Style-Shifting Behaviour of the Manosphere on Reddit

Jai Aggarwal **Suzanne Stevenson**

Department of Computer Science
University of Toronto
{jai, suzanne}@cs.toronto.edu

Abstract

Content warning: misogyny, profanity.
Hate speech groups (HSGs) may negatively influence online platforms through their distinctive language, which may affect the tone and topics of other spaces if spread beyond the HSGs. We explore the linguistic style of the Manosphere, a misogynistic HSG, on Reddit. We find that Manospheric authors have a distinct linguistic style using not only uncivil language, but a greater focus on gendered topics, which are retained when posting in other communities. Thus, potentially harmful aspects of Manospheric style carry over into posts on non-Manospheric subreddits, motivating future work to explore how this stylistic spillover may negatively influence community health.

1 Introduction

A concern for broad social media platforms is the harmful influence of hate speech groups (HSGs). This impact may be wide-reaching because HSG members may post in non-HSG communities, negatively affecting their health (Habib et al., 2022). A potential adverse influence is the distinctive language used by HSGs, which has been characterized in terms of its toxicity or radicalization traits (Ribeiro et al., 2021; Habib et al., 2022). However, other aspects of language may also be distinctive of HSGs, and may negatively affect the tone and topic of discussion in other spaces if spread beyond those groups. Thus, assessing the influence of HSGs requires a more comprehensive understanding both of their linguistic style, and of whether and how that style is used by their members outside of the HSG communities.

Questions concerning how language reflects group membership and how speakers **style shift** – adapting their language across social contexts – are core to the field of sociolinguistics (Bell, 1984; Coupland, 2007; Marwick and Boyd, 2011; Eckert, 2012). Work in computational sociolinguistics has studied these questions online, exploring variation in style across communities (e.g., Zhang et al., 2017; Cork et al., 2020; Lucy and Bamman, 2021), as well as variation in how individuals adjust their style in different contexts (e.g., Danescu-Niculescu-Mizil et al., 2013b; Doyle et al., 2016; Pavalanathan, 2018). However, little work has considered variation at both levels simultaneously – that is, how speakers may carry over their use of a particular community's style outside that community. (To our knowledge, only Koschate et al., 2021, has studied this, limited to style-shifting between a single pair of communities.)

Here, we explore style-shifting among members of the Manosphere, a misogynistic hate speech group (HSG) active on the online platform Reddit (Lilly, 2016; Ribeiro et al., 2021). Extending prior work (Pavalanathan, 2018; Koschate et al., 2021), we investigate whether authors retain aspects of a Manospheric style when posting in a range of 14 other non-Manospheric communities (subreddits). Our approach incorporates a broad set of linguistic features to identify nuanced ways that the style of HSGs may bleed into non-hateful spaces. Table 1 illustrates the stylistic differences that our method taps into. We explore three research questions:
RQ1: What features characterize the Manospheric linguistic style, beyond toxicity?
RQ2: Do Manospheric authors shift their style when posting in non-Manospheric communities?
RQ3: What elements of the Manospheric style are carried over into non-Manospheric communities?

2 Approach, Methodology, and Data

In our work, we develop two kinds of linguistic style classifiers.[1] In RQ1, our goal is to assess whether the Manosphere has a distinct linguistic style, and to identify the important features of this

[1] We make all code and data available at https://github.com/jaikaggarwal/emnlp_2024_styleshifting.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 21976–21989
November 12-16, 2024 ©2024 Association for Computational Linguistics

ID	Post	Score	Posted By
P_1	Why are you fucking with trash women who date trash men?	0.98	M on Manosphere
P_2	The claim that western women are oppressed or really any of my friends political "views" aka shit they see in fb and like or repost.	0.60	M on r/AskReddit
P_3	Whenever people say that racism doesn't exist, and black people have the same opportunities and treatment as white people.	0.40	B on r/AskReddit

Table 1: Manosphericness Scores (range 0–1) of posts written by (M)anospheric or (B)aseline authors. P_2 and P_3 are both responses to the same post: *What's something you often let slide because an argument just isn't worth it?*

style. To do this, we train one platform-level binary logistic regression model to predict whether a post was written inside or outside the Manosphere.

In RQ2, we investigate how much Manospheric individuals shift their style to that of some non-Manospheric subreddit S. Thus, for each subreddit S we consider, we train a binary logistic regression model to predict whether a post was written in the Manosphere or in S, yielding a level of "Manosphericness" of each post (compared to the style of S). We can assess style-shifting for each author by comparing the average Manosphericness of their posts in the Manosphere and in S.

2.1 Linguistic Features

We use three kinds of features to assess style: uncivil language ($n = 3$), syntactic features ($n = 29$), and semantic features ($n = 46$); Appendix A has the full list and how we compute them.

Uncivil language is a key aspect of style in a HSG. We assess toxicity (the focus of prior work on HSGs; e.g., Ribeiro et al., 2021; Habib et al., 2022), as well as subtler features of uncivil language: negativity (valence) and (im)politeness.

Syntactic features are generally relevant for style, as they capture aspects of linguistic expression that can signal group membership (Cork et al., 2020) – e.g., historians on Reddit may use fewer exclamation marks than gamers. Indeed, computational work on style has focused on such non-topical features, such as punctuation or parts of speech, to avoid tying variation in style to variation in topic (e.g., Pavalanathan, 2018; Koschate et al., 2021). Importantly, sociolinguistic work argues that topic and style are inseparable (Eckert, 2012; Zhang et al., 2017), since the choice of what to discuss itself reflects speaker identity. We include both syntactic and semantic features, using LIWC categories of words such as *she/he pronouns* [syntactic] and *female* [semantic] (Pennebaker et al., 2015).

Specifically, our goal is to include general semantic features that reflect the Manospheric linguistic style without overfitting to subreddit-specific topical differences. To do so, we train a platform-level classifier in RQ1 including our full set of semantic features from LIWC, and then identify those semantic features that are most important in classification; we assume these are generally useful in distinguishing Manospheric style from those of the various subreddits, regardless of their specific topics. We include only these general semantic features when training our subreddit-specific classifiers in RQ2.

