Self-reported Demographics and Discourse Dynamics in a Persuasive Online Forum

Agnieszka Falenska[1,*], **Eva Maria Vecchi**[2,*], **Gabriella Lapesa**[3]
[1]Interchange Forum for Reflecting on Intelligent Systems, University of Stuttgart, Germany
[2]Institute for Natural Language Processing, University of Stuttgart, Germany
[3]GESIS - Leibniz Institute for Social Sciences and Heinrich-Heine University of Düsseldorf, Germany
[1,2]first[-middle].last@ims.uni-stuttgart.de, [3]gabriella.lapesa@gesis.org

Abstract

Research on language as interactive discourse underscores the deliberate use of demographic parameters such as gender, ethnicity, and class to shape social identities. For example, by explicitly disclosing one's information and enforcing one's social identity to an online community, the reception by and interaction with the said community is impacted, e.g., strengthening one's opinions by depicting the speaker as credible through their experience in the subject. Here, we present a first thorough study of the role and effects of self-disclosures on online discourse dynamics, focusing on a pervasive type of self-disclosure: author gender. Concretely, we investigate the contexts and properties of gender self-disclosures and their impact on interaction dynamics in an online persuasive forum, ChangeMyView. Our contribution is twofold. At the level of the target phenomenon, we fill a research gap in the understanding of the impact of these self-disclosures on the discourse by bringing together features related to forum activity (votes, number of comments), linguistic/stylistic features from the literature, and discourse topics. At the level of the contributed resource, we enrich and release a comprehensive dataset that will provide a further impulse for research on the interplay between gender disclosures, community interaction, and persuasion in online discourse.

Keywords: Opinion Mining / Sentiment Analysis, Corpus (Creation, Annotation, etc.), Other

1. Introduction

Research on language as interactive discourse demonstrates that demographic parameters – gender, ethnicity, class – are intentionally communicatively used as boundaries to create our own social identities (Gumperz, 1982). The explicit definition of one's social identity impacts, intentionally or not, the role of the speaker, the expectations and standards of the audience, and the subsequent discourse dynamics. More specifically, individuals strategically present themselves to others by controlling the amount of information available to maintain a publicly desirable image, a concept known as impression management (Goffman, 1959), which aims to achieve socially and rhetorically desirable goals such as maintaining reputation (Schlenker and Britt, 1999; Zivnuska et al., 2004) or controlling the degrees of prominence to the information conveyed (Brennan et al., 2010). With respect to argumentation and persuasive discourse, it has been long noted that argumentative success is contingent upon the way arguers manage and project their identities (Kline, 1987). To better understand aspects of this identity and how they affect/are affected by social, political, and ethical factors, it is necessary to examine the communicative processes by which demographic parameters arise and the range of their impact.

Online platforms have indubitably become a predominant setting for social and public discourse. In such settings, explicit disclosure of one's own demographics ("As a millennial...", "As a black woman...") is a widespread phenomenon. The act of revealing one's identity in such a setting is a deliberate communication tool prompted by two incentives: 1) establishing their credibility with respect to the issues at hand ("I can talk about this, because..."), with the aim of strengthening their contribution to the discourse; or 2) establishing themselves as part of a group collective and thus implying implicit support to the speaker, naturally particularly useful when the explicit disclosure is that of a minority collective. Unavoidably, this mechanism yields advantages, as mentioned, as well as consequences: vulnerability. Once the demographic property is expressed as a credential or as collaborative support, this property is open to rebuttal, scrutiny, and even bias in the discourse interaction (Sap et al., 2020; De Candia et al., 2022). However, a thorough study of the role and effects of self-disclosures on online discourse dynamics has yet to be performed.

In this paper, we investigate particular self-disclosures that explicitly mention *author gender* and their influence on interaction dynamics in the online community. Such disclosures provide ground truth about the authors of the posts, as well as a sample of instances with explicitly established one's social identity in discourse. Moreover, they enable an analysis of how explicit demographic signals influence reactions from the community. If a particular post contains an explicit gender mention, "I am a woman...", will it spark more replies?

[*] These authors contributed equally.

LREC-COLING 2024, pages 14606–14621
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

How will the forum community receive the post, with appreciation or with criticism? Will others feel triggered to also reveal their own identities and will their reactions depend on whether they share the same gender as the author of the post?

We investigate this phenomenon in an online forum targeted at persuasion, the `/r/ChangeMyView` subreddit (CMV). In this forum, users post their stances and opinions about a specific issue (e.g., prisons should provide decent living conditions) and invite others to challenge their perspective, effectively changing their view (c.f. Figure 1). Compared to shorter social media texts (e.g., Twitter), CMV posts are representative of everyday discourse (longer and yet informal text), thus allowing for the identification of a broader set of distinctive features. Moreover, the forum structure allows the investigation of user interactions, involving both the communicative goal of the speaker/source and the impact of the message on its addressees. We examine the context and effect of a user explicitly disclosing their gender in their arguments, focusing on three research questions:

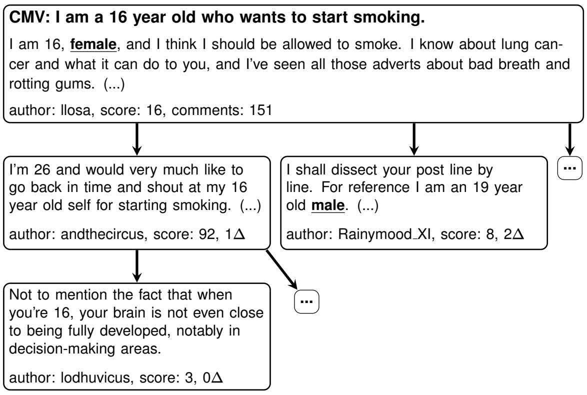
RQ1: When do people disclose their gender?
We analyze the distribution of gender self-disclosures across automatically extracted topics (cf. Section 4).

RQ2: How does the forum community react to gender disclosures?
We implement a regression analysis of the modulation of forum activity (votes, replies) triggered by the self-disclosures (cf. Section 5).

RQ3: What are the stylistic features of posts with self-disclosures, and how do they impact the forum community?
We integrate linguistic (i.e., readability metrics, relative frequency of POS tags) and pragmatic (i.e., toxicity, sentiment) features into the analysis discussed for RQ2 (cf. Section 6).

We find that although self-reported gender appears across diverse topics and with different communication purposes, it strongly affects the community reaction and composition, even contributing to the negative judgment of arguments.

The contributions of our work are twofold. At the level of the investigated phenomenon, we fill a research gap in the understanding of the impact of gender self-disclosures on online interactions. At the resource level, we release an extension of a reference dataset (Tan et al., 2016) with multiple annotation layers previously not available, most notably self-disclosed gender mentions (automatically extracted and manually checked), linguistic and pragmatic features, and topics, establishing a foundation for research topics within Argument Mining, NLP, and Computational Social Science.



CMV: I am a 16 year old who wants to start smoking.

I am 16, **female**, and I think I should be allowed to smoke. I know about lung cancer and what it can do to you, and I've seen all those adverts about bad breath and rotting gums. (...)

author: llosa, score: 16, comments: 151

I'm 26 and would very much like to go back in time and shout at my 16 year old self for starting smoking. (...)

author: andthecircus, score: 92, 1Δ

I shall dissect your post line by line. For reference I am an 19 year old **male**. (...)

author: Rainymood_XI, score: 8, 2Δ

Not to mention the fact that when you're 16, your brain is not even close to being fully developed, notably in decision-making areas.

author: lodhuvicus, score: 3, 0Δ

Figure 1: Example discussion tree from the CMV$_T$ dataset: original post (top) and three selected comments (out of 151).

