"Feels Feminine to Me": Understanding Perceived Gendered Style through Human Annotations

Hongyu Chen[1], **Neele Falk**[2], **Michael Roth**[3], **Agnieszka Falenska**[1,2]
[1]Interchange Forum for Reflecting on Intelligent System, University of Stuttgart
[2]Institute for Natural Language Processing, University of Stuttgart
[3] Natural Language Understanding Lab, University of Technology Nuremberg
{hongyu.chen, agnieszka.falenska}@iris.uni-stuttgart.de
neele.falk@ims.uni-stuttgart.de, michael.roth@utn.de

Abstract

In NLP, language–gender associations are commonly grounded in the author's gender identity, inferred from their language use. However, this identity-based framing risks reinforcing stereotypes and marginalizing individuals who do not conform to normative language–gender associations. To address this, we operationalize the language–gender association as a perceived gender expression of language, focusing on how such expression is externally interpreted by humans, independent of the author's gender identity. We present the first dataset of its kind: 5,100 human annotations of *perceived gendered style*—human-written texts rated on a five-point scale from very feminine to very masculine. While perception is inherently subjective, our analysis identifies textual features associated with higher agreement among annotators: formal expressions and lower emotional intensity. Moreover, annotator demographics influence their perception: women annotators are more likely to label texts as feminine, and men and non-binary annotators as masculine. Finally, feature analysis reveals that text's perceived gendered style is shaped by both affective and function words, partially overlapping with known patterns of language variation across gender identities. Our findings lay the groundwork for operationalizing gendered style through human annotation, while also highlighting annotators' subjective judgments as meaningful signals to understand perception-based concepts.[1]

1 Introduction

Gender as a social construct encompasses identity and expression, two distinct but interrelated dimensions of how individuals experience and present their gender (Bucholtz, 2002; Zimman, 2013). *Gender identity* refers to an individual's internal sense of self and how they identify (e.g., woman, man, non-binary). In contrast, *gender expression* (e.g., feminine, masculine, or gender-neutral) relates to how individuals present their gender externally (Baum and Westheimer, 2015; Ehrensaft, 2018; Pinney et al., 2023). While gender identity and expression might align with binary gender categories, they frequently extend beyond, embracing a diverse spectrum of identities.

Figure 1: Overview of our study: annotators rate texts on a masculine–feminine scale, revealing how specific linguistic cues (e.g., emotion, verbs) shape subjective perceptions of gendered language style.

A prominent medium for gender expression is **gendered style of language**, patterns of language use such as word choice, tone, or sentence structure that are commonly associated with more feminine or masculine ways of communicating. Despite the sociolinguistic understanding that gendered style is not determined by one's identity (Bucholtz, 2002; Bamman et al., 2014), much of NLP work continues to conflate these two dimensions. Tasks such as authorship profiling and attribution (Mishra et al., 2018), text style transfer (Preoţiuc-Pietro et al., 2016; Kang et al., 2019), or even gender prediction from LLM-generated texts (Alowibdi, 2024) treat gendered stylistic variation as a stable source of information about the gender identity of their authors. Such approaches either risk misgendering individuals, especially those who do not conform to stereotypical linguistic patterns (Fosch-Villaronga et al., 2021), or reinforcing normative assumptions

[1]The datasets and experimental code for this work are available at github.com/HongyuChen2022/Gendered-Style-Annotation.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 31459–31480
November 4-9, 2025

about how people "should write", perpetuating cultural biases and marginalizing diverse gender expressions (Dev et al., 2021; Devinney et al., 2022).

Addressing these issues requires both conceptual clarity—distinguishing between gender identity and gender expression—and methodological innovation in how gendered style is modeled and annotated. In this work, we take the first step in this direction by examining **perceived gendered style** as a subjective, socially constructed phenomenon. To this end, we introduce a new dataset—the first of its kind—comprising 5,100 human annotations of perceived gendered style in text (see Figure 1 for an overview). Using this dataset, we answer three key research questions:

RQ1 To what extent do annotators agree in their perception of gendered style and which text features contribute to the agreement?

RQ2 Do perceived gendered style ratings vary by the sociodemographic background of annotators?

RQ3 Which textual features are distinct to perceived gendered style?

We find that perceived gendered style is inherently subjective, with readers frequently disagreeing on whether a given text feels "masculine" or "feminine" (§4). However, we also identify specific linguistic textual features that contribute to higher pairwise agreement among annotators: formal expressions and lower emotional intensity. Moreover, beyond textual properties, we observe a moderate association between annotator background and perception: women annotators are more likely to label texts as feminine, men and non-binary annotators as more masculine (§5). Building on these observations, and in line with recent work that treats label variation as a meaningful signal rather than noise (Cabitza et al., 2023; Plank, 2022), we conduct the first systematic analysis of perceived gendered style. Rather than collapsing annotations into a single label, we analyze the full distribution of annotator responses, investigating which linguistic features are most strongly contributing to perceived gendered styles variation (§6). Our feature analysis highlights that perceived gendered style is shaped by both affective and function properties of text. Specifically, feminine style emphasizes positive emotional features, whereas masculine style relies more on syntactic features and direct, dominance-oriented expressions. Finally, neutral style emerges as distinct, characterized by balanced emotional intensity and structural features.

Our contributions are twofold. First, we present a novel corpus for perceived gendered style, featuring perception-based scale rating that includes a neutral option—moving beyond traditional binary categories. Second, we show the feasibility of shifting from an author identity-based framework to a human perception-driven model of gendered style. Our analysis reveals systematic patterns of agreement across annotators. These insights suggest new directions for building NLP systems that model gender as a socially perceived concept, enabling more inclusive, bias-aware NLP applications.

2 Related Work

2.1 Perceived Gender Expression

In gender studies, along with insights from transgender and queer activism, researchers emphasize the distinction between gender identity and gender expression (Baum and Westheimer, 2015; Larson, 2017; Ehrensaft, 2018; Pinney et al., 2023). Gender expression itself can be understood along two axes: one's self-directed gender expression and how that expression is interpreted or perceived by others (Rubin and Greene, 1991). Research on **perceived gender expression** has largely focused on appearance-based cues, typically measured through perceived characteristics such as the use of subjective adjectives to describe images of women (Hamon, 2004; Hattori et al., 2007; Otterbacher, 2015).

In contrast, work on the **perceived gender expression of written texts** has, to our knowledge, consistently conflated gender style (feminine/masculine) with gender identity (woman/man). This line of research typically asks annotators to *guess* the author's gender based on their texts (Nguyen et al., 2014; Flekova et al., 2016; Preoţiuc-Pietro et al., 2017). For example, Flekova et al. (2016) showed that annotator judgments are strongly influenced by gender-stereotypical associations, such as linking sports-related terms to men and emotional terms to women. Preoţiuc-Pietro et al. (2017) further explored this by controlling for textual mediation and found that male-authored texts containing features stereotypically associated with women were more likely to be misclassified. While these studies consistently conclude that predicting author gender from text is challenging, they fail to engage with what this ambiguity reveals—namely, the variability of gendered expression itself, independent of author identity.

2.2 Gender Identity in Text

While the previous section explored how gender is *perceived* through linguistic style, we now shift focus to how **gender identity is expressed in language use**. Variation in language use across gender identities has been a central topic of sociolinguistic analyses (Becker et al., 2022; Bamman et al., 2014; Morales Sánchez et al., 2022). For example, Bamman et al. (2014) analyze lexical patterns in relation to assigned binary gender. While they identify certain linguistic markers associated with gender, their findings also emphasize that these associations are fluid, context-dependent, and not strictly aligned with binary categories.

Yet, these sociolinguistic nuances are often overlooked in NLP tasks that aim at leveraging gender-related linguistic variation **to infer (usually binary) gender from text**. Prior research has applied such gender prediction in contexts such as authorship profiling and analysis (Gjurković et al., 2021; Zhang, 2024; White and Cotterell, 2021; Skurla and Petrik, 2024; Chen et al., 2024) and feature engineering for gender classification (Mamgain et al., 2019; Bianchi et al., 2022; Onikoyi et al., 2023).

In parallel, a growing body of work has examined how gender identity is encoded in text from the perspective of **bias in NLP models** (Stanczak and Augenstein, 2021). Language models encode gender-related linguistic variation (Lauscher et al., 2022). Knupleš et al. (2024) demonstrate that this encoding is uneven across gender-identities, potentially leading to biased model behavior and downstream harms (Lalor et al., 2022). However, to the best of our knowledge, none of the NLP bias work has focused on gendered language styles as perceived, rather than identity inferred or embedded.

2.3 Subjectivity of Annotation in NLP

Finally, our work can be integrated into related research strand on **perspectivism and human label variation** (Aroyo and Welty, 2015; Plank, 2022; Cabitza et al., 2023): perceived gendered style is inherently subjective and there is no ground truth to how gendered a specific text *should* be perceived, hence reducing any annotations to a binary 'gold' label does not make sense. While modeling the distribution of human judgments might be a valid next step (Uma et al., 2021; Mostafazadeh Davani et al., 2022; Heinisch et al., 2023), this work focuses on understanding human label variation stemming from two sources: (a) linguistic features that characterize the text (linguistic features have been investigated as a source of disagreement, for instance in NLI, see Pavlick and Kwiatkowski, 2019) and (b) characteristics of the annotators themselves—specifically, their gender.

Prior research on the **influence of socio-cultural factors on annotation** outcomes has produced mixed findings. Some studies report significant effects, revealing systematic differences among annotators based on moral values (Mostafazadeh Davani et al., 2024), socio-demographic profiles (Wan et al., 2023; Al Kuwatly et al., 2020) or personal attitudes (Jiang et al., 2024), while others suggests that socio-demographic variables account for only a small fraction of the overall variation in human annotation (Hu and Collier, 2024). Given that our task—perceived gendered style—involves both stylistic aspects of language and gender as a socio-cultural construct, we hypothesize that both linguistic features and annotator's gender identity systematically influence annotation outcomes.

3 Data Selection and Annotation

We collect and annotate texts from three well-established datasets.

3.1 Data Selection

We selected three datasets for analysis: PAN13-EN, BLOG, and PASTEL (see details below). The two first are widely used benchmarks in gender prediction research, with relatively weak associations between text features and author identity (Morales Sánchez et al., 2022; Chen et al., 2024), making them well-suited for studying perceived gendered style. In contrast, PASTEL is used in gendered style transfer and offers more stylistically varied texts:

PAN13-EN is a large-scale dataset introduced as part of a shared task on authorship verification and identification (Rangel et al., 2013). It contains 283,240 conversational texts in English that span a wide range of everyday topics, with language representative of informal social media discourse.