2.2 Reddit Data

Reddit is an online platform where users post in a wide range of communities called subreddits. Ribeiro et al. (2021) identified 51 subreddits as forming the over-arching community of the Manosphere. We investigate how members of the Manosphere (as so defined) style-shift when posting on 14 large, topically-diverse (non-Manospheric) subreddits (given in Appendix B).

We use a 10% sample of the Pushshift Data Dumps (Baumgartner et al., 2020) to collect Reddit data from 2014-2017. (See Appendix C for all data processing details and statistics.) We remove all posts written by Manospheric users prior to their first post on the Manosphere, so that remaining posts reflect their behaviour after participating in the Manosphere. Then, to ensure that we have enough data for our style-shifting analyses – which assess user-level behavior across subreddits – we only retain users with at least 100 posts. We refer to all authors with at least 10 posts in the Manosphere as Manospheric authors, and those who have never posted in the Manosphere as Baseline authors.[2]

Training Data. The training data for each subreddit-specific classifier consists of two sets of posts: posts written by Manospheric authors in the Manosphere, and posts written by Baseline authors in subreddit S. To ensure that we compare authors with similar degrees of engagement in each

[2] In Appendix D, we describe key aspects of how Manospheric authors engage with non-Manospheric spaces.

	All	Unc.	Syn.	M/F	Final
Acc.	0.69	0.56	0.59	0.64	0.68
TPR	0.64	0.31	0.56	0.41	0.60
TNR	0.74	0.81	0.62	0.88	0.75

Table 2: RQ1: Comparisons of classifiers trained using all 78 features, only unc(ivil), only syn(tactic), only male/female (M/F), and our final set of 34 features.

space, we match Baseline authors to Manospheric authors by their posting volume in their respective spaces (e.g., the posting volume of a Baseline author on subreddit S). We sample Manospheric and Baseline authors proportional to their average post score in the Manosphere and in S, respectively, assuming that higher-scoring posts are more representative of a community's style (LaViolette and Hogan, 2019). Each subreddit-specific dataset has 800–2400 authors with 50K–120K posts of each type (see Appendix C.2).

We form the platform-level training dataset using data from the 14 subreddit-specific datasets. However, the subreddit-specific datasets cannot simply be merged, as authors may appear in the training data of multiple subreddits. Instead, we begin with the superset of 2.4K unique Manospheric authors who appear across the 14 subreddit-specific training sets. We then match each Manospheric author to a unique Baseline author, ensuring an equal number of the latter from each of the subreddits. This process yields 158K posts across 2.4K authors in each of the Manospheric and Baseline groups.

Test Data for Style-shifting. For each of the 14 non-Manospheric subreddits S, we first extract all Manospheric and Baseline authors with at least 10 comments on S. We then match Manospheric and Baseline authors by their posting volume in S to ensure users with a similar degree of engagement in S. The test set for S consists of three sets of comments: Baseline authors' comments on S, Manospheric authors' comments on S, and Manospheric authors' comments in the Manosphere.[3]

3 Manospheric Linguistic Style (RQ1)

Using the platform-level dataset, we fit a logistic regression to distinguish posts in the Manosphere (class 1) from those in the 14 non-Manospheric subreddits (class 0). We evaluate our model with 5x2 cross-validation (statistics below are averages).

As seen in Table 2, our model trained on all 78 linguistic features has 69% accuracy, with a true positive rate (TPR) of 0.64 and true negative rate (TNR) of 0.74, showing that the Manosphere has a distinct and detectable linguistic style. (These results are notable given that we use posts as short as 5 tokens.) This style is characterized by discussions of gender (*female*, *male*, use of *she/he* pronouns), toxic language, and the use of 2nd-person pronoun *you*; the latter syntactic feature perhaps captures the confrontational tone of the Manosphere, as in P1 of Table 1 (see Appendix E for further detail). To identify general semantic features relevant to the Manospheric style, we find an elbow in a feature importance graph (Cork et al., 2020); *female* and *male* are the only two highly important semantic features at the platform level. Henceforth, we use only these two of the set of semantic features.

Table 2 also shows that features considered in previous work – only uncivil language ($n = 3$) or syntactic features ($n = 29$) – are much worse at capturing the Manospheric style than the full set of 78 features. Though the model achieves a surprisingly high accuracy with just the two general semantic features, *female* and *male*, we see that it better predicts Baseline posts (high TNR) than Manospheric posts (low TPR). Our interpretation is that because Manospheric speech is dominated by discussions of gender, posts that do not mention gender are much less likely to have been written in the Manosphere (leading to a high TNR). That being said, if the *male/female* categories are mentioned in a post, that does not necessarily mean that the post is Manospheric (leading to a low TPR). This suggests that though features related to gender are important to the Manospheric identity, they do not provide a complete picture of the Manospheric style.

To capture style more comprehensively than in prior work, we combine the uncivil, syntactic, and two general semantic features to create our final feature set ($n = 34$), achieving comparable performance to the full model. Together, these results highlight: (1) the importance of combining topical and non-topical features of language in sociolinguistic analyses of variation, and (2) the importance of considering features beyond toxicity when studying the speech of a HSG.

4 Manospheric Style-Shifting (RQ2)

We now explore whether Manospheric authors shift their style when posting on a non-Manospheric

[3] No authors appear in both training and test data.

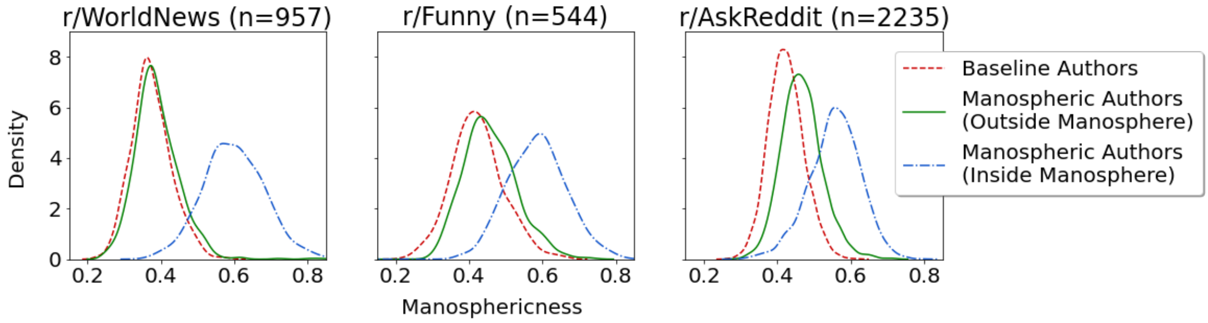

Figure 1: RQ2: Distributions of author-level Manosphericness scores in three subreddit-specific test sets. The three subreddits show low (r/WorldNews), medium (r/Funny), and high (r/AskReddit) degrees of style-shifting.

