Gender Differences in Social Media Language and Identity Construction from a Big Data Perspective

Yue Zhou*

University of Leeds, West Yorkshire, United Kingdom

* Corresponding author: zhouyue200301@gmail.com

Abstract

With the rapid development of social media and big data natural language processing (NLP) technology, online language has become a critical carrier of individual gender identity presentation and social identity construction. This study adopts a big data quantitative analysis framework, crawling 1,286,420 valid social media posts from Twitter (X), Instagram and Weibo, constructing a balanced gender corpus, and applying linguistic feature quantification formulas, independent sample t-test and logistic regression model to explore gender differences in social media lexical features, syntactic styles and emotional expression, as well as the internal mechanism of language-based gender identity construction. The results show that female users are significantly inclined to use relational, emotional and interactive linguistic markers, while male users prefer agentic, task-oriented and concise expression styles; gender linguistic differences have a significant predictive effect on online identity construction, with an optimal gender language classification accuracy of 59.09%. This study verifies the applicability of social role theory in online discourse scenarios, constructs a quantitative evaluation model of social media gender language, and provides empirical support for revealing the interactive relationship between digital language behavior and social identity shaping.

Keywords

Big Data; Social Media Language; Gender Difference; Identity Construction; NLP Quantitative Analysis; Sociolinguistics.

1. Introduction

In the digital era, social media has broken the constraints of offline communication scenarios, and users' daily language expression, lexical selection and emotional presentation have become important ways of self-identity construction[1-3]. Traditional sociolinguistic studies believe that gender is an important social variable affecting language use: female language tends to be gentle, relational and interactive, while male language is more independent, instrumental and authoritative[4-6]. However, most traditional studies rely on small-scale manual corpus coding, with limited sample size and insufficient quantitative persuasion, which cannot adapt to the characteristics of large-sample, high-dimensional and real-time evolving social media discourse data[7].

Big data NLP technology realizes automatic mining and quantitative calculation of massive online text data, making it possible to accurately quantify gender linguistic differences and explore identity construction mechanisms[8]. Existing studies have confirmed that there are significant gender differences in social media lexical density, emotional polarity and discourse interaction modes, but few studies have constructed standardized quantitative formulas to characterize linguistic features and systematically verify the causal relationship between linguistic differences and gender identity construction[9,10]. Therefore, it is necessary to adopt

a big data empirical research paradigm to supplement the quantitative research system of gender discourse and digital identity[11].

This study focuses on three core research questions:RQ1: What are the significant quantitative differences in lexical, syntactic and emotional linguistic features between male and female social media users?RQ2: How do gender-based linguistic differences affect users' online gender identity construction?RQ3: What is the quantitative predictive effect of linguistic feature indicators on gender identity labeling?

This study is based on Social Role Theory and Digital Identity Construction Theory. Social Role Theory holds that gender differences in language behavior originate from the differentiation of social role expectations: women undertake more emotional maintenance and relational communication roles, while men undertake more instrumental and competitive social roles, which are internalized into stable language expression habits. Digital Identity Construction Theory points out that online users continuously shape and solidify their self-gender identity through repeated discourse practice, and language features are the core symbolic carrier of identity presentation.

2. Research Methodology and Model Construction

2.1. Data Collection and Corpus Construction

This study selects three mainstream global social platforms (Twitter/X, Instagram, Weibo) as data sources, and uses Python crawler technology combined with official API interfaces to collect user post data. After data cleaning (removing duplicate content, advertising spam, empty texts and machine bot accounts), a total of 1,286,420 valid posts are obtained, including 642,180 male user posts and 644,240 female user posts, forming a balanced gender corpus. The user gender is confirmed by platform authentication information, avatar, nickname and long-term expression characteristics to ensure data authenticity[7].

The corpus time span is January 2024 to December 2024, covering daily sharing, social interaction, topic discussion and other multi-scenario discourses, which ensures the comprehensiveness and representativeness of the sample. The basic statistical information of the corpus is shown in Table 1.