2. Related Work

This work lies at the intersection between two prominent lines of research in NLP: a) the one which targets the identification of user demographics and the corresponding textual modulation; b) research in computational argumentation, which aims at identifying the textual features of effective arguments. Below, we first summarize both research directions and then identify a clear research gap at their crossing.

Socio-demographic attributes: extraction and analysis The socio-demographic turn in NLP (Hovy, 2018) comes with a clear data need: mapping text to demographic attributes (e.g., age, gender, race) as well as other relevant features (e.g., personality traits). Socio-demographics are used to create representations that are specific to users or their groups (Plepi et al., 2022) and analyzed from the perspective of their influence on NLP models (Hung et al., 2023; Lauscher et al., 2022a).

This work takes a different direction, as we focus on the impact of gender mentions on a broad audience and in connection with their stylistic properties. More similarly, Voigt et al. (2018) explore how differential responses to gender can be measured and analyzed, and Aggarwal et al. (2020) perform a comparative analysis between male and female language in a set of COVID-themed subreddits and topical preferences. The findings corroborate assumptions on (i) distinctions along emotional dimensions between the two genders, demonstrating that these differences are amplified in emotionally-intensive discourse, and (ii) gender-related topical preferences. De Candia et al. (2022) explore the social norms and factors involved in community judgment using data from the r/AITA subreddit, focusing on the age and gender of the author and post topics. The results describe a clear trend: older and male authors receive significantly more negative judgment from the community than younger and female authors.

In line with our methodology, Plepi et al. (2022) map texts to author's demographics in a rule-based fashion, based on explicit mentions (Welch et al., 2020). A comparable methodology is also the one employed by Gjurković et al. (2021) who employ Reddit flairs (self-assigned tags, which are specific to subreddits) as well as explicit mentions, to assign users to personality traits (and other demographic attributes); as a next step, Jukić et al. (2022) investigate topic modulation in relation to these attributes. An important feature shared by all these studies is the projection of the gender information to all posts of the same user.

Discourse dynamics in online persuasive forums Tan et al. (2016) study the mechanics of persuasion using discussions extracted from CMV. The authors study which interaction dynamics are associated with a successful change of opinion, finding that the most consequential variable is *when* a community member joins the discussion. The authors additionally explore differences in language used in counterarguments and linguistic properties of more persuasive arguments. Wiegmann et al. (2022) study the qualities that define the success of debaters (persuasion) in CMV discussions taking into account stylistic aspects and linguistic features. While not focused on persuasion *per se*, our work fits into this line of research by targeting the reaction of the community (replies, likes) to self-disclosures.

Research Gap The variables that impact persuasion in CMV, or similar online settings, have gained attention in recent years (Tan et al., 2016; Morio et al., 2019; Egawa et al., 2020; Dayter and Messerli, 2022). However, no dataset relevant for the task links to the effects of gender, or other demographic properties, *as explicitly expressed by the author* or overtly perceived by the community. Additionally, while some recent research has explored the potential impact of bias in Argument Mining (Spliethöver and Wachsmuth, 2020; Jakobsen et al., 2021; Manzoor et al., 2022), there remains a lack of analysis with respect to how the inclusion of (explicit) demographics by members of the online argument forums effect the argumentation discourse. Finally, most resources indeed annotated for author gender rely on induced gender, e.g., predicted from first names or profile pictures (Verhoeven et al., 2016) or cross-referenced online or through user profiles (see, for example, Voigt et al. (2018)); this contradicts the current understanding of gender as a social concept, and does not consider the intended or unintended impact of self-disclosures.

3. Data and Annotation

Our initial dataset originates from Tan et al. (2016). It encompasses a large collection of discussions sourced from the /r/ChangeMyView subreddit (we refer to this dataset as **CMV**$_T$). Each discussion starts with an Original Post (OP) in which its author expresses their view (c.f. Figure 1) and continues with arguments supporting or opposing the original stance. If a particular reply is especially convincing, it can be rewarded a Δ – clue that the comment made somebody change their mind. CMV$_T$ consists of 20.626 posts and 1.258.035 comments, organized in discussion trees.[1] Each node (OP or comment) includes user name of its author, subsequent replies, and score, i.e., the number of "up" votes for OPs and the sum of "up" and "down" votes for comments.[2] While some of the user names are gendered (i.e., u/mysteryman403), most are ambiguous and do not reveal any demographic features of the authors (e.g., top three CMV posts from 29th of January 2024 were written by u/that_person_658, GardenOrca, and TaoHumor). Therefore, in this work, instead of relying on user names, we focus on manually checked explicit self-disclosures, which provide us with a more robust signal about the authors.

We extract two types of features directly from CMV$_T$ (further referred to as **CMV Features**): (1) scores, associated with the community's overall appreciation for the post, and (2) the total number of comments to the OP, corresponding to the notion of quantity of interaction by the community. We further enrich the dataset with manually (Section 3.1) and automatically predicted (Sections 3.2 and 3.3) information.

3.1. Explicit Gender Mentions

CMV$_T$ comprises over a million comments. Given the impracticality of manually annotating each of them, we automatically filter potential self-reports of gender and then manually annotate them.

Filtering heuristics Based on multiple online sources, such as gender.fandom.com and www.gendergp.com, we collected a list of 83 gender-identity expressions (e.g., agender, man,

[1] Tan et al. (2016) split the data into training and heldout parts, depending on the timestamp of the post. Since we do not perform any analysis over time, we merge the two parts and process them jointly.

[2] For ethical reasons, API of CMV imposes that scores of OPs take only "up" votes into account.

gender questioning, girl).[3] Next, we filtered contexts in which these expressions appear in CMV$_T$ and manually selected 48 phrases that people used to self-report (e.g., “me as a”, “I am”, “I identify as”). Finally, we designed a simple grammar to recognize mentions of the user’s gender. For example, the most frequently matched rule was:

mention → context feature* gender

where `gender` and `context` are non-terminals representing the two sets of expressions described above, and `feature` matches descriptions of human properties (e.g., married, educated, old) that we found in the dataset.

Manual annotation The above-described heuristics filtered 422 posts and 3784 comments with potential gender self-disclosures. In the next step, each of them was manually reviewed by two annotators.[4] They were presented with the entire post/comment and, with the gender-related phrase highlighted, asked to determine whether the author indeed mentioned their gender.

Initially, the annotators disagreed on 19 posts and 371 comments. After manual investigation of these cases, we found that Annotator 2 performed the task with a much higher level of detail, identifying multiple errors of the automatic filtering tool that Annotator 1 overlooked. Among the most common errors were texts with quoted speech (72 cases, e.g., “They don’t want to hear ‘I’m a nonbinary [...]’”), hypothetical situations (45 cases, e.g., “If I am female”), and mentions that crossed sentence borders (15 cases, e.g., “No, I’m an atheist. Male/female is not determined by a god [...]”). The disagreements were resolved manually by an expert – in all the cases, their decision aligned with Annotator 2.

Statistics Table 1 presents the statistics for the final manual annotations (henceforth **CMVGender**). In total, the dataset contains 396 OPs and 3,235 comments with explicit mentions of the author’s gender. Interestingly, these posts are distributed across more than 1.8k discussions and originate from almost 2.5k distinct authors. This distribution indicates that the self-disclosures do not cluster in only a few discussions, but rather cover a broad part of CMV$_T$.

The dataset includes self-disclosures of seven gender identities, listed here by frequency: male, female, transgender female, transgender (without explicit female/male markers), transgender male, genderqueer, and non-binary. The statistics show that CMVGender is male-skewed, with only 24% of posts and 40% of comments written by people of other genders. This result aligns with the general Reddit audience profile, which, according to the reports by Statistica, is in 63.8% male.[5]

	Posts	Replies	Discussions	Authors
male	299	1,953	1,357	1,640
female	89	961	693	674
trans female	4	152	97	76
transgender	2	73	60	58
trans male	1	47	32	27
genderqueer	1	30	24	22
non-binary	0	19	16	16
Total	396	3,235	1,812	2,456

Table 1: Frequency of gender self-disclosures; discussions – unique threads with at least one explicit mention of gender.