subreddit S. To do so, we assess the level of **Manosphericness** of each author's set of posts on the Manosphere, and on S. For each S, we train a binary logistic regression model to predict whether a post was written on the Manosphere or on S, using our final set of 34 features on the subreddit-specific data described in Section 2.2. We quantify the Manosphericness of an author's set of posts (on the Manosphere or on S) as the average of the model's class probability estimate over the set of posts (1 is fully Manospheric).

We assess style-shifting by comparing the Manosphericness of Manospheric authors on S to: (1) their Manosphericness in the Manosphere, and (2) the Manosphericness of Baseline authors on S. The first comparison reveals whether Manospheric authors shift their style relative to how they post in the Manosphere. The second shows whether they completely shift to speaking like other authors on S, or if they retain some degree of the Manospheric style. We conduct paired t-tests for each comparison and report effect size using Cohen's d.

We find consistent and statistically significant results on all 14 subreddits, confirming the pattern exemplified in Table 1. Figure 1 visualizes our findings for three subreddits (full results in Appendix F).[4] Comparing the green (middle) and blue (rightmost) distributions reveals that when posting outside the Manosphere, Manospheric authors shift toward the style of non-Manospheric spaces (Cohen's d of 1.27–3.38 across the 14 subreddits).

Differences between the green and red (leftmost) distributions reveal that Manospheric authors do not completely shift to speaking like other members of non-Manospheric spaces: they use a more Manospheric style than Baseline authors across subreddits. The difference in Manosphericness between Manospheric and Baseline authors is smallest (though still significant) on r/WorldNews ($d = 0.26$); the remaining subreddits showing moderate to large effect sizes, such as r/Funny ($d = 0.49$) and r/AskReddit ($d = 0.83$). These results suggest a potential for harm, since Manospheric authors are carrying over aspects of Manospheric language to other communities on Reddit; cf. Table 1.

Feature	**Post**	**Parent Post**
Toxicity	0.031*** (0.002)	0.027*** (0.002)
Impoliteness	0.035*** (0.002)	0.024*** (0.002)
Negativity	0.043*** (0.002)	0.059*** (0.002)
Female	0.034*** (0.002)	0.042*** (0.002)
Male	-0.016*** (0.002)	0.016*** (0.002)

Table 3: RQ3: Regression coefficient estimates for stylistic spillover effects, with standard error values in parentheses. *** shows significance at $p < 0.001$.

5 Stylistic Spillover (RQ3)

In the previous section, we showed evidence of Manospheric authors retaining some degree of the Manospheric style outside the Manosphere. Here, we assess *which* elements of the Manospheric style spill over into non-Manospheric subreddits.[5] That

[4] Note that the blue (rightmost) distributions of Manosphericness inside the Manosphere vary a bit across subreddits: the Manosphericness score is relative to the style of S, since it is based on a subreddit-specific classifier.

[5] In using the word "spillover", we are not claiming that Manospheric authors necessarily learned to speak in an Manospheric way in the Manosphere, and then started talking that way in other subreddits. Rather, we simply mean that there is a distinct style in the Manosphere compared to the other subreddits, and elements of this style are used by

ID	Post	Posted By
$Parent_1$	What's something other guys do that bugs the crap out of you?	B on r/AskMen
$Reply_{1(B)}$	Too much cologne. Dude, I don't need to smell your old spice fifteen feet away.	B on r/AskMen
$Reply_{1(M)}$	Putting women on a pedestal and treating them like these magical, amazing, otherworldly beings.	M on r/AskMen
$Parent_2$	Spoiled Brat screaming at Grandpa over IPhone Appointment.	B on r/Videos
$Reply_{2(B)}$	[...] This chick needs a reality check.	B on r/Videos
$Reply_{2(M)}$	That bitch needs to be hit in the head with a bag of nickels. [...]	M on r/Videos

Table 4: RQ3: Comparison of (M)anospheric and (B)aseline author responses to parent posts.

is, outside the Manosphere, which features do Manospheric authors use more than Baseline authors? We focus on interpretable features thought to be especially relevant to the Manospheric identity: the use of uncivil language (toxicity, impoliteness, and negativity) and discussions of gender (*female* and *male*).

For each feature, we fit a logistic regression to predict whether or not a post was written by a Manospheric author given the feature's value. Work in sociolinguistics suggests that feature usage may be shaped by the post that a user is responding to (e.g., Giles et al., 1991; Danescu-Niculescu-Mizil et al., 2011). To assess whether Manospheric authors use features *beyond* what might be used in the post they are responding to, we also include the feature value of the post's parent as a control predictor.[6] Our dataset consists of all 823K posts written across the 14 non-Manospheric subreddits.

Table 3 shows the regression coefficients for the features that Manospheric authors both *use* and *respond to* more than Baseline authors in non-Manospheric subreddits. We find that posts written by Manospheric authors are more toxic, impolite, and negative than those written by Baseline authors, and include greater use of *female* words. These findings cohere with the nature of the Manosphere as a misogynist HSG. Interestingly, though we find that the *male* feature is characteristic of the Manospheric style in RQ1, Manospheric authors use fewer *male* words than Baseline authors outside the Manosphere. At the same time, we see they *respond* to posts with greater use of *male* words.

Table 4 provides examples of parent/reply pairs that highlight these patterns. In responding to $Parent_1$, the Baseline author of $Reply_{1(B)}$ continues the conversational focus on men, while the Manospheric author of $Reply_{1(M)}$ shifts the focus towards women by criticizing men who view women too positively. Comparing the two replies to $Parent_2$ further reveals how Manospheric language spills over into other subreddits; relative to the Baseline author of $Reply_{2(B)}$, the Manospheric author of $Reply_{2(M)}$ conveys the idea of a woman "needing a reality check" in a more toxic manner. In sum, these results confirm that potentially harmful elements of the Manospheric style bleed into their posts on non-Manospheric subreddits.