BLOG refers to the Blog Authorship Corpus (Schler et al., 2006), which was constructed in August 2004 using data from `blogger.com`. The corpus comprises approximately 71,000 blogs and 681,284 individual posts.

PASTEL is a parallel stylistic language dataset designed for research on persona-conditioned lan-

guage variation (Kang et al., 2019). It contains approx. 41,000 parallel sentences and 8,300 parallel stories, each annotated across a range of personas.

Data selection started from equally sampling texts from the three datasets. Next, we manually removed any texts containing personal or private information, resulting in a set of 510 texts (see data statistics in Table 6, §A.2). Since PAN13-EN and BLOG were scraped from online sources, we performed minor preprocessing for readability by removing noisy characters and URLs. Finally, to ensure consistency across these two datasets, we truncated each sample to the first 100 characters. For PASTEL, each sample consists of five consecutive sentences, all of which were retained.

To analyze content variation across datasets, we extracted 50 topics using both BERTopic (Grootendorst, 2022) and LDA (Blei et al., 2003). Topic quality was evaluated with two metrics: (1) topic coherence and (2) topic diversity. As shown in §A.4.2, BERTopic outperforms LDA on both measures. We therefore report the top 5 BERTopic topics per dataset in Figure 7a, §A.2.

3.2 Annotation Setup

To obtain a comprehensive understanding of the perceived gendered style, we collected 10 independent ratings for each of the 510 texts. To minimize cognitive and reading fatigue, each annotator rated maximally 30–40 texts within a time frame of 20 to 30 minutes. Annotators rated each text on a 5-point scale: very feminine (1), somewhat feminine (2), neutral (3), somewhat masculine (4), and very masculine (5). To capture annotators' uncertainty for each of the texts, they also indicated their confidence level from 1 (not confident) to 4 (very confident). Finally, to ensure annotation quality, each survey included three attention checks. Annotators who failed at least two or completed the task in under 10 minutes were excluded from the analysis and replaced with new independent annotators. We also applied MACE (Hovy et al., 2013) to assess annotators' overall competence and reliability within the survey ($N = 130, \mu = 0.25,\ \sigma = 0.22$, for the competence distribution, see Figure 8, §A.4). Since all annotators passed the two primary filtering criteria, MACE scores served only as a consistency check and did not lead to further exclusions.

In total, we recruited 130 participants via Prolific[2], selecting only those who reported English as their native language and were located in the United States (for the demographics of the annotators, see Table 7, §A.4). Participants were compensated with an average reward of £9 per hour. They completed the survey either through Google Forms or a custom-built Streamlit app.[3]

Annotation Instructions Participants were asked to provide "their perception on the writing style" (see the exact annotation guidelines in Figure 6, §A.1). In total, we conducted 5 rounds of pilot studies. Based on the feedback from the pilot annotators (see Table 5, §A.2), we added to the guidelines brief "key features" (e.g., patterns commonly associated with linguistic variation across gender identities, such as collaborative tone or textual complexity) and examples for each style as optional references. While this decision reduced annotator confusion, it also introduced a potential confound in our dataset, as some judgments may have been influenced by the examples. To mitigate this effect, participants were explicitly encouraged to rely on their intuition and personal interpretation of the text. They were also asked to report confidence scores and provide open-ended comments to capture their individual perspectives.

Content and style are often difficult to disentangle in annotation studies. Therefore, following (Dollinger, 2015; Chan and Maglio, 2020), we hypothesized that passive phrasing would direct annotators' attention more toward style than content. Accordingly, we employed agent-less wording in most parts of the task framing, asking "is the text perceived" rather than "do you perceive".

Annotator Calibration As suggested by one of the reviewers, we assessed annotators reliability through a re-annotation study, conducted after a six-month interval to minimize potential memory effects. All annotators were invited to participate, and 10 agreed to take part. We then examined (1) the agreement of test–retest rating pairs using weighted Cohen's kappa for each of the 10 annotators, which showed that half of them reached moderate consistency ($N = 10, \mu = 0.51,\ \sigma = 0.17$); and (2) exact-match stability, measured as the average rating shift per re-annotator on the 5-point scale, which was low overall ($N = 10, \mu = 0.20, \sigma = 0.25$). These results suggest that annotators' retest responses were consistent with their

[2] https://www.prolific.com/

[3] Design of survey questions on both platforms is identical. The development of this survey, as well as the analysis code and text proofreading, was supported by AI assistants.

initial ratings, supporting the reliability of our annotations.

3.3 Annotation Results

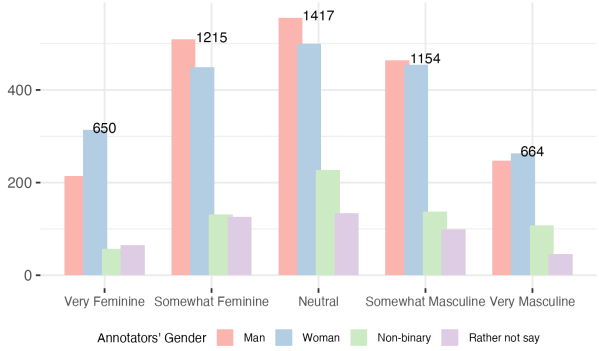


Figure 2: Frequency of gendered style annotations by self-reported gender of the annotators.

As a result of the annotation process, we collected 5,100 judgments of perceived gendered style, with each of the 510 texts receiving 10 style labels and 10 corresponding confidence scores. Figure 2 shows the frequency distribution of style annotations. Overall, the neutral style received the highest number of annotations ($N = 1417$), followed by "somewhat feminine" ($N = 1215$) and "somewhat masculine" ($N = 1154$). The average style rating across all annotations was ($\mu = 2.99,\ \sigma = 1.22$), and the average confidence score was ($\mu = 3.02,\ \sigma = 0.86$) which indicates a wide range of annotations and that the annotators in general felt confident about their judgments.

Finally, since one of our hypotheses is that annotators' own gender may influence their judgments (Wan et al., 2023; Al Kuwatly et al., 2020), we take an initial look at this relationship by grouping annotations based on self-reported gender of the annotators (colors in Figure 2). We find that women annotators contributed more annotations to extreme style categories compared to other gender groups. We come back to this topic in §5.

4 Annotator Agreement

We now turn our focus to **RQ1** and ask to what extent annotators agree with their perception of gendered style.

4.1 Inter-annotator Agreement

To gain a high-level understanding, we quantify inter-annotator agreement (IAA) for our data. Table 1 reports Krippendorff's alpha for the full annotation set, computed across 10 independent annotators for each of the 510 texts. The overall IAA across the five-point style scale is 0.22 highlighting the inherent subjectivity of this phenomenon.

Confidence	Agreement	Number of Annotations
all	0.22	5,100
>1	0.23	4,843
>2	0.25	3,773
>3	0.31	1,681

Table 1: Inter-annotator agreement scores: Krippendorff's alpha with ordinal level of measurement by confidence level and corresponding amount of annotations.

To further understand variation in agreement, we group annotations by self-reported confidence levels. Prior work has shown that confidence can serve as a proxy for annotator disagreement or uncertainty (Troiano et al., 2021). In line with this, we observe a positive association between confidence and agreement: annotators with the highest confidence (> 3) achieve a higher IAA (0.31) than those with moderate confidence (> 2, IAA 0.25). Pairwise observed agreement scores for individual texts are provided in Figure 9, §A.4.

In summary, while overall annotator agreement is generally low, higher self-reported confidence tends to indicate greater agreement.

4.2 Textual Features as Predictors of Agreement

As explained by Plank (2022), the variation in agreement is of analytical interest. To better understand the factors that contribute to this variation, we examine the role of textual features in shaping the agreement of gendered style.

Observed Agreement For each text instance, we calculate the raw consensus of pairwise observed agreements.[4] This measure captures the proportion of annotator pairs who assigned the same label to the same instance, without correcting for agreement expected by chance (for metrics details, see §A.3).

Feature Extraction We extract a total of 192 textual features from each annotated text using the ELFEN package with default parameters (Maurer, 2025). The features span several linguistic and stylistic dimensions, including surface-level metrics (e.g., token count), part-of-speech tags (e.g,

[4]This is part of Fleiss' kappa (Bem, 1974)

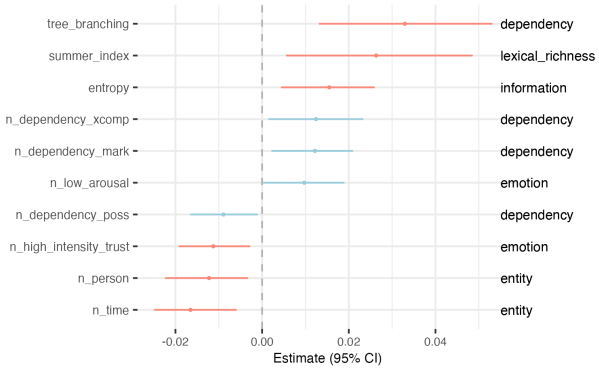


Figure 3: Forest plot showing the average bootstrap-estimated effects of the 10 most explanatory features in predicting annotator agreement across 1,000 resamples (linear regression, model fit: R^2 = 11.5%); horizontal lines show the corresponding 95% bootstrap confidence intervals. The estimates measure how strongly each feature affects the agreement (color in blue: $p < 0.01$; color in coral: $p < 0.05$)

number of adverbs), lexical richness (e.g., Sichel's index), readability scores (e.g., number of polysyllabic words), information density (e.g., compressibility), named entities (e.g., time entities), emotional tone (e.g., joy intensity), as well as semantic features like hedges (see Table 11, §A.4.3 for further details). We exclude 78 features due to missing values, high collinearity, or near-zero variance. In total, 114 features are retained for analysis (full list of features in Tables 12 and 13, §A.4.3).

Analysis Method We examine the explanatory power of textual features in predicting annotator agreement on gendered style using a linear regression model. The dependent variable (DV) is the pairwise observed agreement for each text, ranging from 0.111 to 0.644 ($\mu = 0.275$, $\sigma = 0.096$). The independent variables (IVs) consist of 114 textual features introduced in the section above. We evaluate model fit using R^2 and perform feature selection based on the Akaike Information Criterion (AIC), adding a feature only if the more complex model achieves a lower AIC. To obtain estimates, we applied nonparametric bootstrapping (1,000 resamples) to the AIC-selected model and report the mean for the coefficients and confidence intervals.

