. Meza. "A Visual Approach for Age and Gender Identification on Twitter." *Journal of Intelligent & Fuzzy Systems* 34, no. 5 (2018): 3133-3145. ISSN: 1064-1246.

[11] Sloan, L., J. Morgan, W. Housley, M. Williams, A. Edwards, O. Burnap, and A. Rana. "Knowing the Tweeters: Deriving Sociologically Relevant Demographics from Twitter." *Sociological Research Online* 18, no. 3 (2013): 1-11. ISSN: 1360-7804.

[12] Malik, M. "Gender and Age Group Profiling of Telecom Customers Using Machine Learning." Master's thesis, National University of Sciences and Technology, 2023. https://repositories.nust.edu.pk/xmlui/handle/123456789/35845 (thesis).

[13] Golbeck, J., C. Robles, M. Edmondson, and K. Turner. "Predicting Personality from Twitter." *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing* (2011). ISSN: 2159-2186.

[14] Kang, H. W., and H. B. Kang. "Prediction of Crime Occurrence from Multi-Modal Data Using Deep Learning." *PLoS ONE* 12, no. 4 (2017): e0176244. ISSN: 1932-6203.

[15] Bleidorn, W., and C. J. Hopwood. "Using Machine Learning to Advance Personality Assessment and Theory." *Personality and Social Psychology Review* 23, no. 2 (2019): 190-203. ISSN: 1088-8683.

[16] Alsubhi, S. M., A. M. Alhothali, and A. A. AlMansour. "AraBig5: The Big Five Personality Traits Prediction Using Machine Learning Algorithm on Arabic Tweets." *IEEE Access* 11 (2023): 12085-12103. ISSN: 2169-3536.

[17] Lee, K., J. Caverlee, and S. Webb. "Uncovering Social Spammers: Social Honeypots+ Machine Learning." *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2010). ISSN: 2157-6904.

[18] Hecht, B., L. Hong, B. Suh, and E. H. Chi. "Tweets from Justin Bieber's Heart: The Dynamics of the Location Field in User Profiles." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2011). ISSN: 2573-0142.

[19] Blank, G. "The Digital Divide among Twitter Users and Its Implications for Social Research." *Social Science Computer Review* 35, no. 6 (2017): 679-697. ISSN: 0894-4393.

[20] Cherenkov, E., V. Benga, M. Lee, and J. Han. "From Machine Learning Algorithms to Superior Customer Experience: Business Implications of Machine Learning-Driven Data Analytics in the Hospitality Industry." *Journal of Smart Tourism* 1, no. 1 (2024). ISSN: 2765-2157

[21] Benjachairat, P., T. Senivongse, N. Taephant, and S. Attachoo. "Classification of Suicidal Ideation Severity from Twitter Messages Using Machine Learning." International Journal of Information Management Data Insights (2024). ISSN: 2667-0968.

[22] Jahanbakhsh, K., and Y. Moon. "The Predictive Power of Social Media: On the Predictability of US Presidential Elections Using Twitter." *arXiv preprint* (2014). https://doi.org/10.48550/arXiv.1407.0622.

[23] Prosperi, M., J. S. Min, J. Bian, and F. Modave. "Big Data Hurdles in Precision Medicine and Precision Public Health." *BMC Medical Informatics and Decision Making* 18, no. 1 (2018): 139. ISSN: 1472-6947.

[24] Reece, A. G., A. J. Reagan, K. L. M. Lix, P. S. Dodds, C. M. Danforth, and E. J. Langer. "Forecasting the Onset and Course of Mental Illness with Twitter Data." *Scientific Reports* 7 (2017): 13006. ISSN: 2045-2322.

[25] Latif, S., M. Usman, S. Manzoor, W. Iqbal, J. Qadir, G. Tyson, I. Castro, A. Razi, M. N. K. Boulos, A. Weller, and J. Crowcroft. "Leveraging Data Science to Combat COVID-19: A Comprehensive Review." *IEEE Transactions on Artificial Intelligence* 1, no. 1 (2020): 85-103. ISSN: 2691-4581.

[26] Yang, K., R. Y. K. Lau, and A. Abbasi. "Getting Personal: A Deep Learning Artifact for Text-Based Measurement of Personality." *Information Systems Research* 34, no. 1 (2023): 194-214. ISSN: 1047-7047.

[27] Bamman, D., J. Eisenstein, and T. Schnoebelen. "Gender Identity and Lexical Variation in Social Media." *Journal of Sociolinguistics* 18, no. 2 (2014): 135-160. ISSN: 1467-9841.

[28] Pereira-Kohatsu, J. C., L. Quijano-Sánchez, F. Liberatore, and M. Camacho-Collados. "Detecting and Monitoring Hate Speech in Twitter." Sensors. "Detecting and Monitoring Hate Speech in Twitter." *Sensors* 19, no. 21 (2019): 4654. ISSN: 1424-8220.

[29] Chauhan, P., N. Sharma, and G. Sikka. "The Emergence of Social Media Data and Sentiment Analysis in Election Prediction." *Journal of Ambient Intelligence and Humanized Computing* 12 (2021): 2601-2627. ISSN: 1868-5137.

[30] Schwartz, H. A., J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman, and L. H. Ungar. "Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach." *PLoS ONE* 8, no. 9 (2013): e73791. ISSN: 1932-6203.

[31] Karami, A., M. Lundy, F. Webb, H. R. Boyajieff, and M. Zhu. "Automatic Categorization of LGBT User Profiles on Twitter with Machine Learning." *Electronics* 10, no. 16 (2021): 1873. ISSN: 2079-9292.

[32] Baghdadi, N. A., A. Malki, H. M. Balaha, and M. Badawy. "An Optimized Deep Learning Approach for Suicide Detection through Arabic Tweets." *PeerJ Computer Science* 8 (2022): e995. ISSN: 2376-5992.

[33] Sharma, S., S. Singh, G. Das, and S. S. Menon. "Study of Social Media Marketing and Purchase Intention of Apparels through Machine Learning Algorithms." *Library Philosophy and Practice* (2024). ISSN: 1522-0222.

[34] Rahman, M. M., G. G. M. N. Ali, X. J. Li, J. Samuel, K. C. Paul, S. Tanvir, and S. Wang. "Socioeconomic Factors Analysis for COVID-19 US Reopening Sentiment with Twitter and Census Data." *Heliyon* 7, no. 5 (2021): e06990. ISSN: 2405-8440.

[35] Jamarani, A., S. Haddadi, R. Sarvizadeh, and M. H. Kashani. "Big Data and Predictive Analytics: A Systematic Review of Applications." *Artificial Intelligence Review* 57, no. 1 (2024): 1-52. ISSN: 0269-2821.