6 Conclusion

We find that members of the Manosphere, a prominent online hate speech group, have a distinct linguistic style. Moreover, when posting outside the Manosphere, Manospheric authors retain elements of this style, including greater use of female-gendered terms and use of more uncivil language. These findings suggest concrete ways a hate speech group may shape discussions in other spaces. Future work can build on our sociolinguistically-inspired analyses to further explore the impact of hate speech groups. For example, causal analyses could reveal whether the act of participating in the Manosphere changes the style that authors use in non-Manospheric spaces, as well as how this stylistic spillover may harm community health.

7 Acknowledgments

We acknowledge the support of NSERC of Canada (through grant RGPIN-2017-06506 to SS), as well as the support of the Data Sciences Institute, University of Toronto (through a Catalyst Grant to SS).

Manospheric authors in those other subreddits. We leave temporal or causal analyses to future work.

[6] We did not use the parent post as a predictor in RQ2 because there the aim was to simply capture whether authors were using the Manospheric style, not to account for reasons behind that (such as properties of the parent post). In RQ3, however, we want to see which features an author introduces from the Manospheric style above and beyond what is in the post that they are replying to.

We also thank the Perspective team for graciously increasing our query limit for their API.

8 Limitations

In this section, we note several limitations of our approach, as well as how we mitigate these concerns as best as possible.

LIWC One major limitation of LIWC is that it does not account for the context in which words are used. For example, if a word in the "Certainty" category is preceded by a negation, it may instead connote uncertainty; the LIWC would simply count this as the use of a "Certainty" word. This concern is mitigated for our syntactic features (which are more robust to this issue) and for our uncivil language features (which we infer using neural-based methods that better account for context).

A second limitation is that LIWC was constructed in a top-down fashion. As such, both the categories and their respective word lists are subject to the biases of the researchers. The top-down nature also means that the word lists may be incomplete. This is especially true given that we use LIWC-2015, as the more recent LIWC-2022 was not available when we began our research. Thus, the word lists do not include novel words that emerged in the last decade.

Though using LIWC features offers some degree of interpretability for aspects of style, future work may jointly consider these features along with latent aspects of style derived from methods beyond count-based approaches (see Zhu and Jurgens, 2021 for one such example).

Model Accuracy We make inferences about style-shifting using regression models that achieve accuracies between $65 - 75\%$. Though these accuracies are notable for the reasons described in the main text, they show that we do not perfectly capture the Manospheric style. As mentioned previously, future work may investigate whether capturing additional, potentially latent, aspects of style result in improved accuracy on this task.

Generalizability It is unclear whether our results generalize to populations beyond the Manosphere. Our claims about style-shifting involving general topical features may not be true of other HSGs, as it would require there being suitable semantic features that distinguish their discourse from the rest of Reddit. Moreover HSGs are particularly extreme groups; style-shifting between less extreme groups may not show the same patterns (c.f. Koschate et al., 2021). Even within the Manosphere, our results hold for a set of active users on Reddit (those with at least 100 posts on Reddit). This constraint was important for gathering sufficient user-level data to perform our analyses, but it is unclear whether our style-shifting results hold for less active users.

Data Access In May of 2023, the Pushshift Data Dumps were made unavailable at their original link, limiting the future accessibility of our data. Future work will need to use Reddit's official API to re-extract our data (we will release all comment ids used in our paper upon publication).

9 Ethical Considerations

Data privacy is a major ethical consideration when using online data, as we do here. Though all posts in the Pushshift Data Dumps were publicly accessible at the time of collection by Baumgartner et al. (2020), it is critical that we support a user's right to be forgotten. This is especially important when using online hate speech data; individuals who posted such content in the past may later choose to have their data redacted. Prior to being made unavailable, the maintainers of the Pushshift API offered one solution to this issue by allowing Redditors to have their usernames and posts redacted upon request. On our end, we exclude data from any users that deleted their account (despite their posts remaining in the data dumps).

An open question is how best to support users whose data remains in our dataset, but who may want to redact their data in the future. As suggested by Proferes et al. (2021), we only release the comment ids, anonymized user ids, and feature vectors for the posts we use. Future researchers may re-extract the post text and user ids using the official Reddit API. Though this may lead to incomplete data, we err on the side of data privacy, and offer maximal reproducibility given this constraint.

A second ethical concern relates to automatically inferring emotional properties (including valence, politeness, and toxicity) from online text. Performing automatic emotion recognition risks misrepresenting the views of users when the inferred values do not match the user's intended emotions. At the same time, work on the language of the Manosphere requires the study of such features, given their potential to negatively influence the health of communities outside the Manosphere. To address this concern as best we can, we anonymize user-level information for any of the posts in our dataset, thereby de-linking users from the emotions we infer from their language.

References

Jai Aggarwal, Brian Diep, Julia Watson, and Suzanne Stevenson. 2023. Investigating online community engagement through stancetaking. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5814–5830.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit Dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 830–839.

Allan Bell. 1984. Language style as audience design. *Language in society*, 13(2):145–204.

Julian Brooke. 2014. *Computational Approaches to Style and the Lexicon*. Ph.D. thesis, Citeseer.

Jonathan P Chang, Caleb Chiam, Liye Fu, Andrew Z Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. Convokit: A toolkit for the analysis of conversations. *arXiv preprint arXiv:2005.04246*.

Alicia Cork, Richard Everson, Mark Levine, and Miriam Koschate. 2020. Using computational techniques to study social influence online. *Group Processes & Intergroup Relations*, 23(6):808–826.

Nikolas Coupland. 2007. *Style: Language variation and identity*. Cambridge University Press.

Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. 2011. Mark my words! Linguistic style accommodation in social media. In *Proceedings of the 20th international conference on World wide web*, pages 745–754.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013a. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259.

Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013b. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web*, pages 307–318.

Gabriel Doyle, Dan Yurovsky, and Michael C Frank. 2016. A robust framework for estimating linguistic alignment in Twitter conversations. In *Proceedings of the 25th international conference on world wide web*, pages 637–648.