Results Figure 3 presents the bootstrapped results of our linear regression model. The model explains 11.5% of the variance in annotator agreement and includes 27 features. Among the predictors, features from five categories—part of speech, named entities, emotion, dependency structures, and lexical richness—were significantly associated with variation in agreement levels ($p < 0.05$).

Table 2 shows example texts with the most explanatory individual features (marked in blue) and the corresponding agreement scores. On the first place, the number of temporal entities (n_time) contributed 2.62% of the variance and is negatively associated with agreement. Such references to time (e.g., '3:00 am', '45 minutes' in Example (1)) can imply individuals' living patterns or actions and introduce personal contexts, potentially leading to diverse interpretations among annotators.

Similarly, on the emotion side, trust intensity (n_high_intensity_trust) explained 1.10% of variance and is also negatively correlated with agreement. Such components (e.g., 'faith' or 'a friend in need' in Example (2)) may convey reliability and bonds in a cultural context, likely contributing to lower agreement among annotators.

High agreement is strongly associated with emotion features such as low arousal (n_low_arousal), explaining 1.36% of variance. These constructions (e.g., 'Are you aware that' and 'Even though' in Example (3)) convey a neutral and explanatory tone that may promote shared interpretation.

Regarding structural features, we find that frequencies of dependency markers (n_dependency_mark) are positively associated with annotator agreement, explaining 1.04% of the variance. Texts with fewer subordinator cues tend to adopt a more instructional or formal tone (e.g., 'if you want...', 'who awaits...' in Example 4), likely contributing to higher agreement.

Overall, in response to **RQ1**, we find that annotator agreement is higher for texts that are emotionally neutral (n_low_arousal) and formally framed (n_dependency_mark), and lower for those that contain temporal references (n_time) or strong expressions that depends on cultural and contextual settings (n_high_intensity_trust).

5 Annotator Socio-Demographic and Perceived Gendered Style

The previous analysis provided insight into overall patterns of annotator agreement. We now turn our focus to how annotators perceive gendered style specifically (**RQ2**). Socio-demographic factors are known to influence perception and may, in our context, shape how individuals annotate perceived gendered styles. For example, annotators identifying with a particular gender may be more likely to per-

	Feature	Text Example	Feature Value	Agreement
(1)	n_time number of time entity	...I woke up at approximately 3:00 am and now it's 5:00 am... My usual pattern is that I'll fall into my eventual slumber, say 45 minutes before I have to wake up.	10.32	0.13
(2)	n_high_intensity_trust high trust intensity	Where love is there is faith... Love is the salt of life... A broken friendship may be soldered, but will never be sound. A friend in need is a friend indeed. Better alone that in bad company!!!	5.13	0.20
(3)	n_low_arousal low arousal	Are you aware that camels do not have only a thick row of eyelashes but also two layers of eyelids in order to protect their eyes from the desert sand? Even though this seems unnecessary in the beginning, human lashes actually serve a very similar function for keeping out dust and other particular...	4.00	0.49
(4)	n_dependency_marker dependency marker	If you want to succeed in the world must make your own opportunities as you go on. The man who waits for some seventh wave to toss him on dry land ... You can commit no greater folly than to sit by the roadside until someone comes along...	3.66	0.49

Table 2: Text examples from the dataset with normalized feature values of features that significantly influence observed agreement. Words contributing to key feature values are highlighted in blue.

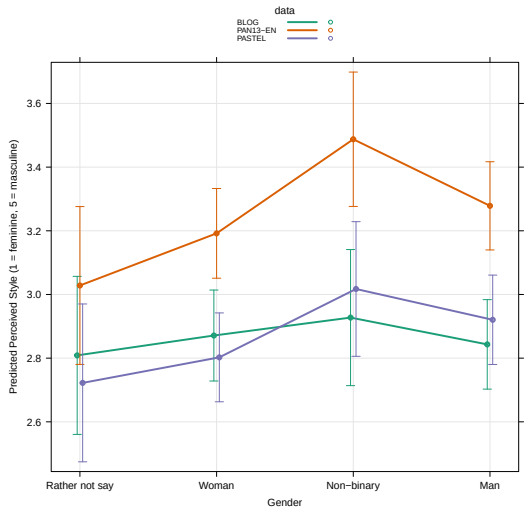


Figure 4: Marginalized effect of annotators' gender on perceived style. Error bars show 95% CIs; the y-axis is cropped for clarity. Style perception differs systematically by gender, with texts in PAN13-EN rated more neutral to masculine. Marginal $R^2 = 3\%$, Conditional $R^2 = 28\%$.

ceive and highlight gender-specific traits in texts. Therefore, we investigate the relationship between annotators' socio-demographic features and their perception of gendered style.

Analysis Method We examine the impact of annotators' self-reported socio-demographics using generalized mixed effect models. The perceived style of a single annotator is predicted on a scale from 1 (very feminine) to 5 (very masculine) and annotator's socio-demographics serve as fixed effects. To account for grouping structure, we include random effects for annotator ID and text ID, and examine how annotators' demographics interact with confidence and data source (e.g., whether the text is from the PASTEL or BLOG dataset).

Results Figure 4 visualizes how the self-ascribed gender impacts the style ratings when comparing the different data sources. Comparing the datasets, the plot shows, that texts in PAN13-EN in general receive higher style ratings than in the other two datasets, so being perceived more masculine, compared to PASTEL and BLOG irrespective of annotator's gender (orange line in Figure 4). This difference could either stem from different linguistic properties of the texts in that dataset or difference in frequently occurring topics. While BLOG and PASTEL focus more on personal and leisure topics (music videos, books, party), PAN13-EN contains more profession-oriented topics (business, medical, research) that are often more attributed to neutrality or masculinity (overview of frequent topics per dataset in Figure 7a, §A.2).

Regarding the relation between self-ascribed gender and perception, we can see most variation in the PAN13-EN and PASTEL dataset (orange and violet line): annotators identifying as 'rather not say' or 'woman' on average rate the style of texts as more feminine, while non-binary annotators or those identifying as 'man' perceive texts more neutral or masculine. This effect becomes stronger when we consider annotation confidence: the more confident an annotator is, the more their ratings shift towards the extremes, influenced by their self-identified gender. So when confident about a text, women tend to give more 'feminine' ratings, while men and non-binary annotators more 'masculine' (effect plot that visualizes this interaction can be found in Figure 11, §A.4).

6 Text Features and Perceived Gendered Style

Given that the previous analysis showed less variance coming from annotators' socio-demographics and more from the texts themselves, we now focus

Feature Category	Feature	Feminine vs. Masculine $R^2 = 11.39\%$	Feminine vs. Neutral $R^2 = 12.04\%$	Neutral vs. Masculine $R^2 = 4.3\%$
Dependency	n_dependency_dobj	+0.13 [***]		
	n_dependency_xcomp		+0.08 [***]	
	n_dependency_attr			-0.05 [***]
	n_dependency_amod			+0.05 [**]
	n_dependency_advcl			+0.06 [***]
Emotion	n_high_intensity_joy	-0.15 [***]		
	avg_valence	-0.14 [***]		
	avg_intensity_joy	-0.06 [***]	-0.03 [*]	
	avg_arousal	+0.07 [***]		+0.07 [***]
	avg_dominance	+0.12 [***]	+0.08 [***]	
	n_low_intensity_anger		+0.02 [*]	
	n_high_intensity_sadness			-0.04 [***]
	n_low_intensity_surprise			-0.04 [***]
	n_high_intensity_surprise			-0.04 [***]
	n_high_dominance			+0.06 [***]
Part of Speech	n_lexical_tokens	-0.38 [***]		
	n_adv	+0.07 [**]		
	n_pron		-0.05 [*]	
	n_intj		-0.03 [**]	
Surface	avg_word_length	-0.11 [***]		
Readability	smog	+0.05 [**]		+0.14 [***]
	n_polysyllables		+0.10 [***]	
Entity	n_org			+0.03 [**]

Table 3: Average bootstrap-estimated effects of the most explanatory features from three linear regression models that predict style rating (each comparing two gendered styles). Features are categorized into feature-type. Top row indicates model fit in terms of R^2. Coefficients are based on 1,000 bootstrap resamples; Significance levels (p < 0.1, ** p < 0.05, *** p < 0.01) are derived from bootstrap-based two-sided tests.

	Feature	Text Example	Feature Value	Style Perception
(1)	n_intj high interjections	hey everyone! wow....this warm weather is gettin the parties started...jay, u know what im talkin bout haha...never again...well not for a while...	4.35	4 × Feminine
(2)	n_high_dominance high dominance	How well your body works for you depends on what you put into it. It is vital to understand and practice proper nutrition in order to live a healthy life. Use these ideas and incorporate them into your daily nutrition regimen...	3.37	5 × Masculine
(3)	n_dependency_xcomp open clausal complement	The house was far from view. I tried to look up more photos of it. Every photo I clicked on said unavailable. I was starting to get frustrated. It seemed as if I wasn't going to be able to find anything.	3.00	5 × Neutral

Table 4: Text examples from the dataset with normalized feature values of features that significantly influence style perception. Words contributing to key feature values are highlighted in blue.

on the latter and investigate which text features are associated with perceived gendered style (**RQ3**).

6.1 Methods

To analyze how specific textual features correlate with different stylistic tendencies, we conduct three pairwise linear regression analyses, each comparing two gendered styles on a continuous scale: feminine vs. masculine (F vs M), feminine vs. neutral (F vs N), and neutral vs. masculine (N vs M). In all models, we use the textual features introduced in §4.2 as independent variables (IVs), and the numerical gendered style ratings from 5,100 annotations as the dependent variable (DV): 5,100 ratings for F vs M, 3,282 ratings for F vs N, and 3,235 ratings for N vs M. We perform feature-selection using AIC, and similar to the previous analysis (§4), we applied nonparametric bootstrapping (1,000 resamples) on the AIC-selected models.

6.2 Results

Table 3 presents estimated effects of the most explanatory features (full results in §A.4). The final regression models explain 11.39% of the variance in F vs M, 12.03% in F vs. N, and 4.3% in N vs M comparisons. Overall, features from six linguistic categories (dependency structures, emotion, entity, part-of-speech tags, readability, and surface-level attributes) influence perceived gendered text style.

We now discuss each of the styles individually. As an example, Table 4 presents one significant feature for each of them.

