# Among Us: Language of Conspiracy Theorists on Mainstream Reddit

**Francesco Corso[1,2,3], Giuseppe Russo[4], Francesco Pierri[1,*], Gianmarco De Francisci Morales[2,3*]**

[1]Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milan, Italy
[2]CENTAI, Turin, Italy
[3]Intesa Sanpaolo Innovation Center, Turin, Italy
[4]EPFL, Lausanne, Switzerland
[*]Equal contribution **Correspondence:** francesco.pierri@polimi.it

## Abstract

The interaction between fringe subcultures and mainstream online communities poses significant challenges for understanding discourse on social media. In this work, we investigate whether users active in conspiracy-focused communities exhibit detectable linguistic signatures when participating in general-interest spaces, such as news, humor, or hobbyist forums. We analyze a large-scale longitudinal dataset of over 500 million comments spanning 10 years of Reddit activity, examining the communication patterns of these users across diverse social contexts independent of the topics they discuss. We show that these users exhibit distinctive linguistic patterns that enable machine learning models to reliably distinguish them from the general population within individual communities (averaging 87% accuracy across more than 20 binary classification tasks). Crucially, no single aggregate model captures these patterns across communities, as community-specific models outperform global classifiers by up to 17 percentage points. This result suggests that while these users are distinct, their linguistic expression is dynamic and highly responsive to the social norms of the environment they inhabit. Our findings suggest the need for tailored interventions in online spaces, as linguistic signals associated with conspiracy and fringe subcultures vary across communities and cannot be effectively addressed by uniform detection or moderation strategies.

## 1 Introduction

Social media platforms facilitate the rapid spread of information and the formation of communities around shared ideologies (Loru et al., 2025). Within this ecosystem, mainstream narratives compete with alternative interpretations of reality (Benkler et al., 2018), allowing subcultures to develop distinct epistemic norms. Among these, conspiracy theories have shown significant resilience and reach in online environments (Starbird, 2017; Monti et al., 2023). Conspiracy theories are not merely fringe beliefs; they are influential, alternative narratives that explain events through the actions of secretive, malevolent groups (Douglas et al., 2019). While often dismissed as speculation, their real-world consequences can be substantial: they have been linked to vaccine hesitancy and public health risks (Enders et al., 2022), prompting official responses from public institutions.[1] In more extreme cases, they can pose a threat to democratic institutions themselves—as exemplified by the January 6th, 2021 attack on the U.S. Capitol, which was partly fueled by conspiracy-driven rhetoric.[2] These narratives can also serve as gateways to more radical ideologies, gradually reshaping individuals' perception of reality (Basit, 2021).

Despite extensive research on conspiracy theories on online platforms (Tangherlini et al., 2020; Faddoul et al., 2020; Samory and Mitra, 2018b; Korenčić et al., 2024a; Corso et al., 2025a; Attanasio et al., 2026; Corso et al., 2025b), how conspiracy theory believers express themselves in mainstream online spaces remains underexplored. Research suggests conspiracy theorists do not just hold different beliefs, but also communicate differently by employing specific rhetorical styles and vocabulary (Samory and Mitra, 2018b). However, it remains unclear whether these linguistic patterns are confined to online conspiracy spaces or are intrinsic features of users' communication style, visible in mainstream discourse.

In this study, we investigate whether users active in conspiracy communities exhibit specific patterns of self-expression that distinguish them from ordinary users. We use Reddit as our primary case study due to its large-scale, publicly accessible dataset and a unique structure based on topical communities, which enables the analysis

[1]EU Commission, 2021
[2]Washington Post, 2021



*Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12996–13017

of users' behavior across both fringe and conventional communities. We articulate our contributions within a comprehensive research question: **Do users who engage with conspiracy theories display distinctive linguistic patterns in mainstream online communities, compared to users without such engagement?**

We employ a dataset of over 500 million comments and hundreds of thousands of users spanning over 10 years of activity on over 20 mainstream communities. We then extract the psycholinguistic features of each comment, which are then aggregated into a user linguistic vector. We then use these vectors to train a supervised classifier, which shows remarkable accuracy and stability in distinguishing users active in conspiracy-related communities from those who are not.