Penelope Eckert. 2012. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual review of Anthropology*, 41(1):87–100.

Howard Giles, Nikolas Coupland, and Justine Coupland. 1991. Accomodation theory: Communication, context, and consequences. In *Contexts of accommodation: developments in applied sociolinguistics/ed. by Howard Giles*, pages 1–68. Cambridge Univ. Press.

Hussam Habib, Padmini Srinivasan, and Rishab Nithyanand. 2022. Making a radical misogynist: How online social engagement with the manosphere influences traits of radicalization. *Proceedings of the ACM on human-computer interaction*, 6(CSCW2):1–28.

Miriam Koschate, Elahe Naserian, Luke Dickens, Avelie Stuart, Alessandra Russo, and Mark Levine. 2021. ASIA: Automated social identity assessment using linguistic style. *Behavior Research Methods*, 53:1762–1781.

Jack LaViolette and Bernie Hogan. 2019. Using platform signals for distinguishing discourses: The case of men's rights and men's liberation on Reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 323–334.

Mary Lilly. 2016. *'The World is Not a Safe Place for Men': The Representational Politics Of The Manosphere*. Ph.D. thesis, Université d'Ottawa/University of Ottawa.

Li Lucy and David Bamman. 2021. Characterizing English variation across social media communities with BERT. *Transactions of the Association for Computational Linguistics*, 9:538–556.

Alice E Marwick and Danah Boyd. 2011. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New media & society*, 13(1):114–133.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184.

Umashanthi Pavalanathan. 2018. *Computational approaches to understanding stylistic variation in online writing*. Ph.D. thesis, Georgia Institute of Technology.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of LIWC2015. Technical report.

Nicholas Proferes, Naiyan Jones, Sarah Gilbert, Casey Fiesler, and Michael Zimmer. 2021. Studying Reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media+ Society*, 7(2):20563051211019004.

Nils Reimers, Iryna Gurevych, Nils Reimers, Iryna Gurevych, Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Manoel Horta Ribeiro, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini, Summer Long, Stephanie Greenberg, and Savvas Zannettou. 2021. The evolution of the Manosphere across the web. In *Proceedings of the international AAAI conference on web and social media*, volume 15, pages 196–207.

Amaury Trujillo and Stefano Cresci. 2022. Make Reddit great again: Assessing community effects of moderation interventions on r/the_donald. *Proceedings of the ACM on Human-computer Interaction*, 6(CSCW2):1–28.

Justine Zhang, William Hamilton, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. Community identity and user engagement in a multi-community landscape. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 377–386.

Jian Zhu and David Jurgens. 2021. Idiosyncratic but not arbitrary: Learning idiolects in online registers reveals distinctive yet consistent individual styles. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 279–297, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Linguistic Features

Here, we explain how we extract the features used in our analyses. The full list of libraries and versions we use can be found in the codebase attached to this submission. All artifacts are used in a manner consistent with intended use (as are all artifacts that we create); see the licenses mentioned through this section for further context.

A.1 Uncivil Language Features

We infer values for three uncivil language features: toxicity, negativity, impoliteness.

In line with previous work, we compute toxicity using the "Severe Toxicity" field from Google's Perspective API (Ribeiro et al., 2021; Trujillo and Cresci, 2022).[7] The API defines the "Severe Toxicity" metric as speech that is "very hateful, aggressive, disrespectful [...] or otherwise very likely to make a user leave a discussion or give up on sharing their perspective". We use the Python `google-api-python-client` (version 2.39.0) to call the Perspective API. In cases where the API was unable to return a toxicity value, we assigned the post a toxicity of 0; this occurs for only 0.06% of all posts.

We compute both negativity and impoliteness using the methodology of Aggarwal et al. (2023), who also inferred these values for sentences on Reddit. In their methodology, negativity is computed using the psycholinguistic construct of valence (positivity). They begin with the NRC-VAD lexicon (Mohammad, 2018), which provides human-annotated valence values for 20K English words.[8] They then train a Beta regression model to predict the valence score of each word using its SBERT embedding (Reimers et al., 2019) (according to the `bert-large-nli-mean-tokens` SBERT model, released under an Apache 2.0 License). Beta regression is used as the values are confined to the $[0, 1]$ interval. The regression model is fit using an $80/20$ train/test split stratified over quintiles of the valence scores. The model is evaluated by computing the Pearson correlation of the model's predictions with the ground truth valence score annotations. We repeat this procedure 10 times and use the best performing model, which achieved a Pearson correlation of 0.85, as in Aggarwal et al. (2023).

To infer valence scores for each post in our dataset, we first split each post into its constituent sentences. Then, we use our regression model to infer the valence score of each sentence given its SBERT representation. The valence of a post is computed as the average valence of its sentences.

Aggarwal et al. (2023) also built an SBERT-based logistic regression model to predict the politeness of documents. As an overview, they trained their model on the Wikipedia text subcorpus released as part of the Stanford Politeness Corpus (Danescu-Niculescu-Mizil et al., 2013a).[9] The model was evaluated using 3x10 cross-validation on the politeness requests in the top and bottom quartile of annotated politeness scores, and achieved a mean accuracy of 84.1%. They also tested the cross-domain generalizability of their model using the StackExchange subcorpus in the Stanford Politeness Corpus, and achieved an accuracy of 65.2%; both accuracies are comparable

[7] `https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages`

[8] The lexicon is freely available for research purposes at `https://saifmohammad.com/WebPages/nrc-vad.html`

[9] The corpus was released as part of the Convokit Python library (Chang et al., 2020) under a CC BY License v4.0.

to the models in Danescu-Niculescu-Mizil et al. (2013a). Aggarwal et al. (2023) use the log-odds of the classifier's predicted probability score as their politeness score, where higher values indicate more polite posts. We replicate their procedure; our politeness model achieves the same cross-validation accuracy of 84.1% and cross-domain accuracy of 65.2%.

The SBERT-based features were extracted using a NVIDIA Titan Xp GPU, and used 9 GPU hours total. Access to the Perspective API was rate-limited to 180 queries per minute, requiring 277 total hours for our entire dataset.