3.2. Textual Features

In order to broaden the scope of research questions that can be explored using CMVGender (e.g. Voigt et al. (2018)) and to make it more accessible to fellow researchers, we enrich the whole CMV$_T$ with various layers of automatically generated textual features (see Table 2 for an overview). In total, 85 new features are incorporated.[6]

We first annotated all OPs and comments with the features reported in Falk and Lapesa (2022). These features are of interest as the set consists of a comprehensive range of linguistic, stylistic and pragmatic attributes, from part-of-speech frequencies to lexical sophistication and sentiment scores.

We then incorporated textual features described in Tan et al. (2016), which the authors use to examine which linguistic aspects impact the persuasiveness of arguments. These features are not part of the publicly available CMV$_T$. While many overlap with the features of Falk and Lapesa (2022), others needed to be computed, such as the use of definite or indefinite articles and the frequency of website links.

Finally, to consider the more social and interactive aspects of the text, we implemented toxicity and sentiment analysis (positive, neutral, negative) classifiers. To annotate the toxicity scores, we implemented the RoBERTa-based toxicity classifier presented in Dale et al. (2021), fine-tuned on an Jigsaw dataset[7] with sentence-level toxicity labeling. Sentiment scores were extracted using a

[3] We release all developed code and data at https://github.com/emvecchi/bias_in_am.

[4] Annotators are MSc students specializing in Computational Linguistics.

[5] www.statista.com/statistics/1255182/distribution-of-users-on-reddit-worldwide-gender/, report from January 2022.

[6] The full set of textual features is provided in Appendix Table 5.

[7] www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data

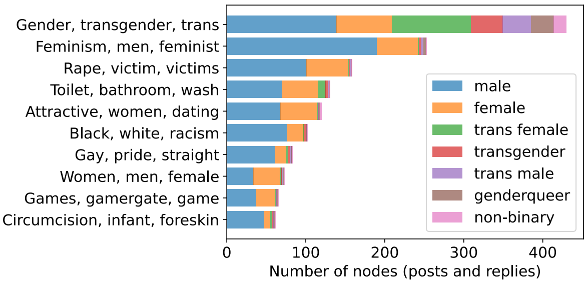


Figure 2: Most common topics of posts with explicit mentions of gender.

RoBERTa-based model,[8] which was fine-tuned for sentiment analysis with the TweetEval benchmark (Barbieri et al., 2020) and outperformed a large set of models in the task of social media sentiment analysis (Loureiro et al., 2022).

3.3. Topics

Finally, to gain insight into the content of the CMV$_T$ discussions, we enrich the dataset with automatically predicted topics. Concretely, we apply BERTopic (Grootendorst, 2022) with default parameters and annotate original posts with their subjects. The procedure results in 264 topics.

4. When Do People Explicitly Mention their Gender?

As discussed in the Introduction, gender self-disclosure via an explicit textual mention is a specific communication move. To grasp a better understanding when CMV users use such a move in persuasive texts, we start from studying topics and examples of posts with gender self-reports.

Topics Figure 2 displays the ten most frequent topics among the posts and replies with explicit gender mentions. A large majority of these topics relate to gender (e.g., transgender, feminism, LGBTQ), specific situations in which gender plays a role (e.g., rape, toilet use), or very specific activities or practices (e.g., gaming, circumcision). The race topic is mid-ranked, and it probably stems from co-occurrence with gender report, given that it is very common to package gender and race when characterizing one's own identity (i.e., "I am an Asian woman", "I am a white man"). Interestingly, the two most frequent gender identities – male and female – appear across all the topics. In other words, we did not find a topic that only one of these groups would address.

Function of self-disclosures Now that we know in which topics users commonly mention their gender, we discuss what the function of such a move can be. The literature on computational argumentation provides a set of potential candidates. Falk and Lapesa (2022) investigate roles of personal reports in argumentation, one of which is establishing the speaker's credibility with respect to the topic at issue. Such a function of gender self-reports can be observed in examples (3) and (4) in Table 3. Interestingly, however, self-disclosures can also serve a complementary function to establish speakers' credibility, that of an implicit rebuttal (see example (6)). Such a rebuttal is a core component in most popular argument models (Habernal and Gurevych, 2017): when constructing an argument, people often explicitly mention a possible counter-argument they expect from their opponents and address it, effectively pre-empting it. In this case, for example, a male speaking about feminism may want to disclose his identity to weaken possible related counter-argument ("You are not a victim of gender discrimination, so your opinion does not count") just by showing that they are aware of being in the majority group, but that this does not weaken their opinion. While interpreting this type of self-disclosure as a rebuttal may be subject to discussion, they minimally serve as a signal of hedging (Medlock and Briscoe, 2007), as in example (1).

How about women speaking against undesirable features of feminist rhetoric, as in example (2)? In this case, self-disclosure strengthens the argument because the speaker discloses a position at odds with the majority of the group to which they belong. Such a function is a combination of a credibility ("I am a woman, and I can talk about this") with a concession ("Even if this may go against my immediate interest...") (Musi, 2018).

5. Who Answers to Whom?

In the previous section, we saw that self-reported gender disclosures can appear across diverse topics and serve different communication purposes. We now shift our focus to RQ2: how do these mentions influence the community? Our hypothesis is that disclosing explicit gender information within a CMV post can have immediate and far-reaching implications, potentially influencing readers and the composition of community interacting. Moreover, the self-defined social identity presented might remain in the memory of the community, shaping their responses to subsequent comments made by the same author, even in cases where this author refrains from mentioning their gender again. In some exceptional cases, readers may even recall an author's gender from previous discussions, as

[8] www.huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest

CMV Features		value
score	sum of “up” and “down” votes for the post	$[0, 1618]$
num comments	total # of comments to the post	$[1, 2777]$
avg comment score	average scores of all comments to post	$[-0.92, 15.34]$
Author Gender Features		value
CMVGender	explicit mentions of author’s gender in post	M, F
extCMVGender	extended author gender information (cf. Section 5)	M, F
CMVGender in comments [m|f]	% of comments with explicit mentions of gender	%, [male|female]
extCMVGender in comments [m|f]	% of comments where extCMVGender is known	%, [male|female]
gender source	author gender was explicitly mentioned (1) or extended (0)	$\{0, 1\}$
Textual Features		source
syntactic features	parts of speech	Falk and Lapesa (2022)
surface features	length, word complexity, readability	Falk and Lapesa (2022)
lexical diversity	variants of the type/token ratio, less sensitive to text length	Falk and Lapesa (2022)
lexical sophistication	based on word/co-occurence information	Falk and Lapesa (2022)
sentiment features	based on sentiment, social-positioning and cognition dictionaries	Falk and Lapesa (2022)
CMV_T lexical features	additional lexical features previously used for persuasiveness	Tan et al. (2016)
toxicity classifier	[neutral, toxic] probability scores for the post	Dale et al. (2021)
sentiment classifier	RoBERTa-based model fine-tuned for sentiment analysis	Loureiro et al. (2022)

Table 2: Overview of features examined in Section 5 and 6.