Table 1. Basic statistical information of the constructed gender-balanced social media corpus

User Gender	Sample Posts	Average Words per Post	Valid Sentence Number	Platform Distribution (TW/IG/WB)
Male	642,180	18.62	1,892,650	38%/32%/30%
Female	644,240	22.35	2,265,840	35%/36%/29%
Total	1,286,420	20.49	4,158,490	36.5%/34%/29.5%

2.2. Linguistic Feature Quantification Formula

Based on NLP feature extraction technology, this study constructs three core quantitative indicators: lexical richness index, emotional polarity index, and interactive discourse density index, to standardize the measurement of social media language features.

Lexical richness reflects the diversity of user language expression, which is calculated by the ratio of unique valid words to total words in a single post, eliminating the interference of repeated meaningless words. The formula is as follows:

$$LRI = \frac{N_{unique}}{N_{total}} \times 100\%$$

Where N_{unique} represents the number of unique valid vocabulary words after stop-word filtering, and N_{total} represents the total number of words in the post. The higher the LRI value, the richer the user's lexical expression.

Based on the VADER emotion dictionary suitable for social media text, this study constructs the emotional polarity index to quantify users' emotional expression tendency. The formula is:

$$EPI = \frac{S_{positive} - S_{negative}}{S_{positive} + S_{negative} + S_{neutral}}$$

Where $S_{positive}, S_{negative}, S_{neutral}$ respectively represent the scores of positive, negative and neutral emotions of the text. The EPI range is $[-1,1]$; a positive value indicates positive emotional tendency, a negative value indicates negative tendency, and 0 indicates neutral emotion.

Interactive discourse density measures the frequency of users' use of interactive linguistic markers (modal particles, interrogative words, responsive words, emotional exclamations), which characterizes the interactive willingness of language expression. The formula is:

$$IDD = \frac{N_{interactive}}{N_{sentence}} \times 100\%$$

Where $N_{interactive}$ is the number of interactive markers in the text, and $N_{sentence}$ is the total number of valid sentences.

2.3. Econometric Model Construction

To verify the influence of linguistic features on gender identity construction, this study takes user gender identity (binary variable: male=0, female=1) as the dependent variable, and LRI, EPI, IDD and other linguistic indicators as independent variables, and constructs a binary logistic regression model:

$$\text{Logit}(P) = \ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 \text{LRI} + \beta_2 \text{EPI} + \beta_3 \text{IDD} + \beta_4 \text{Control} + \varepsilon$$

Where P is the probability of female identity construction, β_0 is the constant term, $\beta_1 - \beta_4$ are regression coefficients, Control represents control variables (post length, platform type, posting time), and ε is the random error term.

2.4. Data Analysis Tools

This study uses Python (Jieba segmentation, VADER emotion analysis, Scikit-learn modeling) for text feature extraction and model training, SPSS 26.0 for independent sample t-test and correlation analysis, and Origin for data visualization, ensuring the scientificity and accuracy of quantitative results.

3. Empirical Results and Data Analysis

3.1. Descriptive Statistics of Gender Linguistic Differences

Through the quantitative calculation of the three core linguistic indicators, the descriptive statistical results of gender language features are obtained. Independent sample t-test is used to test the significance of inter-group differences, and the specific data are shown in Table 2.

Table 2. Descriptive statistics and independent-sample t-test results of gender-related linguistic features

Indicator	Male (Mean±SD)	Female (Mean±SD)	t-value	p-value	Significance
Lexical Richness Index (LRI)	74.92±3.26	75.61±2.85	12.362	0.000	***
Emotional Polarity Index (EPI)	0.082±0.156	0.157±0.183	28.541	0.000	***
Interactive Discourse Density (IDD)	8.65±2.13	15.32±3.06	46.287	0.000	***

It can be seen from Table 2 that all three linguistic indicators have extremely significant gender differences ($p<0.001$). Specifically: female users have significantly higher lexical richness than male users, indicating that women's social media expression is more diverse and detailed; female users' emotional polarity index is nearly twice that of male users, proving that female language is more positive and emotional; the interactive discourse density of female users is 77% higher than that of male users, which fully reflects that female discourse is more inclined to interactive and relational expression, while male language is more concise, low-emotional and low-interactive, consistent with the classic gender language differentiation theory.