A.2 Syntactic and Semantic Features

We extract these features using the text analysis software LIWC-15 (Pennebaker et al., 2015).[10] The LIWC categories are structured hierarchically; for example, the categories for *Anger* and *Sadness* are part of the category of *Negative Emotions*, which itself is in the category of *Affect Words*. The values of the lower-level categories inform the counts of the higher-level categories, leading to a large number of correlated features. Moreover, LIWC includes 4 summary variables that are computed using the counts of the other features. To avoid multicollinearity as a result of these highly-correlated features we only keep the lowest-level categories and remove the summary features. LIWC feature extraction was completed in 1.5 hours.

Table A.1 shows the LIWC features that are included in our set of syntactic features (including function words), and Table A.2 shows the LIWC features included as our set of semantic features. Note that part-of-speech categories that reflect content words, including verbs, adverbs, and adjectives, are considered as semantic features.

We additionally include type-token ratio (TTR) as a syntactic feature as it has been used to assess style previously (Brooke, 2014). We calculate TTR as the number of unique tokens in a post divided by the total number of tokens.

B Selecting Non-Manospheric Subreddits

We study Manospheric linguistic behaviour on non-Manospheric subreddits that had more than 300 Manospheric authors who posted at least 10 comments on the subreddit. We excluded 2 sub-

LIWC Code	**Category Description**
WC	Word Count
WPS	Words per Sentence
Sixltr	Six-letter Words
i	1st Person Singular Pronouns
we	1st Person Plural Pronouns
you	2nd Person Pronouns
shehe	3rd Person Singular Pronouns
they	3rd Person Plural Pronouns
ipron	Impersonal Pronouns
article	Articles
prep	Prepositions
auxverb	Auxiliary Verb
conj	Conjunctions
negate	Negations
interrog	Interrogative Words
number	Numbers
quant	Quantifiers
Period	Periods
Comma	Commas
Colon	Colons
SemiC	Semicolons
QMark	Question Marks
Exclam	Exclamation Marks
Dash	Dashes/Hyphens
Quote	Quotation Marks
Apostro	Apostrophes
Parenth	Parentheses
OtherP	Other Punctuation

Table A.1: LIWC categories used to compute function words and syntactic features.

[10] The license for LIWC can be found at https://www.liwc.app/help/eula.

LIWC Code	Category Description
verb	Verbs
adverb	Adverbs
adj	Adjectives
posemo	Positive Emotion
anx	Anxiety
anger	Anger
sad	Sadness
family	Family
friend	Friend
female	Female Referents
male	Male Referents
insight	Insight
cause	Cause
discrep	Discrepancies
tentat	Tentativeness
certain	Certainty
differ	Differentiation
see	Seeing
hear	Hearing
feel	Feeling
body	Body
health	Health/Illness
sexual	Sexuality
ingest	Ingesting
affiliation	Affiliation
achieve	Achieve
power	Power
reward	Reward
risk	Risk
focuspast	Past Focus
focuspresent	Present Focus
focusfuture	Future Focus
motion	Motion
space	Space
time	Time
work	Work
leisure	Leisure
home	Home
money	Money
relig	Religion
death	Death
swear	Swear words
netspeak	Netspeak
assent	Assent
nonfl	Nonfluencies
filler	Fillers

Table A.2: LIWC categories used to compute semantic features.

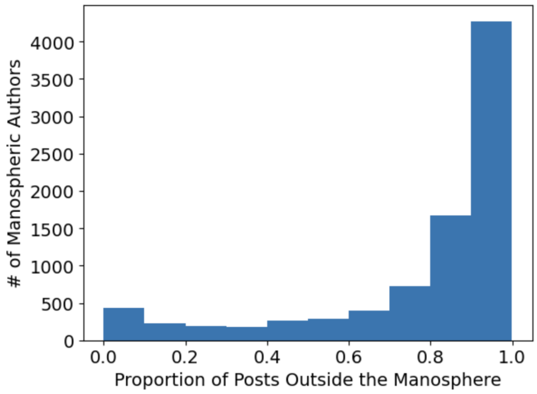

Figure B.1: Proportion of posts by Manospheric authors (n=8650) that are written outside the Manosphere.

reddits (r/kotakuinaction and r/the_donald) as we wanted to assess style-shifting on mainstream subreddits. This led to a set of 14 topically diverse subreddits: r/AskReddit, r/News, r/WorldNews, r/TodayILearned, r/AskMen, r/Movies, r/Politics, r/Technology, r/AdviceAnimals, r/Videos, r/Pics, r/Funny, r/WTF, and r/Gaming.

C Data Extraction

C.1 Preprocessing and Filtering

To create our dataset, we use English-language posts written between 2014-2017. We use 2017 as our endpoint to control for potential changes in author behaviour due to r/incels being banned at the end of 2017. We use 2014 as our starting point to ensure we have sufficient data for our analyses.

To preprocess our data, we remove all deleted posts and those written by deleted users, the "Automoderator" or "Autotldr" accounts, or usernames ending in "bot" (regardless of case). We also substitute out all mentions of hyperlinks, usernames, and subreddit names for LINK, USER, and REDDIT tokens, and enforce a minimum length of 5 tokens (not counting punctuation).

As our analysis in Section 4 requires each post to have a parent post, we additionally extract the post preceding each comment in our dataset, creating pairs of *parents* and *replies*. If the reply was a top-level comment, its parent was a submission; otherwise, parents were other comments. Our final dataset consists of all remaining (parent, reply) pairs where both posts meet our filtering criteria (descriptive statistics in Table C.1).

	# Users	# Posts
Baseline	636K	198M
Manospheric	8650	4.4M

Table C.1: Descriptive statistics for our Baseline and Manospheric authors across all of Reddit.

C.2 Sampling Manospheric Training Data

Our filtered dataset contains data for 36 of the 51 subreddits released by Ribeiro et al. (2021).[11] Ribeiro et al. (2021) divide these Manospheric subreddits into 5 mutually exclusive *Manospheric subcultures* (e.g. Incels, or Pick Up Artistry). As these subcultures may have their own distinct styles, we additionally ensure that the subcultural makeup of each subreddit's training data matches the subcultural makeup of its testing data.