	Topic	Post
(1)	Gender, transgender, trans, sex	CMV - I don’t think people should be able claim they want to be referred to as he/her or they/them. Personally, I am a straight **cisgender male**, but I do have many friends who are LGBTQ. (...)
(2)	Feminism, men, feminist, women	CMV: I think the feminist movement was detrimental to society. Firstly I’d just like to point out that I am **female**. Secondly I’d like to clarify that I’m all for equality between all people. However, (...)
(3)	Rape, victim, victims, raped	If I was raped or sexually assaulted I probably wouldn’t report it... Care to CMV? Preface: I’m a 23 year old **woman**. I believe that my life would be far worse off if I reported a rape to if I didn’t. (...)
(4)	Toilet, bathroom, wash, seat	CMV - public toilets should be unisex It really grinds my gears when I have to wait to use the toilet when it is clear that there are perfectly good toilets for the other gender unoccupied. I’m **a guy**, and this must be an even bigger issue for the ladies.
(5)	Black, white, racism, racist	CMV: The ”Model Minority” and ”Positive” Asian Stereotypes are Dangerous and Racist People don’t believe Asian stereotypes are harmful. We are the “model minority” (...) Edit: I am an Asian **woman**.
(6)	Feminism, men, feminist, women	I think that feminism currently uses hate speech as a way to advance its goals. In fact, this attitude hurts the advancement of women. CMV I’ll start by saying I’m 26 **male**.

Table 3: Examples of posts with explicit gender mentions.

	Posts	Comments	Discussions	Authors
male	2,253	227,261	18,515	1,634
female	396	53,042	10,119	664
Total	2,649	280,303	19,016	2,298

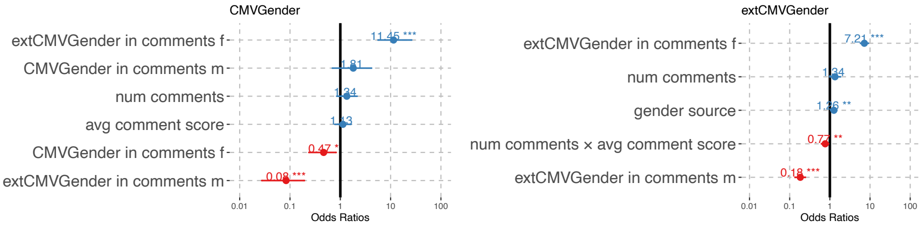
Table 4: Frequency of posts written by the authors who self-reported their gender in CMV_T.

their IDs are unique throughout the CMV dataset.

Method We examine which variables indicating type of interaction (i.e. quantity and degree of post appreciation) and community composition (i.e. amount of males and/or females in the comment community) are explanatory in a binary (male/female) logistic regression task.[9] We first focus on the self-reported instances (CMVGender) to determine how the definition of one’s identity, as an intentional communication mechanism, corresponds to these features. Next, to determine the extended impact that one’s disclosed identity has on CMV_T users, we expand the author’s gender information to other posts (henceforth, **extCMVGender**). In simpler terms, if an author explicitly identifies as a female in one post or comment (cf. Section 3.1), we assume that all other posts by the same author were also written by a female. We exclude 14 authors who provided varying gender information across different posts from this analysis. For detailed statistics on posts and comments from authors who mention their gender at least once, please refer to Table 4. Finally, we investigate which properties of the community interaction most influence the community’s appreciation for a post, quantified by score.

We therefore examine three dependent variables (DVs): CMVGender, extCMVGender, and score; while the independent variables (IVs) are the collection of CMV Features and additional Author Gender Features from Section 3 (with all pairwise interactions). Step-wise model selection was

[9] Due to data sparsity, our analysis in Sections 5 and 6 is limited to male and female genders.



(a) CMVGender (Pseudo $R^2 = 52\%$) (b) extCMVGender (Pseudo $R^2 = 58\%$)

Figure 3: **Gender $\sim$ CMV Features.** Standardized beta values for significant $(Pr(|z|) < 0.05)$ terms for each selected logistic model. Positive beta values correspond to higher feature values for females; negative beta values correspond to higher feature values for males.

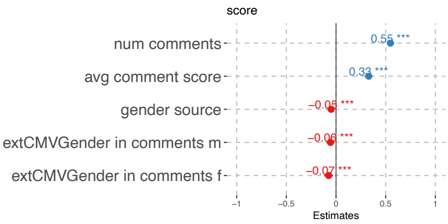


Figure 4: **score $\sim$ CMV Features.** $(R^2 = 64\%)$ Standardized beta values for significant $(Pr(|t|) < 0.05)$ terms for the most explanatory linear model.

implemented via AIC, in both directions.[10] The forest plots in Figures 3 and 4 report the standardized beta values of significant IVs, as well as the adjusted R^2 values for each model.[11]

Results The use of explicit mentions of gender in a post strongly defines the community that interacts with this post and the quality of that interaction. More specifically, in Figure 3a, we find that the explicit mention of female gender in a post is attributed to a larger female population among the comment authors (extCMVGender in comments f), however, it is also indicative of many comments by male authors who explicitly mention their gender (CMVGender in comments m). A similar trend is seen for explicit mentions of male authors. This finding indicates a notion of solidarity or shared interest between the post and the authors of the comments on the one hand, while on the other shows that **comment authors use the explicit mention of their gender to counter the explicit gender of the OP author** – a tactic that ascribes self-assignment of gender as a credential in the argument-counterargument discourse. Both the quantity (num comments) and quality of community interaction (avg comment score) are both significantly higher when a female author explicitly discloses their gender.

When considering the extended effects of attribution of gender (i.e. extCMVGender, cf. Figure 3b), we find distinctive patterns in interaction behavior based on author gender. Most notably, male authors predominantly engage with posts made by fellow males, even when not explicitly mentioned, and likewise, female authors tend to interact more with posts from female counterparts. Additionally, a notable predictor for female authorship is a higher amount of community interaction (num comments). Intriguingly, the interplay between the extent of community engagement with the post and the appreciation of comments (i.e. the interaction between num comments and the average comment score), a stronger likelihood of male authorship emerges with higher values. Importantly, our model exhibits robust predictive power ($R^2 = 58\%$) even in the absence of textual or stylistic features, underscoring the significance of gender-based behavioral patterns in these online interactions.

The results in Figure 4 address which factors influence community appreciation for a post. The quantity of community interaction and community appreciation for each post, measured by num comments and the average comment score respectively, emerge as the strongest predictors for a higher score. Moreover, posts garnering a higher score are often associated with comments where the author gender is unknown, as evidenced by negative values for both male and female extCMVGender in comments. Undoubtedly, the question of why people decide to self-disclose their gender is multifaceted, incorporating aspects such as controversiality, topic, and personal preferences. This finding, for example, suggests that disclosures of gender might, more often, be in combination with controversial statements. When it comes to controversiality, we consider a range of possibly associated textual features, such as toxicity, use of

[10] The analysis code is made available here: https://github.com/emvecchi/bias_in_am/tree/main/scripts/analyze.

[11] We used the StepAIC R package. A full set of terms for selected models, as well as a full model summary, is provided in Appendix Table 6.

negations, and sentiment components, such as polarity, certainty, politeness, and respect (cf. Section 6). Interestingly, the presence of explicit gender mentions in posts (gender source) was linked to a *decrease* in post appreciation, indicating that **self-disclosure of gender identity negatively impacts the community's assessment of the post**.

In the next section, we incorporate the effects of textual features. We observe improvements (higher R^2) across gender-predicting models with the inclusion of Textual Features, showing that encompassing stylistic and pragmatic attributes raises the predictive capacity for author gender. The finding that males engage predominantly with comments authored by males and females similarly engage more with comments authored by females, as discussed for Figures 3a and 3b, remains a consistent and the strongest predictor for both models.

6. What Style Elicits What Answer?

Now that we know that the self-reported gender of the author has a strong influence on the interaction that their post receives, we shift to the final research question: which stylistic and pragmatic features affect this reaction? For this, we consider the Textual Features in addition to those examined in the previous section.

