SocialDemoExtract: A Tool for Extracting Self-Reported Age and Gender from Social Media Text

Salim Sazzed*, Md Ehashan Rabbi Pial and Farhan Noor Dehan
Department of Computer Science, Georgia Southern University, Statesboro, GA 30460, USA
*@georgiasouthern.edu)

***Abstract*—The increasing prevalence of online health discussions provides valuable insights into how individuals from diverse backgrounds communicate about health experiences. However, manually extracting demographic attributes such as age and gender—key factors for demography-specific analyses—is highly resource-intensive, limiting researchers' ability to conduct population-specific studies and develop inclusive digital health interventions. This study introduces SocialDemoExtract, an open-source tool designed to automatically extract self-reported age and gender information from health-related social media text. The tool integrates natural language processing, pattern recognition, and rule-based methods to accurately identify demographic information in informal online text. SocialDemoExtract demonstrates high efficacy across three mental health datasets, with precision in age extraction ranging from 92% to 97% and precision in gender identification ranging from 88% to 95%. By enabling demography-aware analyses, this tool facilitates a deeper understanding of patterns of health issues across diverse populations and supports the development of data-driven health systems.**

I. Introduction

The growing volume of health-related discussions on various social media platforms presents a rich source of data for understanding how people share experiences, seek support, and express concerns about their physical and mental health issues [1], [2], [3], [4], [5], [6], [7]. These discussions often reveal nuanced, real-world perspectives and self-reported experiences that are not typically available in clinical data [8], [9]. However, extracting demographic information such as age and gender from large-scale, unstructured social media data is challenging [10], [11], [12], as these attributes are often ambiguously expressed and lack standardized formats [13]. Extracting such information manually is both time-consuming and highly resource-intensive, making it impractical for large corpora. As a result, researchers are often unable to fully utilize demographic information when analyzing health-related social media text, which limits the depth and inclusiveness of their findings [14].

Although the need to incorporate demographic information when analyzing social media text continues to grow [15], [16], particularly in computational social science and digital health research, publicly available and reliable automated tools capable of accurately extracting such information remain scarce [17]. Although some studies have explored machine learning approaches to infer age and gender, these methods typically focus on binary gender classification, age group categorization, or detecting the presence of age information, rather than extracting precise age values [18], [19], [5], [20], [21], [22], [23], [24]. The lack of robust automated methods limits researchers' ability to perform demography-aware analyses [25] and hampers progress toward the development of data-driven digital health interventions across diverse populations.

To address these challenges, this study introduces *SocialDemoExtract*, an open-source tool designed to automatically extract self-reported age and gender information from social media text. *SocialDemoExtract* integrates natural language processing, pattern recognition, and rule-based techniques to identify explicit demographic references within user-generated content. We applied *SocialDemoExtract* to three mental health–related datasets. The tool successfully extracted age information with precision ranging from 92% to 98% and accurately identified gender with precision between 88% and 95%. By enabling scalable, demography-aware analyses, *SocialDemoExtract* allows researchers to examine how health experiences vary across age and gender groups, facilitating the development of inclusive, evidence-based digital health interventions that address population-specific needs.

The key contributions of this work are as follows:

- **Novel demographic extraction tool:** *SocialDemoExtract* is the first open-source tool designed to automatically extract self-reported age and gender information from informal, health-related social media posts by combining curated rules with advanced NLP techniques. It will be made publicly available on GitHub to facilitate broad adoption [1].
- **High precision and scalability:** The tool demonstrates high precision, supporting large-scale demographic analyses without requiring resource-intensive manual review.
- **Enabling demography-aware research:** By providing reliable age and gender information, *SocialDemoExtract* allows researchers to perform stratified analyses of online health discourse across diverse populations.