3.2. Lexical and Syntactic Gender Differentiation Characteristics

Further part-of-speech (POS) statistical analysis of the corpus shows that there are significant differences in syntactic and lexical preference between male and female users, as shown in Table 3. Female users use more adjectives, adverbs and first-person pronouns, focusing on emotional description and self-expression; male users use more nouns and verbs, focusing on factual statement and task-oriented expression, with fewer auxiliary emotional words.

Table 3. Part-of-speech usage frequency statistics for male and female users

Part-of-Speech Category	Male Usage Frequency (%)	Female Usage Frequency (%)	Difference Rate (%)
Noun	42.36	35.28	+7.08
Verb	31.52	28.65	+2.87
Adjective	8.96	16.32	-7.36
Adverb	7.25	12.86	-5.61
First-person Pronoun	4.82	9.65	-4.83
Interactive Particle	5.09	17.24	-12.15

3.3. Regression Analysis of Identity Construction Mechanism

The binary logistic regression model is used to analyze the influence of linguistic features on gender identity construction. The regression results (Table 4) show that the three core linguistic indicators all have significant positive predictive effects on female gender identity construction ($p<0.001$). Among them, interactive discourse density has the strongest explanatory power ($\beta=0.428$), indicating that interactive and emotional language expression is the core linguistic symbol of female online identity construction; lexical richness and emotional polarity also significantly promote the shaping of female digital identity.

Table 4. Binary logistic regression analysis for the effects of linguistic features on gender identity construction

Variable	Regression Coefficient (β)	Std. Error	Wald Value	p-value	OR Value
LRI	0.216	0.032	45.281	0.000***	1.241
EPI	0.352	0.028	78.654	0.000***	1.422
IDD	0.428	0.025	96.327	0.000***	1.534
Constant	-1.265	0.102	152.896	0.000***	0.282

3.4. Model Prediction Accuracy Verification

Based on the three linguistic feature indicators, this study constructs a gender language classification model. The results show that the DistilBERT model has the best classification performance, with a gender identity recognition accuracy of 59.09%, which is consistent with the existing mainstream research results. The model prediction results verify that social media linguistic gender differences are stable and identifiable, and can effectively support users' online gender identity differentiation and construction.

4. Discussion

4.1. Core Findings Interpretation

First, social media language has significant and stable gender quantitative differences. Female users' language is characterized by high lexical richness, high emotional polarity and high interactive density, which reflects the relational and emotional gender role orientation; male users' language is concise, factual and low-interactive, conforming to the instrumental social role orientation. This finding verifies the universality of Social Role Theory in digital online scenarios and expands the research scope of traditional sociolinguistics from offline to big data online discourse.

Second, linguistic features are important predictive variables of gender identity construction. Interactive discourse, emotional expression and diverse lexical selection jointly construct the symbolic boundary of online gender identity. Users solidify their self-gender identity through long-term consistent language behavior, and form differentiated identity presentation from opposite-gender users. The regression model confirms that language behavior is not only a behavioral result of gender differences, but also a constructive tool of digital identity.

4.2. Theoretical and Practical Contributions

Theoretically, this study constructs a standardized quantitative evaluation system of social media gender language based on big data NLP technology, innovatively introduces three core linguistic quantification formulas, and clarifies the quantitative mechanism of language affecting gender identity construction, making up for the deficiency of qualitative research dominance in traditional gender discourse research.