Method Given the large number of Textual Features (85), we run a correlation analysis to remove features potentially contributing to co-linearity, which can likely distort the performance of the model (Falk and Lapesa, 2022). We cluster the features based on their Spearman correlation and select only the first of each sub-cluster with a correlation higher than a threshold of Spearman ≥ 0.5. The remaining features are a set of 62 variables.

We then implement the regression models from Section 5, including the selected Textual Features among the IVs and running a step-wise AIC model selection for each DV.[12]

Results In estimating author gender (cf. Figures 5a and 5b), we (i) confirm and, in some cases, reinforce the previous section's findings, and (ii) find that distinct linguistic patterns play a significant role. In CMV discussions, male authors tend to exhibit a writing style characterized by higher toxicity scores and frequent use of nouns denoting certainty, reflecting a sense of assurance. Additionally, they employ higher hypernymy scores, indicative of semantically rich texts, and use failure/power-loss verbs and questions more frequently, a tool often used to engage with readers or challenge perspectives. On the other hand, female authors demonstrate a writing style marked by higher lexical decision accuracy scores and more frequent bigrams, implying clarity and predictability in language use. They also utilize trust verbs, incorporating positive emotions, and employ words related to the economy, indicating potential themes or topics of interest.

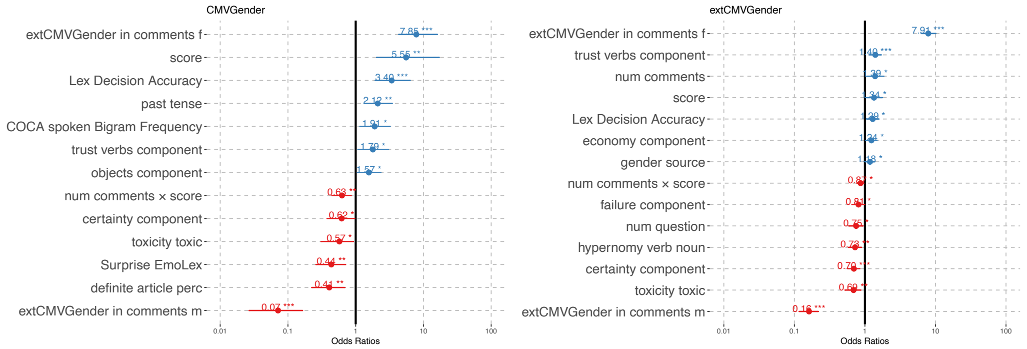
In Figure 6, in addition to our previous results outlining the impact of gender, we find that higher-scoring posts exhibit features like more familiar language, longer text lengths, diverse semantic content, and words of surprise, indicating that engaging, varied, and unexpected language garners more community approval. Conversely, lower-scoring posts relate to emotional and affective language, particularly regarding personal feelings or social relationships, resulting in reduced community appreciation. Most importantly, overt expressions of personal or empathetic qualities have a similar community impact as explicit mentions of gender; **suggesting that writing in a style that is more significantly associated with a particular gender, such as openly displaying personal or empathetic characteristics, can lead to similar consequences in argument appraisal as explicitly stating one's gender identity**.

While traits typically associated with females, like emotional language (Aggarwal et al., 2020), contribute significantly to lower scores (Figure 6), the male-associated feature with a notable impact on score is "Surprise EmoLex" (cf. negative values in Figure 5a), indicating the use of lexicon depicting surprise. Interestingly, posts featuring this trait tend to score higher. However, it's crucial to note that while female-associated features impacting this variable align with previous literature, the extent to which higher scores are associated solely with male traits, like surprise, remains uncertain. Future research will be needed to explore whether emotional or empathetic language, regardless of gender, consistently leads to lower scores, and whether this is influenced by gender-specific language use.

7. Discussion and Conclusion

This work focused on a particular communication phenomenon – explicitly disclosing information about one's gender – and its influence on community interaction in an online persuasive forum. Through a qualitative analysis, we first established that from the user's perspective, including information about their gender can have a variety of intentional and valuable persuasive functions, such

[12] Given the complexity of the analysis and the high number of predictors, including interactions in the analysis is left for future work. Full details are provided in Appendix Table 7.



(a) CMVGender (Pseudo $R^2 = 58\%$)

(b) extCMVGender (Pseudo $R^2 = 61\%$)

Figure 5: **Gender $\sim$ Combined CMV and Textual Features.** Standardized beta values for significant ($Pr(|z|) < 0.05$) terms for each selected logistic model. Positive beta values correspond to higher feature values for females; negative beta values correspond to higher feature values for males.

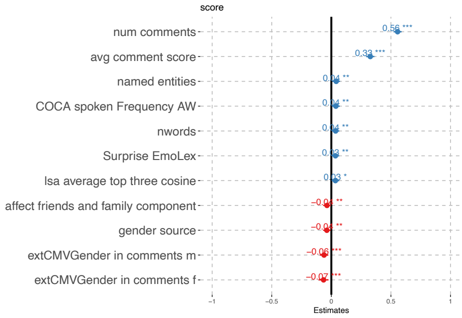


Figure 6: **score $\sim$ Combined CMV and Textual Features.** ($R^2 = 60\%$) Standardized beta values for significant ($Pr(|t|) < 0.05$) terms of the linear model.

as increasing their credibility or weakening possible related counter-arguments. However, such a move might have a broad range of *unintentional* repercussions. We found that posts with explicit mentions of gender will receive significantly more reactions from users of the same self-reported gender. Moreover, explicit mentions of gender significantly produce countered gender mentions in the comments – a tactic that forces gender as a credential in the argument-counterargument discourse. Finally, from the perspective of the community's appreciation, we found that revealing one's gender identity can *lower* the rating of the post. Interestingly, this effect goes beyond explicit mentions of gender and can also be observed when analyzing the content of the posts. Since there is a significant interconnection between the gender of the authors and their writing styles, simply demonstrating a style that is more frequent for one gender, such as overt expressions of personal or empathetic qualities, has a similar community impact as explicit mentions of gender.

The findings of this research support prior studies and theories, particularly those outlined in Section 2 such as Voigt et al. (2018), Aggarwal et al. (2020), Plepi et al. (2022), and De Candia et al. (2022). They confirm gender-based disparities in persuasive discourse, responses to gender mentions, and community judgments on Reddit. From the more theoretical perspective, notably the Social Identity Theory (SIT, Tajfel, 1978; Tajfel and Turner, 1978) suggests individuals shape their identities based on social groups, with Self-Categorization Theory (SCT Turner et al., 1987) further emphasizing depersonalization effects, or eventual deindividualization (Postmes and Spears, 1998), leading to strengthened collective ("ingroup") identities in contexts like subreddit interactions, and relying on these identifications in the case of intergroup conflict. Our findings, indicating gender-skewed responses and the use of explicit gender disclosures as counterarguments, align with these theories, highlighting the role of ingroup definitions in intergroup conflicts.

Beyond the broader understanding of the impact of self-disclosures on CMV, this work contributes CMVGender – new layers of features for CMV$_T$ allowing for evaluating demographic, stylistic, and interactive properties in an online setting. The dataset can serve to better understand biases in Argument Mining (Spliethöver and Wachsmuth, 2020) and in tasks such as assessing Argument Quality. Additionally, the annotations of self-disclosed demographics can support collecting terms and attributes for explicit bias identification and removal (Barikeri et al., 2021; Holtermann et al., 2022). Finally, the self-mention data provides a ground-truth to anchor studies on moral foundations (Alshomary et al., 2022) and human values (Kiesel et al., 2022).

8. Ethical Statement

Our work deals with analyzing and predicting socio-demographic aspects from text, and how they impact discourse, which should be considered sensitive information. Predictive methods can result in potentially harmful applications, e.g., in the context of user profiling. We acknowledge this potential for dual use (Jonas, 1984) of the data sets we use. However, in this work, we are interested in advancing NLP research towards a better understanding of such fine-grained aspects of language and how they are already captured by our technology. We believe that these insights will lead us toward fairer and more inclusive language technology. In contrast, we explicitly discourage the prediction of sensitive attributes from text for harmful purposes.