[1]https://github.com/sazzadcsedu/SocialDemopExtract.git

II. Dataset

This study analyzes three datasets representing different forms of psychological distress sourced from Reddit, a popular social media platform composed of topic-specific forums where users share posts and participate in discussions. Unlike platforms such as Twitter, Reddit's longer post format allows users to share more detailed personal narratives and often includes self-reported demographic information such as age and gender identity [10]. These characteristics make Reddit particularly suitable for evaluating *SocialDemoExtract* in identifying demographic references within informal, health-related social media text. All datasets include only posts authored by individuals, excluding any responses from other people.

a) Suicidal Ideation dataset: The Suicidal Ideation dataset contains 27,433 mental health–related posts from the Reddit forum *r/suicide idea ideaWatch*, which represent individuals' thoughts, experiences, and underlying factors associated with suicidality. These 27,433 posts represent user-generated textual content from December 16, 2008, to January 2, 2021, spanning a broad temporal range. Posts have a mean length of 458.53 words (SD = 349.87), with considerable variability ranging from 90 to 5,973 words and a median of 348 words. In total, the dataset comprises 12,578,793 words.

b) Anxiety dataset: The anxiety dataset contains 30,000 posts from the Reddit forum *r/Anxiety*, representing discussions by individuals experiencing anxiety. Posts have a mean length of 157.16 words (SD = 163.30), ranging from 1 to 4,358 words, with a median of 112 words. In total, the dataset comprises 4,714,779 words.

c) Depression dataset: The depression dataset contains 30,000 posts from the Reddit forum *r/Depression*, representing peer support for individuals experiencing depressive disorders. Posts have a mean length of 172.71 words (SD = 187.65), ranging from 1 to 4,161 words, with a median of 120 words. In total, the dataset comprises 5,181,453 words.

III. Methods

This section describes the methodological details of *SocialDemoExtract* for extracting self-reported age and gender from mental health–related social media posts. *SocialDemoExtract* employs a combination of rule-based and pattern-driven techniques that focus on explicit, self-referential mentions.

A. *Age Extraction*

Extracting age information from user-generated social media posts poses significant challenges due to the unstructured, informal, and linguistically diverse nature of user-generated content. Social media posts lack standardized formatting, resulting in age being expressed in a wide variety of forms, such as *'I'm 30'*, *'thirty years old'*, or *'turned twenty-two last month.'* Numeric values may also refer to unrelated concepts, including weight, height, or temporal durations, which increases the risk of false positives (e.g., *'I am 200 pounds of nothing'* or *'I was 23 years old when I lost my best friend to suicide idea'*).

SocialDemoExtract is designed to identify and interpret age references expressed in both textual and numeric formats. The algorithm leverages contextual information to distinguish the author's current age from references to others or past events, ensuring robust performance across diverse posts. Irrelevant numeric data are systematically filtered, while age-related expressions are captured through a multi-stage framework that integrates rule-based methods, pattern recognition, and contextual phrase detection.

Key components of the age extraction module include:

a) Word-to-number conversion: A predefined dictionary maps textual numbers to their corresponding integers (e.g., "twenty-five" $\rightarrow$ 25), supporting both single-word forms (e.g., "five") and multi-word compounds (e.g., "twenty-one").

b) Pattern recognition: Regular expressions identify candidate numeric values while minimizing irrelevant numbers. For instance, the pattern `\b{1,3}\b` matches standalone numbers of 1–3 digits (e.g., 9, 25) for further evaluation.

c) Contextual phrase detection: The system captures age-related linguistic patterns surrounding numbers, such as "I'm [number]", "[number] years old", or "age [number]", to improve extraction precision.

d) Exclusion Patterns: The algorithm employs specific exclusion patterns to filter out numeric data that are irrelevant to age identification. These patterns are designed to match common formats representing temporal references, weight, and height, ensuring that such data do not interfere with accurate age extraction. Key exclusion patterns include the following (non-exhaustive list):

- Temporal references: Matches phrases representing time intervals, such as `\b\d+\s*(minutes?|hours?|months?|weeks?)\b`. This pattern captures numeric values associated with temporal units, e.g., "45 minutes," "3 hours," or "2 months," and excludes them as they do not pertain to age.
- Weight data: Filters out expressions related to body weight, represented by patterns such as `\b\d+\s?lbs\b`. This ensures that references to weight, such as "150 lbs" or "70 lbs," are excluded from the analysis.
- **Height representations:** Identifies and excludes common height formats, such as `\b\d{1,2}\'\d{1,2}\b`. For example, "`5'8''`" or "`6'2''`" would be matched and filtered out, as they are unrelated to age.

e) Subject Identification: The algorithm incorporates subject identification to exclude ages mentioned for individuals other than the narrator. For example, in the sentence *"I'm a 22-year-old male in a long-distance relationship with my*

20-year-old girlfriend", the algorithm correctly identifies the narrator's age as 22 while excluding the age of the other individual.

f) Tense Identification:: In addition to identifying the subject, the algorithm evaluates the tense of each statement to ensure that the extracted age corresponds to the narrator's current age. For example, the sentence *"I was 15 years old when I had my first suicidal thought"* is excluded because it refers to a past event rather than the individual's present age. This filtering mechanism ensures that the extracted data accurately reflect the narrator's current demographic information, enhancing the algorithm's reliability and relevance for its intended applications.