Feminine Style Several emotional and syntactic features are perceived as feminine. Emotion features such as frequent expressions of joy (avg_intensity_joy, n_high_intensity_joy) and a mild polarity (avg_valence) are positively associated with feminine style (F vs M). POS features, such as pronouns (n_pron), are prominent, as well as interjections (n_intj in F vs N), e.g., 'wow!',

'hey!' in Example (1). The result aligns with previous findings that women use emotive interjections more frequently (Stange, 2019).

Masculine Style Masculine style is more strongly associated with structural features (e.g., dependency_dobj in F vs M), and certain entities, such as organizations (n_org in N vs M). Lexically, texts that are associated with a more masculine style contain more adverbs (n_adv in F vs M). Interestingly, prior work links adverb use more strongly to female authors (Newman et al., 2008; Park et al., 2016; Chen et al., 2024). In terms of emotional features, texts perceived as more masculine tend to include direct expressions that convey high dominance (n_high_dominance in F vs M), e.g., 'It is vital to understand' and 'Use these ideas...' in Example (2). The result aligns with earlier findings on male authors' language use of direct expressions (Leaper and Ayres, 2007).

Neutral Style Neutral texts show a distinct set of emotional and structural features. While more feminine or masculine styles are characterized by stronger emotional expressions—such as intense joy or high dominance—neutral texts tend to express emotions more subtle and balanced, marked by lower intensity and arousal (n_low_intensity_anger and avg_dominance in F vs N). Compared to texts perceived as more feminine, they are also more readable (n_polysyllables in F vs N) and include more subject-controlled structures (n_dependency_xcomp) indicating a chain of actions or behaviors (cf., '...tried to look up' and 'was starting to get...' in Example (3)). Compared to texts perceived as more masculine, they show a more negative polarity but at the same time a higher presence of surprise-related words, indicating a more balanced use of emotions (n_high_intensity_sadness and n_low/high_intensity_surprise in N vs M).

In response to **RQ3**, distinct linguistic features are systematically associated with perceptions of feminine, masculine, and neutral text styles. Specifically, feminine style is linked to a higher polarity and emotionally positive language (e.g., high-intensity joy), use of function words (n_pron), and interjections. Masculine style is characterized by syntactic features and the use of more direct expressions (dominance). Neutral texts tend to show both reduced and polarized emotional intensity and more complex structures.

7 Discussion and Conclusion

The association between language and gender has long been a central focus in NLP. However, a key ethical and methodological challenge remains: how should gender be operationalized in these tasks? To move toward a more inclusive and perception-aware approach, we examine perceived gendered style through human annotation. Rather than collapsing responses into a single aggregated label, we treat each annotation as a valid, individual perception. While inter-annotator agreement is moderate overall, over 70% of annotations were rated by annotators themselves as "moderate" or "very" confident, indicating that individual judgments are meaningful even in the absence of consensus.

Regarding gendered style itself, our findings reveal that women annotators are more likely to label texts as feminine, men and non-binary annotators as more masculine, indicating a possible shared cultural or social alignment in interpreting style cues. Moreover, particular linguistic features have a stronger impact on their agreement. Finally, our style feature analysis shows that emotion, function words, and syntactic features are the key indicators of gendered styles. These results suggest that annotators' perceptions of gendered style are shaped by both affective and function properties of text. Interestingly, these perceptions only partially map to the identity-based gender signals observed in previous work, which further underscores the distinction of patterns between perceived gendered style and authors' gender identity.

As for neutral style, prior research often conceptualizes neutrality in terms of sentiment, the absence of clearly positive or negative emotion (Son et al., 2022). Our analysis attempts to extend this view by showing that neutral style tends to exhibit distinct emotional intensity: less expressive than feminine, more polarized than masculine style. This suggests that perceptions of neutral style are not fixed, but rather depend on the relative positioning of a text along a continuum between feminine and masculine textual cues.

Combining all the evidence above, our study contributes to the perspective that gender in language is not a fixed, author-based trait, but a socially shaped perception that varies across readers and contexts. This opens the door for future NLP systems that can reason about style with greater nuance.

8 Limitations

Methodologically, our work offers a new perspective for representing language-gender associations in NLP tasks shifting from an author-centered, binary paradigm to a human-centered, perception-driven model of gendered language. However, this approach would benefit from direct comparison to author-identity-based patterns. Aligning perceived styles with actual author gender could offer more intuitive insights into how gender is both expressed and interpreted in text.

Our dataset is limited to 5,100 annotations across 510 texts. While sufficient for preliminary insights, a larger and more diverse dataset would better capture the variability of gendered expression and enhance the generalizability of our findings.

In terms of evaluation, our pairwise agreement metric captures overall agreement but does not disaggregate agreement by style category. Future work could explore what linguistic or contextual factors contribute to higher agreement within each perceived style (e.g., feminine vs masculine vs neutral).

Although our primary aim is to highlight the importance of human perception over identity labels, our work would benefit from a comparison with automatic annotation using state-of-the-art language models. Such comparisons could shed light on how closely machine predictions align—or diverge—from human perception in this task.

Finally, although we introduce a novel dataset to operationalize perceived gendered style, we did not evaluate its utility in downstream tasks—an avenue for future work. While the dataset is too small to train large language models, it represents a crucial first step: linguistic features with high annotator agreement can guide targeted, larger-scale data collection that would be infeasible without initial annotations. Moreover, the dataset can be leveraged to probe large language models for covert gendered-style biases—an area that, to our knowledge, remains underexplored. Beyond NLP, it also offers value for social science by investigating into which linguistic cues are stereotypically linked to femininity or masculinity and how these associations shape social perception across cultural and social contexts.

9 Ethics & Potential Risks

While this study does not conceptualize gender as a binary category, it measures perception of gendered style along a spectrum with the binary poles representing its endpoints (from feminine to masculine). However, gender identity and expression are far more diverse and nuanced. This simplification may have encouraged annotators to rely on gender stereotypes, as they were likely unable to account for the full spectrum of gender diversity in their annotations. Furthermore, gender is inherently intersectional; its expression and perception of gendered style are shaped by intersecting factors such as class, race, and cultural context.

The intent of the dataset presented here was to investigate perceived gendered style. This can help investigate potential stylistic biases in large language models (LLMs). For example, does the style of an LLM align more closely with a gender expression perceived as masculine? Or, in certain contexts, does the generated text reflect stylistic features that are stereotypically associated with specific gendered expressions?

At the same time, the dataset can be used to train models that predict perceived gender expression based on style or language use. However, even perspectivist models—which account for multiple interpretations—can have harmful consequences. For instance, mismatches between the intended gender expression and the predicted or perceived gender expression may reinforce stereotypes or misrepresent the individual's identity.

10 Acknowledgements

This work is supported by the Ministry of Science, Research, and the Arts, Baden-Württemberg through the project IRIS3D (Reflecting Intelligent Systems for Diversity, Demography, and Democracy, Az. 33-7533-9-19/54/5). We would like to thank the anonymous reviewers for their valuable feedback. We also thank Aidan Combs, Amelie Wührl, Aswathy Velutharambath, Chris Jenkins, Cornelia Sinderman, Esra Dönmez, Filip Miletić, Iman Jundi, Franziska Weeber, Madhumitha Arivu Chelvan, Nicola Fanton, Sebastian Padó, Simon Tannert, and Solange Vega for their inputs that helped improve this work.

References

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse*

and Harms, pages 184–190, Online. Association for Computational Linguistics.

Jalal S Alowibdi. 2024. Gender prediction of generated tweets using generative ai. *Information*, 15(8):452.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

Joel Baum and Kim Westheimer. 2015. Sex? sexual orientation? gender identity? gender expression? *Teaching Tolerance*, 50(34–38).

Kara Becker, Sameer ud Dowla Khan, and Lal Zimman. 2022. Beyond binary gender: creaky voice, gender, and the variationist enterprise. *Language Variation and Change*, 34(2):215–238.

Sandra L Bem. 1974. The measurement of psychological androgyny. *Journal of consulting and clinical psychology*, 42(2):155.

Federico Bianchi, Vincenzo Cutrona, and Dirk Hovy. 2022. Twitter-demographer: A flow-based tool to enrich Twitter data. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 289–297, Abu Dhabi, UAE. Association for Computational Linguistics.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Mary Bucholtz. 2002. From 'sex differences' to gender variation in sociolinguistics. *University of Pennsylvania Working Papers in Linguistics*, 8(3):33–45.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Eugene Y Chan and Sam J Maglio. 2020. The voice of cognition: Active and passive voice influence distance and construal. *Personality and Social Psychology Bulletin*, 46(4):547–558.

Hongyu Chen, Michael Roth, and Agnieszka Falenska. 2024. What can go wrong in authorship profiling: Cross-domain analysis of gender and age prediction. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 150–166, Bangkok, Thailand. Association for Computational Linguistics.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of "gender" in nlp bias research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2083–2102, New York, NY, USA. Association for Computing Machinery.

Stefan Dollinger. 2015. *The Written Questionnaire in Social Dialectology: History, theory, practice*. John Benjamins, Amsterdam.

Diane Ehrensaft. 2018. Exploring gender expansive expressions versus asserting a gender identity. In Colt Keo-Meier and Diane Ehrensaft, editors, *The gender affirmative model: An interdisciplinary approach to supporting transgender and gender expansive children*, pages 37–53. American Psychological Association.

Lucie Flekova, Jordan Carpenter, Salvatore Giorgi, Lyle Ungar, and Daniel Preoţiuc-Pietro. 2016. Analyzing biases in human perception of user age and gender from text. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 843–854.

E. Fosch-Villaronga, A. Poulsen, R.A. Søraa, and B.H.M. Custers. 2021. A little bird told me your gender: Gender inferences in social media. *Information Processing & Management*, 58(3):102541.

Matej Gjurković, Vanja Mladen Karan, Iva Vukojević, Mihaela Bošnjak, and Jan Snajder. 2021. PANDORA talks: Personality and demographics on Reddit. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 138–152, Online. Association for Computational Linguistics.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Philippe Hamon. 2004. What is a description? *Bal, M. Narrative Theory: Critical Concepts in Literary and Cultural Studies*, 1:309–340.

Shun Hattori, Taro Tezuka, and Katsumi Tanaka. 2007. Mining the web for appearance description. In *International Conference on Database and Expert Systems Applications*.

Philipp Heinisch, Matthias Orlikowski, Julia Romberg, and Philipp Cimiano. 2023. Architectural sweet spots for modeling human label variation by the example of argument quality: It's best to relate perspectives! In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11138–11154, Singapore. Association for Computational Linguistics.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with mace. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307, Bangkok, Thailand. Association for Computational Linguistics.