First, we assign each Manospheric author to a subculture based on the subculture in which they posted more than 50% of their posts. Then, to create the training data for non-Manospheric subreddit S, we sample Manospheric authors from each subculture proportional to the number of authors per subculture in the testing data for S. We set a minimum of 50 users for the subculture with the fewest number of individuals in the testing data, and sampled individuals from the remaining subcultures proportionally. For each of these users, we only consider their posts in their assigned subculture.

C.3 Training and Testing Dataset Statistics

Table C.2 shows the final number of posts and authors for each of our non-Manospheric subreddits.

D Manospheric Engagement Habits Outside the Manosphere

In this section, we provide additional context about some of the engagement dynamics of Manospheric individuals in these non-Manospheric spaces. To capture the degree to which Manospheric authors are active in non-Manospheric spaces, we compute the proportion of an individual's total posts on Reddit that are posted outside the Manosphere. Figure B.1 shows that Manospheric individuals post broadly across the platform, with an average of 78% (± 26%) of posts being written outside the Manosphere. These results reveal that the Manosphere is not siloed off from the rest of the platform, emphasizing the importance of studying how their style is carried into non-Manospheric spaces.

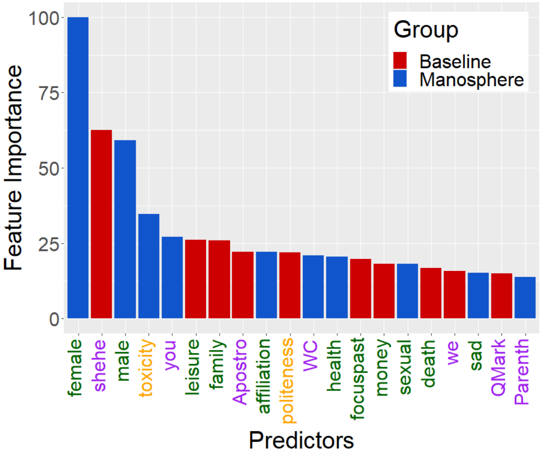

Figure E.1: Top 20 most important syntactic (purple), semantic (green), and uncivil (orange) linguistic features of Manospheric style.

We also study how well-received posts by Manospheric authors are outside of the Manosphere, relative to posts written by Baseline authors. For each of the two groups of authors, we compute the average score of each author's set of posts in a particular subreddit, and then compute the subreddit-level average score as the average across authors. Table D.1 shows that Manospheric authors tend to write posts that receive a lower average score than Baseline authors. Paired t-tests show that this difference is significant across 8/14 subreddits. Our findings suggest that something about the manner in which Manospheric individuals engage with non-Manospheric spaces results in their posts being viewed less favourably than those of Baseline authors. We leave the question of whether this is driven by style to future work.

E RQ1 - Manospheric Style

Figure E.1 shows the features that best predict the Manospheric speech style. Though we evaluate our model with 5x2 cross-validation, the feature importance graph was generated using a model trained on the entire platform-level training set. We see that features pertaining to gender come out to be the most important, including references to *female* and *male* individuals. Though Manospheric individuals use the third-person pronouns *she* and *he* more than Baseline individuals on average, the *shehe* variable is more predictive of Baseline authors after controlling for the other gender features. We also see

[11] Their data was released under a CC BY License v4.0.

	Training Data		Testing Data		
	M	B	M (in Manosphere)	M (on S)	B (on S)
News	95989 posts 1752 authors	93569 posts 1752 authors	44584 posts 932 authors	39642 posts 932 authors	39358 posts 932 authors
AskReddit	50746 posts 827 authors	49864 posts 827 authors	95639 posts 2235 authors	127720 posts 2235 authors	127203 posts 2235 authors
WorldNews	60322 posts 1139 authors	60319 posts 1139 authors	45068 posts 957 authors	33734 posts 957 authors	33701 posts 957 authors
TodayI Learned	59754 posts 1177 authors	59854 posts 1177 authors	37509 posts 794 authors	24889 posts 794 authors	24902 posts 794 authors
AskMen	56820 posts 1094 authors	54434 posts 1094 authors	29117 posts 541 authors	27538 posts 541 authors	27903 posts 541 authors
Movies	39250 posts 813 authors	38280 posts 813 authors	14375 posts 349 authors	11953 posts 349 authors	11949 posts 349 authors
Technology	77942 posts 1551 authors	69588 posts 1551 authors	14474 posts 319 authors	8556 posts 319 authors	8563 posts 319 authors
Politics	67981 posts 1249 authors	67989 posts 1249 authors	35697 posts 869 authors	45788 posts 869 authors	45787 posts 869 authors
Advice Animals	128491 posts 2410 authors	114061 posts 2410 authors	28588 posts 647 authors	20302 posts 647 authors	20302 posts 647 authors
Videos	50516 posts 919 authors	47852 posts 919 authors	27353 posts 643 authors	21663 posts 643 authors	21324 posts 643 authors
Pics	61759 posts 1145 authors	60739 posts 1145 authors	27703 posts 608 authors	14820 posts 608 authors	14824 posts 608 authors
Funny	71708 posts 1391 authors	67511 posts 1391 authors	24516 posts 544 authors	13528 posts 544 authors	13529 posts 544 authors
WTF	61632 posts 1216 authors	55490 posts 1216 authors	14844 posts 353 authors	8638 posts 353 authors	8640 posts 353 authors
Gaming	57073 posts 1112 authors	56796 posts 1112 authors	14625 posts 318 authors	6997 posts 318 authors	6997 posts 318 authors

Table C.2: Post and author counts for the training and testing sets for each of the 14 non-Manospheric subreddits. We show the number of posts written by both M(anospheric) and B(aseline) authors.

Subreddit	B	M
News	46.39	39.10
AskReddit	**83.11**	**56.79**
WorldNews	40.67	31.82
TodayILearned	**58.80**	**39.15**
AskMen	**16.88**	**13.46**
Movies	44.85	30.38
Technology	46.62	34.62
Politics	**28.53**	**13.60**
AdviceAnimals	**39.67**	**24.91**
Videos	**64.19**	**39.56**
Pics	**60.71**	**33.26**
Funny	43.34	34.03
WTF	**51.71**	**31.44**
Gaming	40.39	30.58

Table D.1: Average score of (B)aseline and (M)anospheric authors outside the Manosphere. Bolded rows indicate a significant difference at $p < 0.05$, after applying Bonferroni correction for 14 tests.

that the toxicity variable comes out to be important, as expected. Lastly, Manospheric individuals use the second-person pronoun *you* more than Baseline authors do; inspection of Manospheric posts reveals that this stems from engagement with previous posters/commenters.