B. *Gender Identification*

Similar to age extraction, identifying gender identity from social media text is challenging due to informal writing styles—including abbreviations, lack of standard formatting, and substantial linguistic variability. *SocialDemoExtract* employs a rule-based, pronoun- and keyword-driven approach that focuses on explicit, self-referential gender mentions.

The gender identification module relies on curated lexicons containing gender-indicative terms grouped into three categories. These lexicons were constructed through a multiphase process that involved: (1) manually reviewing more than 1,000 mental health–related Reddit posts to identify naturally occurring self-reported gender cues, such as explicit gender self-disclosures (e.g., "I'm a man," "I'm a woman," "as a male," "as a female"); (2) generating additional candidate terms using a large language model (LLM) to capture broader linguistic variation, including informal expressions commonly used on social media; and (3) validating the final lexicon set with domain experts familiar with social media communication patterns and demographic self-disclosures. Some gender indicator words in the lexicon are provided below:

1) **Male indicators** (e.g., 'male', 'man', 'boy', 'm')
2) **Female indicators** (e.g., 'female', 'woman', 'girl', 'f')
3) **Non-binary indicators** (e.g., 'non-binary', 'trans')

To guarantee that identified gender labels accurately reflect the author's self-identification, our algorithm requires gender indicators to co-occur with unambiguous first-person references (e.g., 'I','me', 'my', 'as a [woman/man]'). In cases of multiple or conflicting indicators, contextual rules prioritize the most self-referential phrase.

IV. Results and findings

This section discusses the results of applying *SocialDemoExtract* to extract self-reported age and gender information from three mental health–related social media datasets (Table I).

A. Age Extraction

We applied *SocialDemoExtract* to our three datasets to automatically identify individuals' ages.

a) Suicidal Ideation dataset: SocialDemoExtract successfully extracted age information from 6,148 posts, representing approximately 22.4% of the total dataset. We assessed the reliability of the extraction through comprehensive manual validation of identified ages. Among the 6,148 extracted ages, 164 were determined to be incorrect, resulting in a precision of 97.33%. Most errors involved numeric values that were contextually ambiguous or unconventional phrasing that led to misidentification, highlighting the subtle challenges of extracting accurate information from informal social media text. Examples of extracted age information, along with the corresponding source text, are presented in Table II.

In addition to validating the successfully extracted ages, we conducted a focused analysis of posts in which *SocialDemoExtract* did not detect any age references. A random sample of 500 such posts was manually reviewed, revealing 86 instances in which age information was present but not captured by the tool. These missed cases were primarily due to typographical errors, unconventional or creative expressions of age, or indirect references that did not conform to the patterns encoded in the extraction rules. For example, expressions such as "i just hit my mid-twenties" posed significant challenges for automated recognition.

b) Anxiety: In the Anxiety dataset, *SocialDemoExtract* identified age information in 1,423 posts, representing approximately 4.74% of the posts in the dataset. To assess the accuracy of these extractions, we conducted a comprehensive manual validation. Of the 1,423 extracted ages, 1,315 were correctly identified, resulting in a precision of 92.41%.

Similar to the Suicidal Ideation dataset, we evaluated the false negative rate for age extraction in the Anxiety dataset—instances where *SocialDemoExtract* failed to detect age references explicitly stated by the author—by manually inspecting a random sample of 253 posts. Manual analysis identified 13 false negatives, corresponding to a rate of 5.1%, highlighting cases in which valid age references were not captured.

c) Depression: In the Depression dataset, *SocialDemoExtract* detected age information in 2,224 posts, corresponding to approximately 7.41% of the posts in the dataset. To assess the accuracy of these extractions, we conducted a thorough manual validation. Of the 2,224 extracted age mentions, 2,140 were correctly identified, resulting in a precision of 96.22%.