Aiqi Jiang, Nikolas Vitsakis, Tanvi Dinkar, Gavin Abercrombie, and Ioannis Konstas. 2024. Re-examining sexism and misogyny classification with annotator attitudes. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15103–15125, Miami, Florida, USA. Association for Computational Linguistics.

Dongyeop Kang, Varun Gangal, and Eduard Hovy. 2019. (male, bachelor) and (female, Ph.D) have different connotations: Parallelly annotated stylistic language dataset with multiple personas. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1696–1706, Hong Kong, China. Association for Computational Linguistics.

Urban Knupleš, Agnieszka Falenska, and Filip Miletić. 2024. Gender identity in pretrained language models: An inclusive approach to data creation and probing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11612–11631, Miami, Florida, USA. Association for Computational Linguistics.

John Lalor, Yi Yang, Kendall Smith, Nicole Forsgren, and Ahmed Abbasi. 2022. Benchmarking intersectional biases in NLP. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3598–3609, Seattle, United States. Association for Computational Linguistics.

Brian Larson. 2017. Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain. Association for Computational Linguistics.

Anne Lauscher, Federico Bianchi, Samuel R. Bowman, and Dirk Hovy. 2022. SocioProbe: What, when, and where language models learn about sociodemographics. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7901–7918, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Campbell Leaper and Melanie M Ayres. 2007. A meta-analytic review of gender variations in adults' language use: Talkativeness, affiliative speech, and assertive speech. *Personality and Social Psychology Review*, 11(4):328–363.

Sunakshi Mamgain, R. Balabantaray, and Ajit Kumar Das. 2019. Author profiling: Prediction of gender and language variety from document. *2019 International Conference on Information Technology (ICIT)*, pages 473–477.

Maximilian Maurer. 2025. Elfen - efficient linguistic feature extraction for natural language datasets. https://github.com/mmmaurer/elfen.

Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. 2018. Author profiling for abuse detection. In *Proceedings of the 27th international conference on computational linguistics*, pages 1088–1098.

Damián Morales Sánchez, Antonio Moreno, and María Dolores Jiménez López. 2022. A white-box sociolinguistic model for gender detection. *Applied Sciences*, 12(5):2676.

Damián Morales Sánchez, Antonio Moreno, and María Dolores Jiménez López. 2022. A white-box sociolinguistic model for gender detection. *Applied Sciences*, 12(5).

Aida Mostafazadeh Davani, Mark Diaz, Dylan K Baker, and Vinodkumar Prabhakaran. 2024. D3CODE: Disentangling disagreements in data across cultures on offensiveness detection and evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18511–18526, Miami, Florida, USA. Association for Computational Linguistics.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Matthew L Newman, Carla J Groom, Lori D Handelman, and James W Pennebaker. 2008. Gender differences in language use: An analysis of 14,000 text samples. *Discourse processes*, 45(3):211–236.

Dong Nguyen, Dolf Trieschnigg, A Seza Dogruöz, Rilana Gravel, Mariët Theune, Theo Meder, and Franciska De Jong. 2014. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 1950–1961. Association for Computational Linguistics.

Babatunde Onikoyi, N. Nnamoko, and Ioannis Korkontzelos. 2023. Gender prediction with descriptive textual data using a machine learning approach. *Nat. Lang. Process. J.*, 4:100018.

Jahna Otterbacher. 2015. Crowdsourcing stereotypes: Linguistic bias in metadata generated via gwap. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*.

Gregory Park, David Bryce Yaden, H Andrew Schwartz, Margaret L Kern, Johannes C Eichstaedt, Michael Kosinski, David Stillwell, Lyle H Ungar, and Martin EP Seligman. 2016. Women are warmer but no less assertive than men: Gender and language on facebook. *PloS one*, 11(5):e0155885.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Christine Pinney, Amifa Raj, Alex Hanna, and Michael D Ekstrand. 2023. Much ado about gender: Current practices and future recommendations for appropriate gender-aware information access. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, pages 269–279.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Daniel Preoţiuc-Pietro, Wei Xu, and Lyle Ungar. 2016. Discovering user attribute stylistic differences via paraphrasing. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 3030–3037. AAAI Press.

Daniel Preoţiuc-Pietro, Sharath Chandra Guntuku, and Lyle Ungar. 2017. Controlling human perception of basic user traits. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2335–2341, Copenhagen, Denmark. Association for Computational Linguistics.

Francisco Rangel, Paolo Rosso, Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. 2013. Overview of the Author Profiling Task at PAN 2013. In *CLEF conference on multilingual and multimodal information access evaluation*, pages 352–365. CELCT.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408.

Donald L Rubin and Kathryn L Greene. 1991. Effects of biological and psychological gender, age cohort, and interviewer gender on attitudes toward gender-inclusive/exclusive language. *Sex Roles*, 24:391–412.

J Schler, M Koppel, S Argamon, and JW Pennebaker. 2006. Effects of Age and Gender on Blogging in Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs. In *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, volume 1.

Adam Skurla and Juraj Petrik. 2024. Authorship profiling in political discourse on twitter: Age and gender determination. In *Proceedings of the International Conference on Computer Systems and Technologies 2024*, CompSysTech '24, page 82–86, New York, NY, USA. Association for Computing Machinery.

Jaebong Son, Hyung-Koo Lee, Hyoungyong Choi, and On-Ook Oh. 2022. Are neutral sentiments worth considering when investigating online consumer reviews? their relationship with review ratings. In *Proceedings of the 55th Hawaii International Conference on System Sciences*.

Karolina Stanczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *Preprint*, arXiv:2112.14168.

Ulrike Stange. 2019. The social life of emotive interjections in spoken british english. *Scandinavian Studies in Language*, 10(1):174–193.

Enrica Troiano, Sebastian Padó, and Roman Klinger. 2021. Emotion ratings: How intensity, annotation confidence and agreements are entangled. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 40–49, Online. Association for Computational Linguistics.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone's voice matters: Quantifying annotation disagreement using demographic information. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):14523–14530.

Jennifer C White and Ryan Cotterell. 2021. Examining the inductive bias of neural language models with artificial languages. *arXiv preprint arXiv:2106.01044*.

Shaomin Zhang. 2024. *Authorship Analysis in Chinese Social Media Texts*. Elements in Forensic Linguistics. Cambridge University Press.

Lal Zimman. 2013. Hegemonic masculinity and the variability of gay-sounding speech: The perceived sexuality of transgender men. *Journal of Language and Sexuality*, 2(1):1–39.

A Appendix

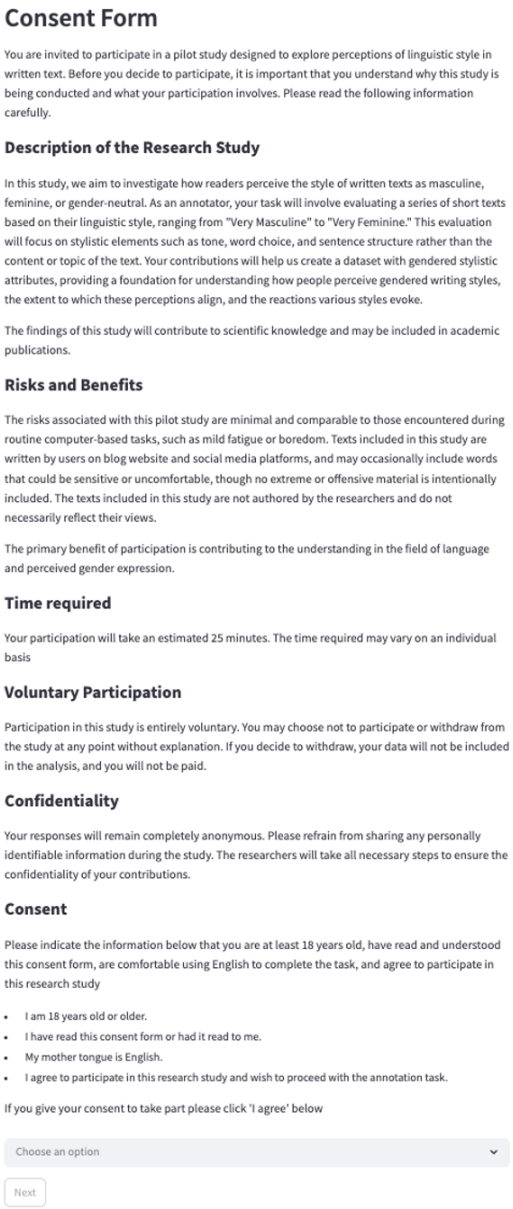
A.1 Annotation Guidelines

Table 5 presents a summary of the pilot studies and the corresponding changes. Overall, we conducted

Round	Main Task	Number of Texts	Changes
0	guessing style and author gender from texts	20	
1	guessing style	30 (including texts from previous round)	launched on Prolific; removed section of gender guessing; added examples and brief feature description
2	guessing style	40	
3	guessing style	40	new survey platform on Streamlit
4	guessing style	40	changed slider to radio buttons

Table 5: Iteration of pilot studies and corresponding changes.

5 rounds of pilot studies using Google Forms and Streamlit. After each round, we revised the annotation instructions and survey design in response to annotators' feedback. For instance, following Pilot 0—where four annotators evaluated 20 texts—we revised the task description and added illustrative examples and brief feature descriptions to help annotators better understand the task. Figure 6 presents the final annotation instructions and Figure 5 the consent form for annotators.



Consent Form

You are invited to participate in a pilot study designed to explore perceptions of linguistic style in written text. Before you decide to participate, it is important that you understand why this study is being conducted and what your participation involves. Please read the following information carefully.

Description of the Research Study

In this study, we aim to investigate how readers perceive the style of written texts as masculine, feminine, or gender-neutral. As an annotator, your task will involve evaluating a series of short texts based on their linguistic style, ranging from "Very Masculine" to "Very Feminine." This evaluation will focus on stylistic elements such as tone, word choice, and sentence structure rather than the content or topic of the text. Your contributions will help us create a dataset with gendered stylistic attributes, providing a foundation for understanding how people perceive gendered writing styles, the extent to which these perceptions align, and the reactions various styles evoke.

The findings of this study will contribute to scientific knowledge and may be included in academic publications.

Risks and Benefits

The risks associated with this pilot study are minimal and comparable to those encountered during routine computer-based tasks, such as mild fatigue or boredom. Texts included in this study are written by users on blog website and social media platforms, and may occasionally include words that could be sensitive or uncomfortable, though no extreme or offensive material is intentionally included. The texts included in this study are not authored by the researchers and do not necessarily reflect their views.