F RQ2 - Style-Shifting

Subreddit	Classifier Accuracy
News	70.1%
AskReddit	67.3%
WorldNews	74.1%
TodayILearned	70.0%
AskMen	65.0%
Movies	74.2%
Technology	73.3%
Politics	72.2%
AdviceAnimals	65.5%
Videos	68.2%
Pics	70.4%
Funny	69.8%
WTF	70.3%
Gaming	74.2%

Table F.1: Classifier accuracy of each subreddit-specific classifier.

Table F.1 shows the accuracies for our 14 subreddit-specific classifiers. Figure F.1 visualizes the style-shifting results for all 14 subreddits, and

Subreddit	vs. Self in Manosphere	vs. Baseline Authors
News	2.42	0.32
AskReddit	1.80	0.83
WorldNews	3.09	0.26
TodayILearned	2.47	0.29
AskMen	1.27	0.81
Movies	3.15	0.64
Technology	3.31	0.29
Politics	2.98	0.28
AdviceAnimals	1.65	0.47
Videos	2.06	0.50
Pics	2.20	0.34
Funny	2.31	0.49
WTF	2.73	0.33
Gaming	3.38	0.51

Table F.2: Effect sizes (Cohen's d) of the style-shifting comparisons for each non-Manospheric subreddit S. The first column compares the Manosphericness of Manospheric authors on S to their Manosphericness in the Manosphere. The second column compares the Manosphericness of Manospheric authors on S to that of Baseline authors on S.

Table F.2 shows the effect sizes for each of our two comparisons across the 14 subreddits. Note that all comparisons are statistically significant at $p < 0.001$, after applying Bonferroni correction for the 28 total tests.

In Figure F.1, we see slight variation in the Manosphericness scores of Baseline authors across subreddits. We leave the question of how subreddit-level differences in tone and topic (e.g., less polite language or greater mentions of the *female* category) shape their relative Manosphericness scores to future work.

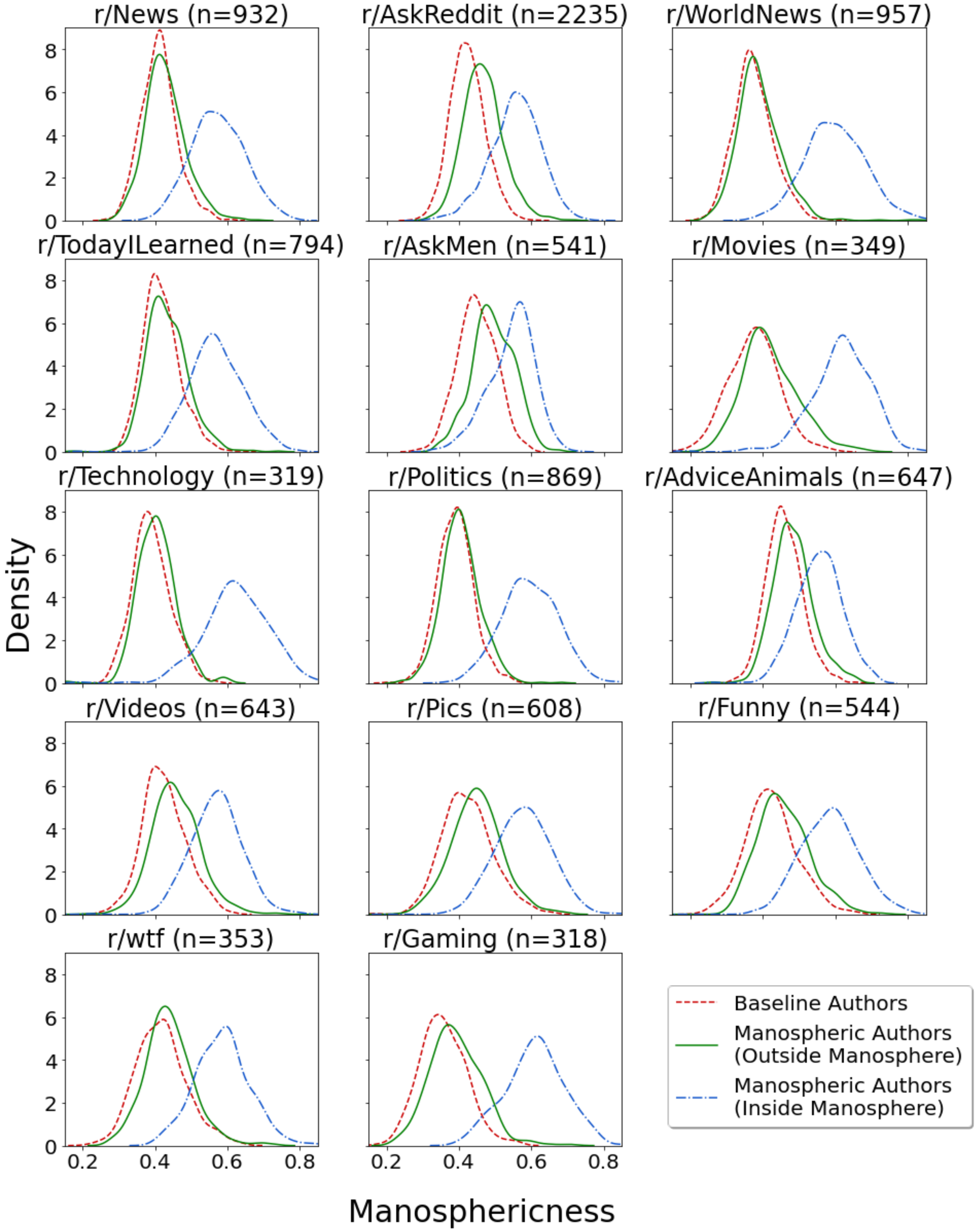

Figure F.1: Style-shifting of Manospheric authors in all 14 non-Manospheric subreddits.