For false negative analysis, among 256 randomly selected posts in which *SocialDemoExtract* predicted no age information, manual inspection revealed that age information was actually present in 9 posts ($\approx 3.5\%$), indicating instances of false negatives where the tool failed to detect the reported age information.

Taken together in Table I, these results demonstrate the overall robustness and high precision of *SocialDemoExtract* in extracting explicitly self-reported age from informal social media content. The residual errors are attributable to specific challenges in identifying clear self-disclosure that extend beyond general linguistic variability. These challenges

TABLE I
PRECISION OF AGE AND GENDER EXTRACTION ACROSS THREE MENTAL HEALTH–RELATED DATASETS USING *SocialDemoExtract*

Dataset (#Num. of Post)	**Age**			**Gender**		
	#Identified	#Correct	Precision	#Identified	#Correct	Precision
Suicidal Ideation (27433)	6,148	5984	97.33%	2,620	2451	93.60%
Anxiety (30000)	1,423	1315	92.41%	787	748	95.05%
Depression (30000)	2,224	2140	96.22%	1,043	914	87.63%

include indirect self-reporting, where age is implied rather than explicitly stated; complex or noisy syntax, such as self-reports embedded in misspelled or fragmented sentences (e.g., 'im 5teen'); and ambiguous self-reference, where age is presented non-literally or with unclear attribution. Overall, these findings underscore the tool's utility for large-scale analyses while highlighting that capturing the full spectrum of self-disclosure, particularly its indirect and non-standard forms, remains an area for continued refinement. Some examples of successfully extracted ages are shown in Table II.

TABLE II
SAMPLE EXCERPTS DEMONSTRATING SUCCESSFUL **AGE** EXTRACTION IN MENTAL HEALTH–RELATED POSTS

Text Excerpts	**Extracted Age**
...im an 18 year old black female ...	18
...16 yrs old suicide idea is a very real option right now i am 16 years old as the title suggests ...	16
...i'd rather die i'm a 16 year old female ...	16
...im already a college dropout on the way to 25 years old unemployed at this point ...	25
...i am a 40 year old male ...	40

B. Gender Extraction

We evaluated *SocialDemoExtract* on the three datasets for automatic extraction of gender information.

a) Suicidal Ideation dataset: Using curated lexicons of gender-indicative terms and a rule- and pattern-based approach (described in the Method section), *SocialDemoExtract* detected gender indications in 2,620 posts, representing approximately 9.55% of the dataset. To assess precision, we manually validated the extracted ages, among which 2,451 were correctly classified, resulting in a precision of 93.6%.

To assess false negatives, we examined a random subset of 250 posts in which the tool did not identify any gender information. Manual review showed that 27 of these posts ($\approx 10.8\%$) did include clear gender cues that the tool failed to capture.

b) Anxiety dataset: In the Anxiety dataset, *SocialDemoExtract* extracted gender information from 787 posts, representing approximately 9.55% of the dataset. Manual validation confirmed that 748 cases were correctly identified, resulting in a precision of 95.05%.

For false negative analysis, we randomly selected 250 posts in which the tool failed to detect gender information. Manual inspection revealed that the tool missed explicitly mentioned gender information in 17 posts ($\approx 6.8\%$), representing instances where gender was present but not identified by the tool.

c) Depression dataset: SocialDemoExtract successfully detected gender indications in 1,043 posts, representing approximately 9.55% of the posts in the dataset. To assess precision, we manually validated these posts and found that 914 were correctly classified, yielding a precision of 87.63%.

For false negative analysis, among 250 randomly selected posts, the tool failed to detect gender in 25 posts ($\approx 10\%$), representing instances where gender information was present but not identified by the tool.