By characterizing how users active in `r/conspiracy` adapt their language across communities, this study contributes to a more nuanced understanding of online conspiracy discourse, with implications for designing context-aware and proportionate interventions on digital platforms.

## 2 Related Work

### 2.1 Detection of Online Conspiracy Theories

Computational approaches to detect online conspiracy theories have primarily focused on identifying explicit narrative structures and content signatures. For instance, Tangherlini et al. (2020) used structural modeling to map the narrative frameworks of theories such as "Pizzagate", distinguishing them from actual conspiracies based on the stability and interconnectivity of their subject-threat relationships. Similarly, Samory and Mitra (2018a) employed topic modeling and n-gram analysis to capture the distinct "conspiratorial language" of dedicated forums, flagging content based on specific vocabulary and topic distributions rather than user style. Early work by Faddoul et al. (2020) investigated the detection of conspiracy-related content in multimodal settings by leveraging textual information associated with videos, such as captions, snippets, and top comments. More recently, Corso et al. (2025a) demonstrated the effectiveness of Large Language Models for identifying conspiratorial content on TikTok, highlighting the potential of recent advances in generative AI to support moderation efforts in multimodal platforms. Relatedly, Diab et al. (2024) explored the capabilities of LLMs for detecting conspiracy-related content on Reddit.

### 2.2 User Pathways To Fringe Communities

Beyond the analysis of content, recent work has modeled online participation in conspiracy discourse as a dynamic, gradual trajectory of engagement. Klein et al. (2019) identified linguistic precursors of conspiracy activity in fringe communities, showing that users who eventually join conspiracy forums display distinct participation and linguistic markers in their posting history before their first direct participation. Phadke et al. (2022) characterize the evolution of conspiracy engagement as a multi-stage process, wherein users progressively adopt the norms and vocabulary of extremist communities through sustained interaction. This effect is further explored by research on gateway communities; for instance, Rollo et al. (2022) and Habib et al. (2022) demonstrate how adjacent ideological spaces, such as the "manosphere", facilitate a user's drift toward more radical, anti-establishment narratives.

### 2.3 The Language of Conspiracy Theories

Psychological research consistently characterizes conspiracy theories as narratives that attribute hidden, intentional agency to powerful actors and that fulfill existential and social motives, particularly under conditions of uncertainty and threat (Douglas et al., 2019; Douglas and Sutton, 2023). Across domains, conspiracy theories are communicated through recurring linguistic patterns, including causal chaining, certainty and authority markers, attribution of malicious intent, and framing that contrasts an informed ingroup with a deceptive outgroup (van Prooijen and Douglas, 2017; Meuer et al., 2023). These communicative features are not incidental: field studies show that conspiratorial discourse is systematically more emotional and distrust-laden than non-conspiratorial explanations of the same events (Fong et al., 2021).

## 3 Methods

### 3.1 Data Collection

Our primary data source is the Pushshift Reddit dataset (Baumgartner et al., 2020), a public archive of posts and comments on the platform, which excludes content created by deleted users or removed by moderators or the authors themselves. Here, we focus on user comments as the fundamental unit

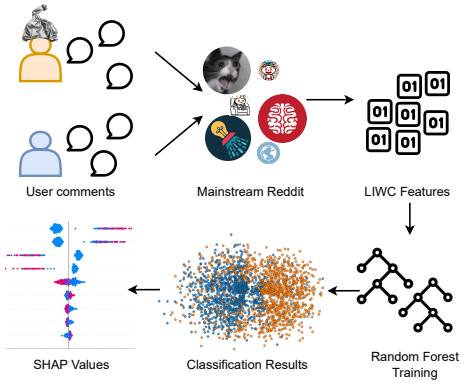



Figure 1: Diagram showing the workflow, from data collection to classification and feature analysis.

of analysis, since they represent the most common form of interaction on Reddit and typically indicate active participation within a community (Naab and Küchler, 2023).

We extracted the full archive of available comments from the largest conspiracy-focused subreddit, `r/conspiracy`, spanning from its creation on 25 January 2008 to 31 December 2023. The resulting corpus includes approximately 25.7 million comments from over 980 000 unique users, with an average of 5k daily comments.