The primary benefit of participation is contributing to the understanding in the field of language and perceived gender expression.

Time required

Your participation will take an estimated 25 minutes. The time required may vary on an individual basis

Voluntary Participation

Participation in this study is entirely voluntary. You may choose not to participate or withdraw from the study at any point without explanation. If you decide to withdraw, your data will not be included in the analysis, and you will not be paid.

Confidentiality

Your responses will remain completely anonymous. Please refrain from sharing any personally identifiable information during the study. The researchers will take all necessary steps to ensure the confidentiality of your contributions.

Consent

Please indicate the information below that you are at least 18 years old, have read and understood this consent form, are comfortable using English to complete the task, and agree to participate in this research study

- I am 18 years old or older.
- I have read this consent form or had it read to me.
- My mother tongue is English.
- I agree to participate in this research study and wish to proceed with the annotation task.

If you give your consent to take part please click 'I agree' below

Choose an option

Next

Figure 5: Consent form for annotators.

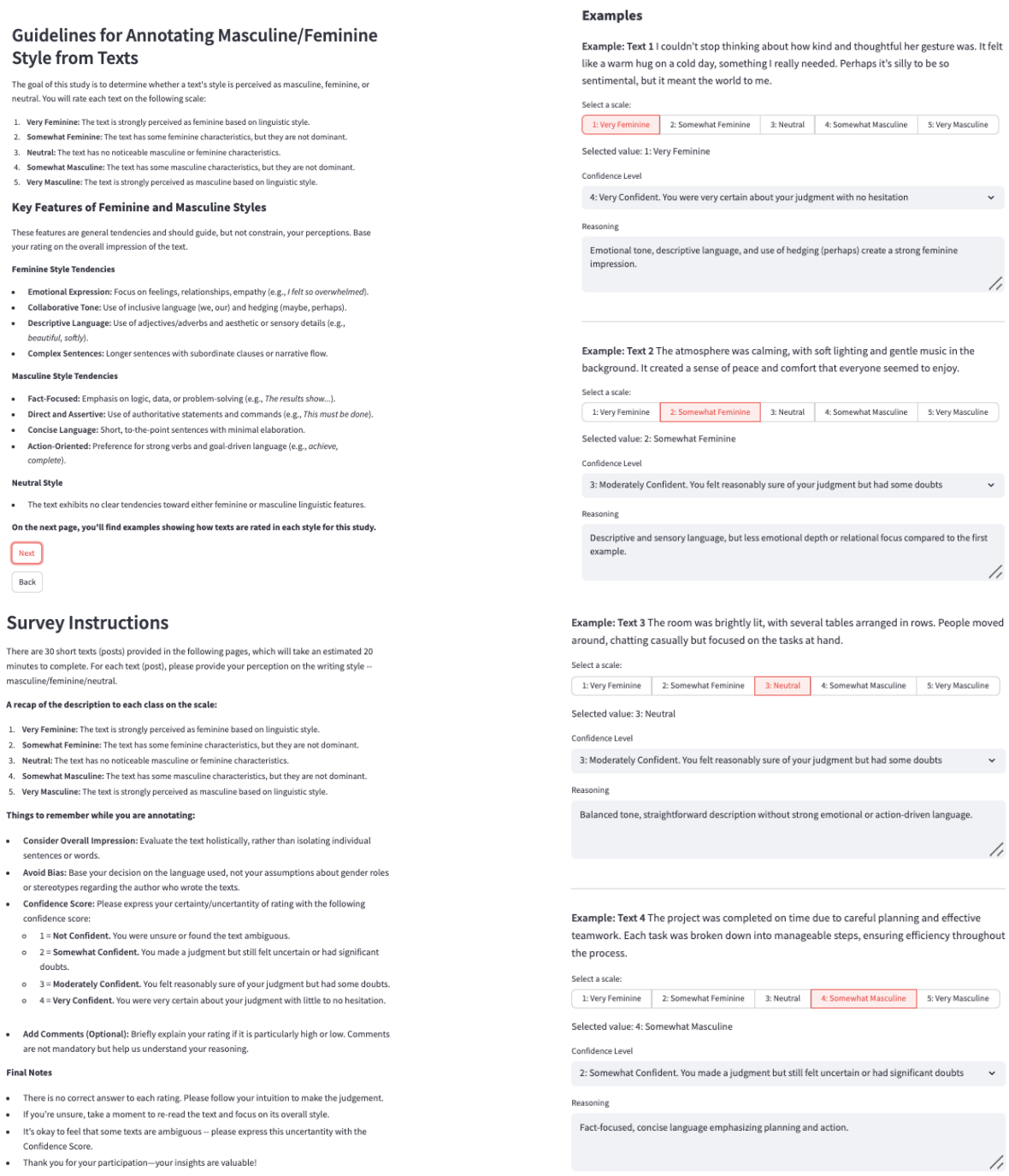


Figure 6: Annotation instructions with explained gendered style (left) and examples illustration (right).

A.2 Data Statistics

	Woman	Man	Non-binary	Total
BLOG	84	83	0	167
PAN13-EN	86	86	0	172
PASTEL	77	85	9	171
				510

Table 6: Sampled data proportion by authors' self-reported genders in each dataset.

Table 6 presents the data proportion by authors' self-reported genders in each dataset. Figure 7a shows the top 5-frequent topic distribution across three datasets. Overall, the datasets contain a comparable proportion of texts ($N = c(124, 140, 139) = 403$) , but the dominant topics differ substantially. In the BLOG dataset, the most frequent topics are related to blogging (8_blog_post_read_comment) and music (7_watch_music_live_video). PAN13-EN is dominated by themes of life (2_life_say_tell_problems) and work-related topics (business_web_design_website). PASTEL highlights topics associated with vacations (0_vacation_beach_trip_view) and memorial-related topics (1_stood_soldiers_lives_trees).

Figure 7b shows the top 5 most frequent topic distributions by author gender, with an equal number of female and male authors and a few non-binary authors ($N = (198, 198, 7) = 403$). Across all genders, the most frequent topics are related to life and social events. For example, life (2_life_say_tell_problems) and memorial-related themes (1_stood_soldiers_lives_trees) are common across groups.

Among female authors, vacation (0_vacation_beach_trip_view) and museum visits (3_museum_piece_sign_art) are especially frequent. At the same time, some topics are more strongly associated with particular genders. For instance, female authors are more likely to discuss food and cooking (4_food_table_ate_dinner), as well as parties and positive emotions (5_costume_party_couple_excited). Male authors, by contrast, more often mention leisure activities (17_game_ball_pool_guitar) and friendship related content (10_friends_good time_friend_relationships). Finally, for non-binary authors, the most frequent topic concerns social events such as performances (6_performance_dressed_city_gay).

A.3 Metrics for Pairwise Observed Agreement

We applied the following metrics to calculate the pairwise observed agreement among annotators. For a text instance i annotated by n annotators, each assigning a label from a set of k possible styles, let n_{ij} denote the number of annotators who assigned style j to item i.

$$\frac{n(n-1)}{2} \tag{1}$$

The number of agreeing annotator pairs for text instance i is computed by summing over all styles:

$$A_i = \sum_{j=1}^{k} \frac{n_{ij}(n_{ij}-1)}{2} \tag{2}$$

The pairwise observed agreement for text instance i is:

$$P_i = \frac{A_i}{\frac{n(n-1)}{2}} = \frac{\sum_{j=1}^{k} n_{ij}(n_{ij}-1)}{n(n-1)} \tag{3}$$

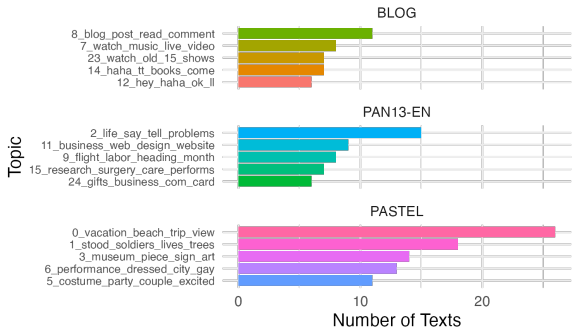
A.4 Analysis

A.4.1 Annotation Statistics

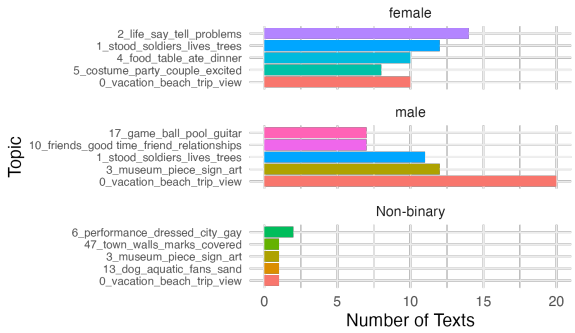
See Table 7 for annotators' socio-demographics statistics, Figure 9 for the distribution of pairwise observed agreement between annotators, and Table 8 for the majority style distribution by authors' gender.

Demographics	Value
age	39 ± 12
annotation time	35 ± 16
sex	female: 71
	male: 59
gender	Woman: 51
	Man: 50
	Non-binary: 17
	Rather Not to Say:12
race	Asian: 4
	Black: 28
	Mixed: 7
	Other: 1
	White: 90
employment status	EXPIRED: 35
	Full-Time: 33
	Not in paid work: 9
	Other: 5
	Part-Time: 40
	Unemployed: 8

Table 7: Summary of annotators' socio-demographics and annotation statistics.



(a) Top 5-frequent topics across datasets ($N = c(124, 140, 139) = 403$).



(b) Top 5-frequent topics across author gender identities.

Figure 7: Topic distribution in the datasets.

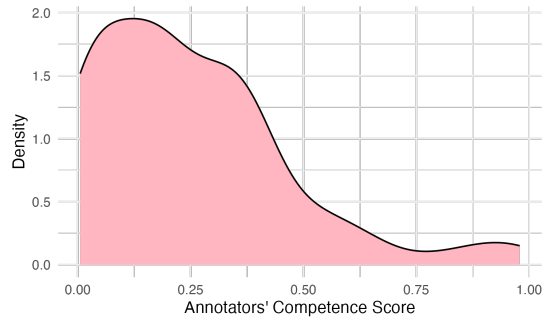


Figure 8: Distribution of individual annotator's competence and reliability within survey (130 annotators in total).