In addition, we acknowledge that our work is limited as data scarcity issues force us to model gender as a binary variable, which does not reflect the wide variety of possible identities along the gender spectrum (Lauscher et al., 2022b). However, we are not aware of other suitable data sets without this limitation. We believe, however, that even the findings derived from a binary view on gender can provide an initial understanding of self-reported demographics impact discourse dynamics, and that any results will hold under a more sophisticated modeling of the problem.

9. Acknowledgements

We would like to acknowledge the reviewers for their helpful comments and Sebastian Padó for giving feedback on this work. Many thanks also to our student annotators who contributed greatly to this work (whom we do not mention here explicitly for anonymity reasons). This work is supported by Bundesministerium für Bildung und Forschung (BMBF) through the project E-DELIB (Powering up e-deliberation: towards AI-supported moderation) and 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).

10. Bibliographical References

Jai Aggarwal, Ella Rabinovich, and Suzanne Stevenson. 2020. Exploration of gender differences in covid-19 discourse on reddit. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

Milad Alshomary, Roxanne El Baff, Timon Gurcke, and Henning Wachsmuth. 2022. The moral debater: A study on the computational generation of morally framed arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8782–8797, Dublin, Ireland. Association for Computational Linguistics.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1941–1955, Online. Association for Computational Linguistics.

Niamh M Brennan, Caroline A Daly, and Claire S Harrington. 2010. Rhetoric, argument and impression management in hostile takeover defence documents. *The British Accounting Review*, 42(4):253–268.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46:904–911.

David Dale, Igor Markov, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Skoltechnlp at semeval-2021 task 5: Leveraging sentence-level pre-training for toxic span detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 927–934.

Daria Dayter and Thomas C Messerli. 2022. Persuasive language and features of formality on the r/changemyview subreddit. *Internet Pragmatics*, 5(1):165–195.

Sara De Candia, Gianmarco De Francisci Morales, Corrado Monti, and Francesco Bonchi. 2022. Social norms on reddit: A demographic analysis. In *Proceedings of the 14th ACM Web Science Conference 2022*, pages 139–147.

Ryo Egawa, Gaku Morio, and Katsuhide Fujita. 2020. Corpus for modeling user interactions in online persuasive discussions. In *Proceedings* *of the Twelfth Language Resources and Evaluation Conference*, pages 1135–1141, Marseille, France. European Language Resources Association.

Neele Falk and Gabriella Lapesa. 2022. Reports of personal experiences and stories in argumentation: datasets and analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5530–5553.

Matej Gjurković, 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.

Erving Goffman. 1959. *Presentation of Self in Everyday Life*. Doubleday Anchor Books.

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

John J Gumperz. 1982. *Language and social identity*. Cambridge University Press.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

Carolin Holtermann, Anne Lauscher, and Simone Ponzetto. 2022. Fair and argumentative language modeling for computational argumentation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7841–7861, Dublin, Ireland. Association for Computational Linguistics.

Dirk Hovy. 2018. The social and the neural network: How to make natural language processing about people again. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 42–49, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Chia-Chien Hung, Anne Lauscher, Dirk Hovy, Simone Paolo Ponzetto, and Goran Glavaš. 2023. Can demographic factors improve text classification? revisiting demographic adaptation in the age of transformers. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1565–1580, Dubrovnik, Croatia. Association for Computational Linguistics.

Terne Sasha Thorn Jakobsen, Maria Barrett, and Anders Søgaard. 2021. Spurious correlations in cross-topic argument mining. In *Proceedings of* SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 263–277.

Hans Jonas. 1984. *The imperative of responsibility: In search of an ethics for the technological age*. University of Chicago press.

Josip Jukić, Iva Vukojević, and Jan Snajder. 2022. You are what you talk about: Inducing evaluative topics for personality analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3986–3999, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the human values behind arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4459–4471, Dublin, Ireland. Association for Computational Linguistics.

Susan L Kline. 1987. Identity management in argumentative discourse. *Argumentation: Perspectives and approaches*, pages 241–251.

Anne Lauscher, Federico Bianchi, Samuel R. Bowman, and Dirk Hovy. 2022a. 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.

Anne Lauscher, Archie Crowley, and Dirk Hovy. 2022b. Welcome to the modern world of pronouns: Identity-inclusive natural language processing beyond gender. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1221–1232, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. Timelms: Diachronic language models from twitter. In

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 251–260.

Emaad Manzoor, Yohan Jo, and Alan Montgomery. 2022. Status biases in deliberation online: Evidence from a randomized experiment on ChangeMyView. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6351–6363, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ben Medlock and Ted Briscoe. 2007. Weakly supervised learning for hedge classification in scientific literature. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 992–999, Prague, Czech Republic. Association for Computational Linguistics.

Gaku Morio, Ryo Egawa, and Katsuhide Fujita. 2019. Revealing and predicting online persuasion strategy with elementary units. 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 6274–6279.

Elena Musi. 2018. How did you change my view? a corpus-based study of concessions' argumentative role. *Discourse Studies*, 20(2):270–288.

Joan Plepi, Béla Neuendorf, Lucie Flek, and Charles Welch. 2022. Unifying data perspectivism and personalization: An application to social norms. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7391–7402, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tom Postmes and Russell Spears. 1998. Deindividuation and antinormative behavior: A meta-analysis. *Psychological bulletin*, 123(3):238.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490.

Barry R Schlenker and Thomas W Britt. 1999. Beneficial impression management: Strategically controlling information to help friends. *Journal of Personality and Social Psychology*, 76(4):559.

Maximilian Spliethöver and Henning Wachsmuth. 2020. Argument from old man's view: Assessing social bias in argumentation. In *Proceedings of the 7th Workshop on Argument Mining*, pages 76–87, Online. Association for Computational Linguistics.

Henri Tajfel and John C Turner. 1978. Intergroup behavior. *Introducing social psychology*, 401:466.

Henri Ed Tajfel. 1978. *Differentiation between social groups: Studies in the social psychology of intergroup relations.* Academic Press.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 613–624, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

JC Turner, MA Hogg, PJ Oakes, SD Reicher, and MS Wetherell. 1987. Rediscovering the social group: A self-categorization theory. basil blackwell.

Ben Verhoeven, Walter Daelemans, and Barbara Plank. 2016. Twisty: a multilingual twitter stylometry corpus for gender and personality profiling. In *Proceedings of the Tenth international conference on language resources and evaluation (LREC'16)*, pages 1632–1637.

Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. 2018. Rtgender: A corpus for studying differential responses to gender. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Charles Welch, Jonathan K. Kummerfeld, Verónica Pérez-Rosas, and Rada Mihalcea. 2020. Compositional demographic word embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4076–4089, Online. Association for Computational Linguistics.

Matti Wiegmann, Khalid Al Khatib, Vishal Khanna, and Benno Stein. 2022. Analyzing persuasion strategies of debaters on social media. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6897–6905, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Suzanne Zivnuska, K Michele Kacmar, LA Witt, Dawn S Carlson, and Virginia K Bratton. 2004. Interactive effects of impression management and organizational politics on job performance.

Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior, 25(5):627–640.