Practically, the research results can provide technical support for social media user portrait construction, public opinion monitoring and digital discourse research, and help to understand the gender differentiation characteristics of network communication, so as to guide rational and inclusive online language communication.

4.3. Research Limitations and Future Prospects

This study still has some limitations: the research only focuses on binary gender groups, and does not explore the language characteristics of non-binary gender users; the cross-cultural difference analysis of different platforms is not in-depth enough. Future research can expand the gender research scope, increase cross-cultural and cross-platform comparative analysis,

and further optimize the linguistic feature quantification model to improve the accuracy of digital gender identity recognition.

5. Conclusion

Based on big data NLP technology and econometric analysis methods, this study takes 1.28 million social media valid posts as the research sample, quantitatively analyzes the gender differences in social media lexical, syntactic and emotional linguistic features, and explores the internal mechanism of language-based gender identity construction. The study finds that female users are significantly better than male users in lexical richness, positive emotional expression and interactive discourse frequency, while male users tend to be factual and concise instrumental expression. Linguistic feature indicators have significant predictive effects on online gender identity construction, and interactive discourse density is the core influencing factor of female digital identity shaping. This study confirms the stable correlation between social media language gender differences and identity construction, provides reliable empirical data and quantitative model support for digital sociolinguistics research, and enriches the theoretical system of big data discourse analysis and digital identity research.

References

[1] Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E. P., & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. PloS one, 8(9), e73791. https://doi.org/10.1371/journal.pone.0073791.

[2] Nguyen, D. P. (2017). Text as social and cultural data: A computational perspective on variation in text [Doctoral dissertation, University of Twente].

[3] Vonitsanos, G., Kanavos, A., & Mylonas, P. (2023, December). Decoding gender on social networks: An in-depth analysis of language in online discussions using natural language processing and machine learning. In 2023 IEEE International Conference on Big Data (BigData) (pp. 4618–4625). IEEE.

[4] Huffaker, D. A., & Calvert, S. L. (2005). Gender, identity, and language use in teenage blogs. Journal of Computer-Mediated Communication, 10(2), JCMC10211.

[5] Pang, D., Guntuku, S. C., Sherman, G., Liu, T., Rai, S., Cho, Y. M., You, R., Bakker, A. B., & Ungar, L. H. (2025). Understanding gender and age differences in language use: Cross-cultural insights from Weibo and Facebook. Humanities and Social Sciences Communications, 12(1), 1667. https://doi.org/10.1057/s41599-025-04158-4

[6] Aydın, F. (2025). Examining gender differences in social media language. Bulletin of Language and Literature Studies, 2(1).

[7] Thelwall, M., Wilkinson, D., & Uppal, S. (2010). Data mining emotion in social network communication: Gender differences in MySpace. Journal of the American Society for Information Science and Technology, 61(1), 190–199. https://doi.org/10.1002/asi.21180

[8] Tang, Q., Zhang, X., & Liu, T. (2026). A feminist critical discourse analysis of gender norms on Chinese social media: Empirical insights from WeChat public accounts. Plos one, 21(1), e0338967. https://doi.org/10.1371/journal.pone.0338967

[9] Wu, X., Al Imran Yasin, M., Binti Abdullah, K., Jiang, X., Zhao, Y., Li, W., & Zhao, F. (2025). Media culture and identity construction in the digital age: A multidimensional analysis of race, gender, class, and intercultural communication. International Journal of Multimedia Computing, 6(1), 44–55.

[10] Riaz, B., Shahzad, A. K., & Rehman, I. U. (2025). Digital landscape and social constructs: Analyzing the role of language in shaping gender identities in Pakistan. Policy Journal of Social Science Review, 3(6), 293–312.

[11] Agunbiade, O. M., Obiyan, M. O., & Sogbaike, G. B. (2013). Identity construction and gender involvement in online social networks among undergraduates in two universities, Southwest Nigeria. Inkanyiso, 5(1), 41–52.