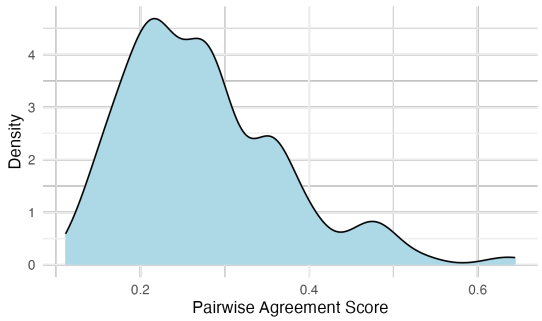


Figure 9: Distribution of pairwise observed agreement between annotators for each text instance (510 texts in total).

	1	2	3	4	5	Total
Female	40	73	80	40	14	247
Male	23	71	89	56	15	254
Non-binary		5	4			9
						510

Table 8: Majority style distribution by authors' gender. style 1 = very feminine to 5 = very masculine

Annotations by Author Gender Table 8 shows the distribution of majority gendered style annotations by author gender. With an approximately balanced number of female and male authors ($N = c(247, 254)$), the most frequent majority rating for both groups was 3 (neutral), followed by 2 (somewhat feminine). Among female-authored texts, nearly half received a majority vote of 1 or 2 (feminine). Very masculine (5) was the least frequent label for female-authored texts – a pattern interestingly mirrored in male-authored texts, where very feminine (1) was also less frequently. For non-binary authors, the sample size is small, but notably, none of their texts received a majority vote of "masculine". Finally, only a small proportion of texts received majority ratings that strongly aligned (1 or 5) with the author's gender, with a slight asymmetry: very feminine ratings for female authors occurred more frequently than very masculine ones for male authors.

Topics by Style Figure 10 shows the top 5 most frequent topic distributions across annotations by style. Vacation-related themes (0_vacation_beach_trip_view) are the most frequent across all styles, often accompanied by memorial-related topics (1_stood_soldiers_lives_trees).

Feminine style, emotion-centered content (16_love_coz_dreams_share) appears most prominently, alongside cooking and food (4_food_table_ate_dinner).

Neutral style, by contrast, highlights collective experiences, such as museum visits and performances (3_museum_piece_sign_art; 6_performance_dressed_city_gay) in 3.

Masculine style is marked by references to music videos (7_watch_music_live_video) in

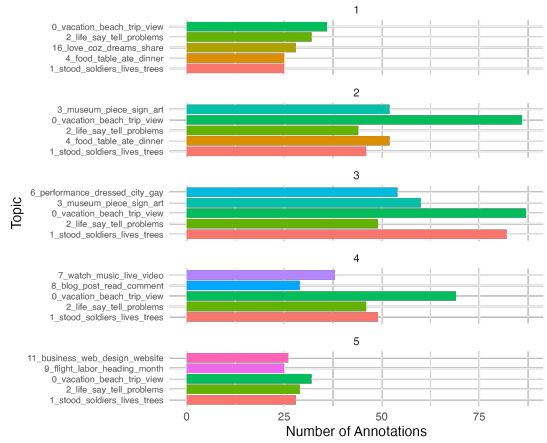


Figure 10: Top 5-frequent topic distribution across styles (annotations, $N = c(513, 963, 1147, 894, 513) = 4030$).

4 as well as professional and work-related themes (11_business_web_design_website and 9_flight_labor_heading_month) in 5.

Overall, while general life activities are present across all styles, feminine annotations tend toward emotions and food, neutral toward social events, and masculine toward work and media. Looking back at the distribution of topics across gender identities of authors (Appendix A.2), the dominant topics across author genders align with those seen in gendered styles overall, especially life, vacation, and memorial-related themes, which may blur distinctions for annotators. However, we also observe correspondences and divergences: the feminine style mirrors female authors (e.g., food and positive emotions), while the masculine style diverges from male authors, emphasizing music and blogging rather than gaming and friendship. This suggests that content patterns by author gender and those perceived as gendered style do not always overlap.

A.4.2 Topic Modeling

We measure topic coherence with normalized pointwise mutual information (NPMI) combined with cosine similarity (Röder et al., 2015), and topic diversity quantified as the proportion of unique words among the top terms of all topics. As shown in Table 9, BERTopic (107 texts with topic "-1" excluded) outperforms LDA on both metrics (coherence: 0.446 vs. 0.300; diversity: 0.947 vs. 0.672), suggesting that topics extracted from BERTopic is more semantically informative than that from LDA.

Table 9 shows the comparison between LDA and BERTopic. Table 10 presents three examples comparing topic content between BERTopic and LDA. Overall, BERTopic provides more semantically informative representations than LDA, and also outperforms LDA in terms of topic coherence and diversity. For example, the text in Example (1) centers on a personal memorial moment in a cemetery during winter. BERTopic captures this with keywords such as "soldiers" and "lives", whereas LDA emphasizes more generic terms like "walk" and "life", which miss the main theme of the text. Similarly, in Example (3), BERTopic highlights content relevant to health and nutrition through keywords such as "healthy" and "protein", while LDA instead yields abstract terms like "life" and "god", which do not accurately reflect the original text.

Model	Coherence (C_v)	Diversity
LDA	0.300	0.672
BERTopic	0.446	0.947

Table 9: Comparison of topic coherence and diversity between LDA and BERTopic.

A.4.3 Textual Features

Table 11 presents the description of extracted text features and Table 13 shows all removed features from the analysis.

A.4.4 Feature Analysis

Figure 11 presents effect plot for the interaction between annotators' confidence and their gender. Tables 14 to 17 show average bootstrap-estimated effect sizes for various experiments.

	Text	LDA Topic_Words	BT_Topic_Words
(1)	One winter's day, I was driving past the cemetery on my way to the airport. I decided to stop for a few minutes and take a walk in the snow. The trees reminded me of a park I visited long ago. I continued to walk through the cold snow. Before I headed back to my car, I decided to walk through the cemetery and pay my respects to those who have died.	long, walk, end, life, snow	stood, soldiers, lives, trees, bird
(2)	Wedding is just about the interpersonal customs of joining two individuals jointly. It is the very first step in raising a family group for this reason in spite of cultural standing up, many individuals devote high of their cash in order to use a respectable marriage ceremony. A few young couples are employing being married limousine to add an expression regarding class in their marriage ceremony....	nice, week, make, give, stress	costume, party, couple, excited, happy
(3)	How well your body works for you depends on what you put into it. It is vital to understand and practice proper nutrition in order to live a healthy life. Use these ideas and incorporate them into your daily nutrition regimen. A great life depends on good nutrition! Altering one's cooking techniques may greatly improve the quality of food. By steaming or boiling your food as opposed to frying it, you will be able to cut down on fat. Preparing your meals in a healthy way allows you to eat more nutritious foods.	life, live, god, bad, watch	healthy, depends, protein, did, ve

Table 10: Examples from topic content comparison between BERTopic and LDA topic models

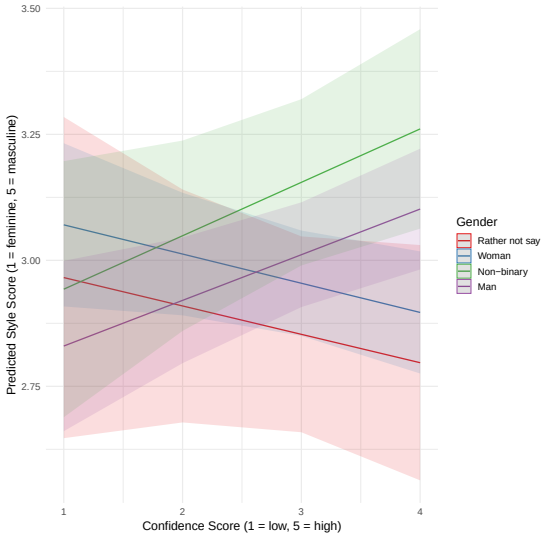


Figure 11: Predicted values of style score across levels of confidence score (1–5), separated by gender. The lines represent the interaction between confidence and gender: differences in slopes indicate that the effect of confidence on style score varies across gender groups. Marginal $R^2 = 1\%$, Conditional $R^2 = 28\%$.

Feature Category	N	Description
surface	4	features including number of tokens, sentences, average word length, etc
pos	14	part of speech features: encompassing the number of tokens with pos tags
lexical_richness	11	includes measures of lexical diversity, lexical sophistication, etc
readability	4	includes metrics that evaluate the readability of texts
information	2	compressibility and entropy
entities	8	number of named entities
semantic	1	number of semantic words: hedge
emotion	35	number of sentiment words: joy, valence, dominance, etc
dependency	35	number of dependencies of type: adjectival complement, attribute; tree branching, etc

Table 11: Description of extracted text features.