A. Textual Features: Overview of full set

Table 5 provides the full set of textual features introduced in Table 2. Syntactic, surface, lexical diversity, lexical sophistication, and sentiment features were extracted as in Falk and Lapesa (2022), while CMV$_T$ lexical features were defined as in Tan et al. (2016). Toxicity features were extracted using the toxicity classifier of Dale et al. (2021); sentiment features with the sentiment classifier of Loureiro et al. (2022).

feature name	explanation	type
adverbs	relative amount of adverbs in the text	syntactic
auxiliary	relative amount of auxiliary verbs in the text	syntactic
first person	relative amount of first personal pronouns in the text	syntactic
named entities	relative amount of named entities in the text	syntactic
past tense	relative amount of past tense verbs in the text	syntactic
subordinate conj	relative amount of subordinate conjunctions in the text	syntactic
flesch readability	flesch score based on average length of a sentence and average number of syllables per word	surface
Gunning Fog simplicity	weighted average of the number of words per sentence and number of long words (words with more than three syllables)	surface
nwords	raw frequency of words in the text	surface
mattr50 aw	Moving average type token ratio (50-word window)	diversity
mtld original aw	computes type token ratio of increased word windows / segments	diversity
token type ratio	for each word type, compute the probability of encountering one of it's tokens	diversity
All AWL Normed	relative amount of academic words	sophistication
Brysbaert Concreteness Combined AW	concreteness norms (Brysbaert et al., 2014)	sophistication
COCA spoken Bigram Frequency	academic bigram frequency scores	sophistication
COCA spoken Frequency AW	frequency scores of words in spoken language	sophistication
content poly	number of senses of content words	sophistication
Lex Decision Accuracy	Average lexical decision accuracy	sophistication
lsa average top three cosine	Natural log of mean LSA cosine of similarity between top3 contexts containing target words; reverses sign	sophistication
KL divergence rel entropy	Co-occurrence probability of word with 500 highly frequent context lemmas	sophistication
WN Mean Accuracy	Average naming accuracy	sophistication
action component	ought verbs, try verbs, travel verbs, descriptive action verbs	sentiment
affect friends and family component	affect nouns, participant affect, kin noun, affiliation nouns	sentiment
certainty component	sureness nouns, quantity	sentiment
economy component	economy words	sentiment
failure component	power loss verbs, failure verbs	sentiment
negative adjectives component	negative adjectives	sentiment
objects component	objects	sentiment
polarity nouns component	polarity nouns, aptitude nouns, pleasantness nouns	sentiment
polarity verbs component	polarity verbs, aptitude verbs, pleasantness verbs	sentiment
politeness component	politeness nouns	sentiment
positive adjectives component	positive adjectives	sentiment
positive nouns component	positive nouns	sentiment
positive verbs component	positive verbs	sentiment
respect component	respect nouns	sentiment
social order component	ethic verbs, need verbs, rectitude words	sentiment
trust verbs component	trust verbs, joy verbs, positive verbs	sentiment
virtue adverbs component	hostility adverbs, rectitude gain adverbs, sureness adverbs	sentiment
well being component	well-being words	sentiment
∗EmoLex	sentiment features for each emotion category	sentiment
Dominance∗	dominance terms based on ANEW	sentiment
pleasantness	sentiment features based on SenticNet	sentiment
attention	sentiment features based on SenticNet	sentiment
sensitivity	sentiment features based on SenticNet	sentiment
aptitude	sentiment features based on SenticNet	sentiment
polarity	sentiment features based on SenticNet	sentiment
hu liu ∗	positive/negative terms based on Hu and Liu (2004)	sentiment
Valence	valence terms based on ANEW	sentiment
hypernomy verb noun	Average hypernymy score for nouns and verbs (average for all senses, all paths)	sentiment
edu link count	amount of .edu links in the text	CMV$_T$
com link count	amount of .com links in the text	CMV$_T$
definite article perc	the	CMV$_T$
indefinite article perc	a, an	CMV$_T$
second person	relative amount of second personal pronouns in a text	CMV$_T$
first person pl	relative amount of first personal plural pronouns in a text	CMV$_T$
hedge words perc	maybe perhaps possibly potentially likely probably could might may can should	CMV$_T$
num question	text contains ?	CMV$_T$
num quotations	single or double quotations	CMV$_T$
example count	example, for ex, eg	CMV$_T$
entropy	based on word probabilities using NLTK's word frequency distributions	CMV$_T$
paragraph count	total number of sentences in text	CMV$_T$
sentence count	total number of paragraphs (newlines) in text	CMV$_T$
toxicity toxic	probability score of toxicity	toxicity
sentiment neutral	probability score of neutral sentiment of post	sentiment
sentiment negative	probability score of negative sentiment of post	sentiment

Table 5: Overview of textual features with short description and features type.

DV	Coefficients	Est.	Std. Err.	z|t	sig.	
CMVGender	(Intercept)	-1.373e+00	8.592e-01	-1.598	0.11007	
	num_comments	-7.399e-03	3.355e-03	-2.205	0.02742	*
	CMVGender_in_comments_m	2.307e+02	1.045e+02	2.208	0.02723	*
	CMVGender_in_comments_f	5.097e+01	2.199e+01	2.318	0.02042	*
	extCMVGender_in_comments_m	-1.267e+01	3.077e+00	-4.117	3.84e-05	***
	extCMVGender_in_comments_f	1.358e+01	5.344e+00	2.540	0.01107	*
	avg_comment_score	5.144e-01	1.923e-01	2.674	0.00748	**
	CMVGender_comments_m:extCMVGender_comments_m	-7.591e+02	3.468e+02	-2.189	0.02862	*
	CMVGender_comments_m:CMVGender_comments_f	-2.477e+03	1.480e+03	-1.674	0.09411	.
	extCMVGender_in_comments_f:avg_comment_score	-4.013e+00	1.493e+00	-2.688	0.00719	**
	extCMVGender_comments_m:extCMVGender_comments_f	2.826e+01	1.537e+01	1.839	0.06599	.
	num_comments:extCMVGender_in_comments_f	9.770e-02	3.184e-02	3.068	0.00215	**
	num_comments:CMVGender_in_comments_f	-1.061e+00	3.917e-01	-2.710	0.00672	**
	num_comments:CMVGender_in_comments_m	5.548e-01	2.915e-01	1.903	0.05702	.
	pseudo R^2					0.5208
extCMVGender	(Intercept)	-1.9990358	0.3608545	-5.540	3.03e-08	***
	num_comments	0.0049795	0.0018741	2.657	0.00788	**
	score	0.0028403	0.0018368	1.546	0.12202	
	extCMVGender_in_comments_m	-8.7120522	0.8484328	-10.268	$<$ 2e-16	***
	extCMVGender_in_comments_f	16.8665176	1.1418913	14.771	$<$ 2e-16	***
	avg_comment_score	0.0614874	0.0837085	0.735	0.46262	
	gender_source	1.0070499	0.3585474	2.809	0.00497	**
	score:extCMVGender_in_comments_f	0.0255225	0.0106531	2.396	0.01658	*
	extCMVGender_in_comments_f:gender_source	-4.6344153	2.0420524	-2.269	0.02324	*
	score:extCMVGender_in_comments_m	-0.0188109	0.0101315	-1.857	0.06336	.
	num_comments:avg_comment_score	-0.0012103	0.0004591	-2.636	0.00839	**
	pseudo R^2					0.5823
score	(Intercept)	-100.90317	9.52442	-10.594	$<$ 2e-16	***
	extCMVGender	-48.27249	14.60083	-3.306	0.000960	***
	num_comments	0.87351	0.03698	23.620	$<$ 2e-16	***
	CMVGender_in_comments_m	861.27211	343.84175	2.505	0.012318	*
	CMVGender_in_comments_f	1004.91427	579.92350	1.733	0.083257	.
	extCMVGender_in_comments_m	100.94041	22.76491	4.434	9.68e-06	***
	extCMVGender_in_comments_f	185.50420	49.73693	3.730	0.000196	***
	avg_comment_score	46.33698	3.26367	14.198	$<$ 2e-16	***
	gender_source	-21.78840	6.67599	-3.264	0.001116	**
	num_comments:CMVGender_in_comments_m	-18.59266	2.04252	-9.103	$<$ 2e-16	***
	extCMVGender_in_comments_m:avg_comment_score	-43.03104	9.30816	-4.623	3.99e-06	***
	num_comments:extCMVGender_in_comments_f	-0.57073	0.24994	-2.283	0.022495	*
	num_comments:extCMVGender_in_comments_m	-0.91275	0.13216	-6.906	6.40e-12	***
	CMVGender_in_comments_f:avg_comment_score	-431.79413	201.22059	-2.146	0.031987	*
	extCMVGender:avg_comment_score	25.36646	4.77216	5.316	1.17e-07	***
	extCMVGender_in_comments_f:avg_comment_score	-92.36578	20.45492	-4.516	6.63e-06	***
	extCMVGender:num_comments	-0.15794	0.04991	-3.164	0.001575	**
	CMVGender_in_comments_m:gender_source	488.67093	338.81106	1.442	0.149350	
	CMVGender_in_comments_m:avg_comment_score	-257.85256	157.69616	-1.635	0.102159	
	extCMVGender:CMVGender_in_comments_m	704.62572	467.67771	1.507	0.132037	
	num_comments:gender_source	0.06553	0.03947	1.660	0.096952	.
	R^2					0.6425