TABLE III
SAMPLE EXCERPTS ILLUSTRATING SUCCESSFUL **GENDER** EXTRACTION IN MENTAL HEALTH–RELATED POSTS

Text Excerpt	**Extracted Gender**
...i'm a 22 year old female who has been blessed with good opportunities in life ...	Female
...i am a 16 male who is currently in his 1st year of college ...	Male
...i'm the god damn odd man out who didn't spend his high school years looking for nothing ...	Male
...im too scared to do it myself i hate being bisexual in a homophobic family ...	Non-binary

Table III presents examples of extracted gender information, along with the corresponding text snippets from which they were derived. Most errors occurred in posts with indirect or contextually ambiguous gender references. Non-standard phrasing or community-specific terminology (e.g., *'as a guy who's struggling'*) occasionally led to misclassification, as did posts containing multiple gender mentions or quoted speech from others, despite the algorithm's focus on first-person self-references. Overall, these results demonstrate that *SocialDemoExtract* effectively extracts gender information from informal social media text, although linguistic variability and unconventional phrasing can occasionally reduce its accuracy. The tool's rule- and pattern-based design ensures transparency and interpretability, which is particularly valuable for research on online health discussions, enabling systematic analysis of gender-specific patterns while maintaining clarity about the basis for each classification. These findings underscore both the strengths and limitations of automated demographic extraction and point to opportunities for future refinement,

such as expanding lexicons or integrating hybrid methods that incorporate machine learning to better capture less explicit gender expressions.

V. Summary and Conclusion

This study presents *SocialDemoExtract*, a tool for the automatic extraction of self-reported age and gender information from informal social media posts, with a particular focus on mental health–related content. Social media text is inherently unstructured, informal, and linguistically diverse, which poses significant challenges for demographic information extraction. *SocialDemoExtract* addresses these challenges through a combination of rule-based methods, pattern recognition, and contextual analysis, enabling robust and context-aware identification of self-reported demographic information.

We applied the tool to approximately 87,000 posts from three datasets. Age information was successfully extracted with a precision of 92%–97%, while gender information was identified with a precision of 88%–95%. The false negative rate for age and gender extraction across the three datasets typically ranges from 3% to 15%, based on manual review of a small subset of posts for each dataset.

Manual validation revealed that most errors stemmed from indirect or contextually ambiguous references, typographical errors, or unconventional phrasing, underscoring the linguistic variability of user-generated social media content. These results demonstrate the effectiveness and reliability of *SocialDemoExtract* for large-scale, demography-aware analyses of online health discussions. The tool enables researchers to explore population-specific patterns in mental health discourse, supporting more inclusive and contextually informed studies.

Future work will focus on extending coverage to less explicit demographic expressions and developing hybrid frameworks that integrate rule-based and machine learning techniques to further improve accuracy, generalizability, and adaptability across diverse online communities. Additionally, we plan to utilize additional datasets to rigorously evaluate the tool's performance and robustness.

References

[1] J. A. Naslund, K. A. Aschbrenner, L. A. Marsch, and S. J. Bartels, "The future of mental health care: peer-to-peer support and social media," *Epidemiology and psychiatric sciences*, vol. 25, no. 2, pp. 113–122, 2016.

[2] S. Sazzed, "Deciphering emotional and linguistic patterns in reddit suicidal discourse," in *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. Springer, 2024, pp. 133–143.

[3] B. Ridout and A. Campbell, "The use of social networking sites in mental health interventions for young people: systematic review," *Journal of medical Internet research*, vol. 20, no. 12, p. e12244, 2018.

[4] S. Sazzed, "Substance use and emotional states associated with suicidal ideation: Insights from social media posts," in *2025 19th International Conference on Semantic Computing (ICSC)*. IEEE Computer Society, 2025, pp. 247–250.

[5] J. Singh, J. Bedi, and M. Kaur, "SMM4H'24 Task6: Extracting Self-Reported Age with LLM and BERTweet: Fine-Grained Approaches for Social Media Text," in *Proceedings of The 9th Social Media Mining for Health Research and Applications (SMM4H 2024) Workshop and Shared Tasks*, 2024, pp. 106–109.

[6] S. Sazzed, "Unraveling affective responses and core determinants in health and trauma-driven suicidal narratives," in *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2024, pp. 6533–6540.

[7] M. Alvarez-Jimenez, J. Gleeson, S. Rice, C. Gonzalez-Blanch, and S. Bendall, "Online peer-to-peer support in youth mental health: seizing the opportunity," *Epidemiology and psychiatric sciences*, vol. 25, no. 2, pp. 123–126, 2016.