Feature	Feature Area	Name in extracted dataframe
Raw sequence length/total number of characters	surface	raw_sequence_length
Number of tokens	surface	n_tokens
Number of sentences	surface	n_sentences
Number of token per sentence	surface	tokens_per_sentence
Number of characters	surface	n_characters
Characters per sentence	surface	characters_per_sentence
Raw sequence length per sentence	surface	raw_length_per_sentence
Average word length	surface	avg_word_length
Number of types	surface	n_types
Number of long words	surface	n_long_words
Number of lemmas	surface	n_lemmas
Token frequencies	surface	token_freqs
Number of lexical tokens	pos	n_lexical_tokens
POS variability	pos	pos_variability
Number of tokens with upos tag {pos}	pos	n_{pos}
Lemma token ratio	lexical_richness	lemma_token_ratio
Type token ratio	lexical_richness	ttr
Root type token ratio	lexical_richness	rttr
Corrected type token ratio	lexical_richness	cttr
Herdan's C	lexical_richness	herdan_c
Summer's type token ratio/ index	lexical_richness	summer_index
Dugast's Uber index	lexical_richness	dugast_u
Maas' text token ratio/index	lexical_richness	maas_index
Number of local hapax legomena	lexical_richness	n_hapax_legomena
Number of global token hapax legomena	lexical_richness	n_global_token_hapax_legomena
Number of global lemma hapax legomena	lexical_richness	n_global_lemma_hapax_legomena
Number of hapax dislegomena	lexical_richness	n_hapax_dislegomena
Number of global token hapax dislegomena	lexical_richness	n_global_token_hapax_dislegomena
Number of global lemma hapax dislegomena	lexical_richness	n_global_lemma_hapax_dislegomena
Sichel's S	lexical_richness	sichel_s
Global Sichel's S	lexical_richness	global_sichel_s
Lexical density	lexical_richness	lexical_density
Giroud's index	lexical_richness	giroud_index
Measure of Textual Lexical Density (MTLD)	lexical_richness	mtld
Hypergeometric Distribution Diversity (HD-D	lexical_richness	hdd
Moving-average type token ratio (MATTR)	lexical_richness	mattr
Mean segmental type token ratio (MSTTR)	lexical_richness	msttr
Yule's K	lexical_richness	yule_k
Simpson's D	lexical_richness	simpsons_d
Herdan's Vm	lexical_richness	herdan_v
Number of syllables	readability	n_syllables
Number of monosyllables	readability	n_monosyllables
Number of polysyllables	readability	n_polysyllables
Flesch reading ease	readability	flesch_reading_ease
Flesch-Kincaid Grade Level	readability	flesch_kincaid_grade
Automated Readability Index (ARI)	readability	ari
Simple Measure of Gobbledygook (SMOG)	readability	smog
Coleman-Liau Index (CLI)	readability	cli
Gunning-fog Index	readability	gunning_fog
LIX	readability	lix
RIX	readability	rix
Compressibility	information	compressibility
Entropy	information	entropy
Number of named entities	entities	n_entitites
Number of named entities of type {ent}	entities	n_{ent}
Number of hedge words	semantic	n_hedges
Hedges token ratio	semantic	hedges_ratio
Average number of synsets	semantic	avg_n_synsets
Number of words with a low number of synsets per pos	semantic	n_low_synsets_{pos}
Number of words with a high number of synsets per pos	semantic	n_high_synsets_{pos}
Number of words with a low number of synsets	semantic	n_low_synsets
Number of words with a high number of synsets	semantic	n_high_synsets
Average valence	emotion	avg_valence
Number of low valence tokens	emotion	n_low_valence
Number of high valence tokens	emotion	n_high_valence
Average arousal	emotion	avg_arousal
Number of low arousal tokens	emotion	n_low_arousal
Number of high arousal tokens	emotion	n_high_arousal
Average dominance	emotion	avg_dominance
Number of low dominance tokens	emotion	n_low_dominance
Number of high dominance tokens	emotion	n_high_dominance
Average emotion intensity for {emotion}	emotion	avg_intensity_{emotion}
Number of high intensity tokens for {emotion}	emotion	n_high_intensity_{emotion}
Number of low intensity tokens for {emotion}	emotion	n_low_intensity_{emotion}
Sentiment score	emotion	sentiment_score
Number of negative sentiment tokens	emotion	n_negative_sentiment
Number of positive sentiment tokens	emotion	n_positive_sentiment
Dependency tree width	dependency	tree_width
Dependency tree depth	dependency	tree_depth
Tree branching factor	dependency	tree_branching
Tree ramification factor	dependency	ramification_factor
Number of noun chunks	dependency	n_noun_chunks
Number of dependencies of type {type}	dependency	n_dependency_{type}

Table 12: Detailed description of extracted text features.

Feature	Reason
n_conj	has_missing_values
hdd	has_missing_values
n_law	has_missing_values
n_language	has_missing_values
synsets	has_missing_values
synsets_noun	has_missing_values
synsets_verb	has_missing_values
synsets_adj	has_missing_values
synsets_adv	has_missing_values
avg_n_synsets	has_missing_values
avg_n_synsets_noun	has_missing_values
avg_n_synsets_verb	has_missing_values
avg_n_synsets_adj	has_missing_values
avg_n_synsets_adv	has_missing_values
n_high_synsets	has_missing_values
n_low_synsets	has_missing_values
n_high_synsets_noun	has_missing_values
n_high_synsets_verb	has_missing_values
n_high_synsets_adj	has_missing_values
n_high_synsets_adv	has_missing_values
n_low_synsets_noun	has_missing_values
n_low_synsets_verb	has_missing_values
n_low_synsets_adj	has_missing_values
n_low_synsets_adv	has_missing_values
tree_depth	has_missing_values
n_dependency_nounmod	has_missing_values
n_dependency_npmod	has_missing_values
n_dependency_root	has_missing_values
n_tokens	high collinearity
n_types	high collinearity
n_characters	high collinearity
maas_index	high collinearity
n_hapax_legomena	high collinearity
n_global_token_hapax_legomena	high collinearity
n_hapax_dislegomena	high collinearity
n_global_lemma_hapax_dislegomena	high collinearity
n_global_token_hapax_dislegomena	high collinearity
n_syllables	high collinearity
flesch_reading_ease	high collinearity
flesch_kincaid_grade	high collinearity

Feature	Reason
ari	high collinearity
cli	high collinearity
gunning_fog	high collinearity
lix	high collinearity
rix	high collinearity
n_dependency_advmod	high collinearity
n_dependency_prep	high collinearity
n_dependency_punct	high collinearity
raw_sequence_length	high collinearity
lemma_token_ratio	high collinearity
n_lemmas	high collinearity
cttr	high collinearity
ttr	high collinearity
herdan_c	high collinearity
rttr	high collinearity
mattr	high collinearity
yule_k	high collinearity
n_cconj	high collinearity
n_det	high collinearity
n_dependency_auxpass	high collinearity
n_adp	high collinearity
n_sym	near_zero_variance
n_x	near_zero_variance
n_money	near_zero_variance
n_product	near_zero_variance
n_percent	near_zero_variance
n_work_of_art	near_zero_variance
n_quantity	near_zero_variance
n_norp	near_zero_variance
n_loc	near_zero_variance
n_event	near_zero_variance
n_fac	near_zero_variance
n_dependency_agent	near_zero_variance
n_dependency_csubjpass	near_zero_variance
n_dependency_meta	near_zero_variance
n_dependency_oprd	near_zero_variance
n_dependency_parataxis	near_zero_variance
n_dependency_preconj	near_zero_variance
n_dependency_quantmod	near_zero_variance

Table 13: A list of all removed features from the analysis with reasoning

Term	original	mean	median	ci_low	ci_high	p_value	explvar
n_time	-0.02	-0.02	-0.02	-0.02	-0.01	0.00	2.62
tree_branching	0.03	0.03	0.03	0.01	0.05	0.00	1.77
n_low_arousal	0.01	0.01	0.01	0.00	0.02	0.04	1.36
n_high_intensity_trust	-0.01	-0.01	-0.01	-0.02	-0.00	0.01	1.10
n_dependency_mark	0.01	0.01	0.01	0.00	0.02	0.02	1.04
n_dependency_xcomp	0.01	0.01	0.01	0.00	0.02	0.03	0.88
summer_index	0.03	0.03	0.03	0.01	0.05	0.00	0.85
entropy	0.02	0.02	0.02	0.00	0.03	0.01	0.85
n_person	-0.01	-0.01	-0.01	-0.02	-0.00	0.01	0.81
n_dependency_poss	-0.01	-0.01	-0.01	-0.02	-0.00	0.03	0.69

Table 14: Average bootstrap-estimated effect (relative amount of R^2) of the 10 most predictive linguistic features (sorted by variance) of the linear regression model predicting annotator's agreement.

Term	original	mean	median	ci_low	ci_high	p_value	explvar
avg_word_length	-0.11	-0.11	-0.11	-0.18	-0.05	0.00	2.43
avg_dominance	0.12	0.12	0.12	0.07	0.17	0.00	0.89
n_lexical_tokens	-0.38	-0.38	-0.38	-0.52	-0.25	0.00	0.87
avg_valence	-0.14	-0.14	-0.14	-0.19	-0.09	0.00	0.86
avg_intensity_joy	-0.06	-0.06	-0.06	-0.11	-0.02	0.00	0.86
n_high_intensity_joy	-0.15	-0.15	-0.15	-0.20	-0.11	0.00	0.77
n_dependency_dobj	0.13	0.13	0.13	0.08	0.17	0.00	0.49
avg_arousal	0.07	0.07	0.07	0.01	0.13	0.02	0.43
smog	0.05	0.05	0.05	0.00	0.10	0.05	0.38
n_adv	0.07	0.07	0.07	0.01	0.12	0.04	0.35

Table 15: Average bootstrap-estimated effect sizes (relative amount of R^2) of the 10 most predictive linguistic features (sorted by variance) of the linear regression model predicting style ratings (from 1 (very feminine) to 5 (very masculine).

Term	original	mean	median	ci_low	ci_high	p_value	explvar
n_polysyllables	0.10	0.10	0.10	0.04	0.16	0.01	1.49
n_pron	-0.05	-0.05	-0.05	-0.10	0.00	0.06	1.38
n_intj	-0.03	-0.03	-0.03	-0.06	-0.00	0.04	1.14
n_lexical_tokens	-0.06	-0.06	-0.06	-0.14	0.01	0.12	1.09
avg_intensity_joy	-0.03	-0.03	-0.03	-0.06	0.00	0.08	1.04
n_high_intensity_joy	-0.07	-0.08	-0.08	-0.11	-0.04	0.00	0.66
n_high_valence	-0.04	-0.04	-0.04	-0.09	0.01	0.15	0.52
avg_dominance	0.08	0.08	0.08	0.04	0.12	0.00	0.42
n_dependency_xcomp	0.08	0.09	0.09	0.05	0.12	0.00	0.37
n_low_intensity_anger	0.02	0.03	0.03	-0.00	0.05	0.06	0.31

Table 16: Average bootstrap-estimated effect sizes (relative amount of R^2) of the 10 most predictive linguistic features (sorted by variance) of the linear regression model predicting style ratings (from 1 (very feminine) to 3 (neutral).

Term	original	mean	median	ci_low	ci_high	p_value	explvar
n_high_dominance	0.06	0.06	0.06	0.02	0.10	0.00	0.48
smog	0.14	0.14	0.14	0.07	0.20	0.00	0.40
avg_arousal	0.07	0.07	0.07	0.03	0.12	0.00	0.39
n_high_intensity_sadness	-0.04	-0.04	-0.04	-0.06	-0.01	0.00	0.27
n_dependency_advcl	0.06	0.06	0.06	0.02	0.09	0.00	0.27
n_dependency_amod	0.05	0.05	0.05	0.01	0.09	0.01	0.25
n_dependency_attr	-0.05	-0.05	-0.05	-0.07	-0.02	0.00	0.24
n_high_intensity_surprise	-0.04	-0.04	-0.04	-0.07	-0.02	0.00	0.24
n_low_intensity_surprise	-0.04	-0.04	-0.04	-0.07	-0.02	0.01	0.22
n_org	0.03	0.03	0.03	0.00	0.06	0.05	0.19

Table 17: Average bootstrap-estimated effect sizes (relative amount of R^2) of the 10 most predictive linguistic features (sorted by variance) of the linear regression model predicting style ratings (from 3 (neutral) to 5 (very masculine).