Table 6: **CMV Features as Independent Variables.** Summary of the most explanatory regression models for predicting CMVGender, extCMVGender, or score (DV) with estimates and statistical significance. Signif. codes: 0 '$***$' 0.001 '$**$' 0.01 '$*$' 0.05 '.' 0.1 ' ' 1

DV	Coefficients	Est.	Std. Err.	z|t	sig.	
CMVGender	(Intercept)	-2.484e+02	6.047e+01	-4.107	4.00e-05	***
	num_comments	1.490e-03	2.266e-03	0.658	0.510852	
	score	1.520e-02	5.190e-03	2.928	0.003412	**
	extCMVGender_in_comments_m	-1.465e+01	2.592e+00	-5.654	1.57e-08	***
	extCMVGender_in_comments_f	1.873e+01	2.846e+00	6.582	4.66e-11	***
	avg_comment_score	-3.390e-01	1.922e-01	-1.764	0.077757	.
	auxiliary	-2.552e+01	1.311e+01	-1.946	0.051595	.
	named_entities	-3.429e+01	1.645e+01	-2.085	0.037055	*
	trust_verbs_component	6.169e+00	2.951e+00	2.090	0.036588	*
	past_tense	4.126e+01	1.457e+01	2.832	0.004620	**
	certainty_component	-9.943e+00	5.364e+00	-1.854	0.063784	.
	COCA_spoken_Bigram_Frequency	1.197e-02	5.304e-03	2.257	0.023978	*
	Lex_Decision_Accuracy	2.617e+02	6.381e+01	4.102	4.10e-05	***
	objects_component	9.057e+00	3.874e+00	2.338	0.019398	*
	Anger_EmoLex	-2.150e+01	1.506e+01	-1.427	0.153457	
	Surprise_EmoLex	-1.015e+02	3.035e+01	-3.343	0.000828	***
	definite_article_perc	-6.081e+01	1.950e+01	-3.118	0.001823	**
	edited	-8.593e-01	5.181e-01	-1.658	0.097228	.
	toxicity_toxic	-2.188e+00	1.079e+00	-2.028	0.042599	*
	sentiment	-7.016e-01	4.251e-01	-1.650	0.098866	.
	num_comments:score	-2.675e-05	1.061e-05	-2.522	0.011684	*
	pseudo R^2					0.5801
extCMVGender	(Intercept)	-4.145e+01	1.880e+01	-2.205	0.027469	*
	num_comments	2.567e-03	1.041e-03	2.467	0.013632	*
	score	2.612e-03	1.105e-03	2.363	0.018120	*
	extCMVGender_in_comments_m	-1.051e+01	9.107e-01	-11.545	$<$ 2e-16	***
	extCMVGender_in_comments_f	1.844e+01	1.078e+00	17.106	$<$ 2e-16	***
	avg_comment_score	-1.025e-01	7.287e-02	-1.407	0.159438	
	gender_source	4.641e-01	2.309e-01	2.010	0.044406	*
	failure_component	-5.352e+00	2.721e+00	-1.967	0.049156	*
	well_being_component	-2.015e+00	9.988e-01	-2.018	0.043615	*
	trust_verbs_component	3.353e+00	9.795e-01	3.424	0.000618	***
	economy_component	1.154e+00	5.052e-01	2.285	0.022310	*
	polarity_verbs_component	6.783e-01	4.215e-01	1.609	0.107550	
	positive_verbs_component	-5.723e-01	2.237e-01	-2.558	0.010523	*
	certainty_component	-7.365e+00	2.022e+00	-3.643	0.000270	***
	hypernomy_verb_noun	-7.749e-01	2.559e-01	-3.028	0.002462	**
	Lex_Decision_Accuracy	4.849e+01	1.949e+01	2.488	0.012833	*
	Gunning_Fog_simplicity	-6.589e-02	3.958e-02	-1.665	0.095978	.
	Dominance	-3.092e-01	1.641e-01	-1.885	0.059450	.
	num_question	-1.430e-01	5.926e-02	-2.414	0.015793	*
	toxicity_toxic	-1.971e+00	6.947e-01	-2.837	0.004548	**
	num_comments:score	-6.494e-06	2.585e-06	-2.512	0.012011	*
	pseudo R^2					0.6111
score	(Intercept)	-2.474e+02	6.588e+01	-3.756	0.000177	***
	extCMVGender	1.339e+01	7.448e+00	1.798	0.072338	.
	num_comments	5.340e-01	1.418e-02	37.675	$<$ 2e-16	***
	extCMVGender_in_comments_m	-5.101e+01	1.320e+01	-3.865	0.000114	***
	extCMVGender_in_comments_f	-8.433e+01	2.411e+01	-3.498	0.000477	***
	avg_comment_score	3.247e+01	1.525e+00	21.296	$<$ 2e-16	***
	gender_source	-1.538e+01	5.498e+00	-2.798	0.005185	**
	positive_nouns_component	5.978e+00	4.215e+00	1.418	0.156213	
	named_entities	3.509e+02	1.197e+02	2.931	0.003413	**
	nwords	2.010e-02	7.546e-03	2.664	0.007767	**
	affect_friends_and_family_component	-3.940e+01	1.522e+01	-2.588	0.009718	**
	lsa_average_top_three_cosine	2.451e+02	9.819e+01	2.496	0.012632	*
	All_AWL_Normed	1.590e+02	9.591e+01	1.658	0.097476	.
	KL_divergence_rel_entropy	-3.683e+01	2.378e+01	-1.549	0.121594	
	COCA_spoken_Frequency_AW	6.469e-03	2.447e-03	2.644	0.008259	**
	mattr50_aw	9.190e+01	6.041e+01	1.521	0.128298	
	Surprise_EmoLex	5.158e+02	1.999e+02	2.580	0.009948	**
	Dominance	6.878e+00	3.837e+00	1.793	0.073179	.
	hu_liu_pos_nwords	-2.289e+02	1.386e+02	-1.652	0.098729	.
	hedge_words_perc	4.042e+02	2.325e+02	1.738	0.082273	.
	avg_comment_sentiment	4.347e+01	2.632e+01	1.652	0.098729	.
	R^2					0.6052

Table 7: **Combined CMV and Textual Features as Independent Variables.** Summary of the most explanatory regression models for predicting CMVGender, extCMVGender, or score (DV) with estimates and statistical significance. Signif. codes: 0 '$***$' 0.001 '$**$' 0.01 '$*$' 0.05 '.' 0.1 ' ' 1