[8] S. Sazzed, "A Comparative Study of Affective and Linguistic Traits in Online Depression and Suicidal Discussion Forums," in *Proceedings of the 34th ACM Conference on Hypertext and Social Media*, ser. HT '23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3603163.3609059

[9] L. McDonald, B. Malcolm, S. Ramagopalan, and H. Syrad, "Real-world data and the patient perspective: the PROmise of social media?" *BMC medicine*, vol. 17, no. 1, p. 11, 2019.

[10] S. Sazzed, "Psychosocial challenges and substance use among suicidal autistic individuals on social media: LLM-assisted keyword generation with human-in-the-loop refinement," *International Journal of Medical Informatics*, p. 106110, 2025.

[11] S. Golder, R. Stevens, K. O'Connor, R. James, and G. Gonzalez-Hernandez, "Methods to establish race or ethnicity of Twitter users: scoping review," *Journal of medical Internet research*, vol. 24, no. 4, p. e35788, 2022.

[12] P. Shrestha, N. Rey-Villamizar, F. Sadeque, T. Pedersen, S. Bethard, and T. Solorio, "Age and gender prediction on health forum data," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016, pp. 3394–3401.

[13] K. O'Connor, S. Golder, D. Weissenbacher, A. Z. Klein, A. Magge, 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*, vol. 26, p. e47923, 2024.

[14] N. Cesare, C. Grant, J. B. Hawkins, J. S. Brownstein, and E. O. Nsoesie, "Demographics in social media data for public health research: does it matter?" *arXiv preprint arXiv:1710.11048*, 2017.

[15] F. Bianchi, V. Cutrona, and D. Hovy, "Twitter-demographer: A flow-based tool to enrich twitter data," *arXiv preprint arXiv:2201.10986*, 2022.

[16] G. Jagfeld, F. Lobban, P. Rayson, and S. J. Jones, "Understanding who uses reddit: Profiling individuals with a self-reported bipolar disorder diagnosis," in *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, 2021, pp. 1–14.

[17] A. Z. Klein, A. Magge, and G. Gonzalez-Hernandez, "ReportAGE: automatically extracting the exact age of Twitter users based on self-reports in tweets," *PloS one*, vol. 17, no. 1, p. e0262087, 2022.

[18] Y. Liu, L. Singh, and Z. Mneimneh, "A Comparative Analysis of Classic and Deep Learning Models for Inferring Gender and Age of Twitter Users," in *Proceedings of the 2nd International Conference on Deep Learning Theory and Applications-DeLTA,*, 2021.

[19] Z. Wang, S. Hale, D. I. Adelani, P. Grabowicz, T. Hartman, F. Flöck, and D. Jurgens, "Demographic inference and representative population estimates from multilingual social media data," in *The world wide web conference*, 2019, pp. 2056–2067.

[20] T. Sankar, D. Suraj, M. Reddy, D. Toshniwal, and A. Agarwal, "IITRoorkee@ SMM4H 2024 Cross-Platform Age Detection in Twitter and Reddit Using Transformer-Based Model," in *Proceedings of The 9th Social Media Mining for Health Research and Applications (SMM4H 2024) Workshop and Shared Tasks*, 2024, pp. 101–105.

[21] N. Cesare, C. Grant, Q. Nguyen, H. Lee, and E. O. Nsoesie, "How well can machine learning predict demographics of social media users?" *arXiv preprint arXiv:1702.01807*, 2017.

[22] R. Sadeghi, A. Akbari, and M. M. Jaziriyan, "Exaauac: Arabic twitter user age prediction corpus based on language and metadata features," *Discover Artificial Intelligence*, vol. 4, no. 1, p. 48, 2024.

[23] A. A. Morgan-Lopez, A. E. Kim, R. F. Chew, and P. Ruddle, "Predicting age groups of twitter users based on language and metadata features," *PloS one*, vol. 12, no. 8, p. e0183537, 2017.

[24] C. M. Black, W. Meng, L. Yao, and Z. B. Miled, "Inferring the patient's age from implicit age clues in health forum posts," *Journal of Biomedical Informatics*, vol. 125, p. 103976, 2022.

[25] F. Karimi, C. Wagner, F. Lemmerich, M. Jadidi, and M. Strohmaier, "Inferring gender from names on the web: A comparative evaluation of gender detection methods," in *Proceedings of the 25th International conference companion on World Wide Web*, 2016, pp. 53–54.