We additionally collected the complete activity of the top subreddits ranked by the number of subscribed users as of June 2025.[3] We selected 22 of the top communities, excluding the following three subreddits: `r/AskReddit` and `r/worldnews`, due to their broad and heterogeneous content, and `r/gaming`. This selection strategy allowed us to observe the behavior of conspiracy users outside of `r/conspiracy` while reducing the bias introduced by the uneven activity distribution of conspiracy users across mainstream subreddits. We found instead that selecting subreddits based purely on where `r/conspiracy` users were most active outside their primary community (a case of selection on the outcome) produced an anomalous, non-decreasing activity distribution—unlike the typical long-tailed activity distributions seen on Reddit and other platforms (Valensise et al., 2019). Additional details on this selection rationale are provided in Appendix A.1.

We limited our analysis to the period from 2013 to 2023 (inclusive), discarding earlier activity. We made this decision because several mainstream subreddits considered, such as `r/ShowerThoughts` (created on February 5, 2012), were established well after the creation of `r/conspiracy`.

The final dataset contains approximately 510 million comments. From this corpus, we removed comments authored by known bots or suspicious accounts using predefined lists (Rollo et al., 2022). Furthermore, for each mainstream subreddit, we excluded users who posted fewer than 20 comments within that community. This threshold was applied to ensure that our analysis includes users with a strong and consistent signal of engagement. A more fine-grained breakdown of the collected data is presented in Table 2 in the Appendix.

### 3.2 Psycholinguistic Features and Dataset Construction

To characterize the psycholinguistic patterns in user comments from the selected mainstream subreddits, we computed user-level embeddings based on LIWC-22 (Boyd et al., 2022), a widely employed dictionary-based tool that extracts linguistic features associated with psychological, cognitive, and social processes. LIWC has been extensively applied in psychological and social media research to infer user traits and mental states from language use (Silva et al., 2021; Tausczik and Pennebaker, 2010). We applied the standard LIWC-22 processing pipeline to each comment collected during the data acquisition phase, resulting in 115 linguistic and psycholinguistic features per comment. We excluded five features related to punctuation, retaining 110 features for subsequent analysis.

To derive a single representation for each user within each mainstream subreddit, we computed the mean feature vector across all comments posted by that user in that community. This aggregation yields a user-level embedding that captures their overall psycholinguistic profile based on language use within the subreddit. This method allowed us to create comprehensive user representations that capture psycholinguistic traits and provide a unified framework for comparing language patterns across conspiracy and non-conspiracy users.

### 3.3 Experimental Design

We operationalize our research question by evaluating the ability of a Random Forest model to distinguish between the linguistic features of users active in the `r/conspiracy` subreddit and those of users who have not engaged with `r/conspiracy`, based on their language use within *mainstream subreddits*. Given the structured and interpretable nature

[3] https://www.reddit.com/best/communities/1

of our feature space, Random Forests provide a robust modeling approach (Haddouchi and Berrado, 2024). Moreover, our objective in this study is not to build a high-performance classifier per se, but to use the classifier as a proxy to measure the distinguishability of language.

The rationale behind this procedure is grounded in prior literature: several studies suggest that conspiracy theorists are influenced by a "conspiratorial mindset", a psychological predisposition to interpret the events happening in the world as the result of hidden machinations orchestrated by secret, malicious entities (Sutton and Douglas, 2020). The presence of this mindset is tied with the concept of the "Monological World View" (Sutton and Douglas, 2014; Swami et al., 2011), which suggests that believers in conspiracy theories perceive events in their lives through a markedly different lens compared to other individuals. This worldview, a defining feature of conspiracy theorists, is thought to pervade all aspects of their lives (Miani et al., 2022), leading us to hypothesize that it can also manifest in their linguistic patterns. Though we can not prove that the online users under study possess this type of mindset, our work offers meaningful support to this hypothesis.

In this work, we use the term conspiracy-engaged users to refer specifically to users who have interacted with the subreddit `r/conspiracy`. Thus, this operationalization captures observable engagement rather than underlying beliefs, and may overrepresent users willing to affiliate with this label while underrepresenting conspiratorial discourse expressed elsewhere. Users are assigned to the positive class if they authored at least one comment in the `r/conspiracy` subreddit; all other users are assigned to the negative class. This inclusive operationalization is intended to capture a broad set of users who engage with conspiracy-focused content; as shown in later analyses, restricting the positive class to higher levels of engagement yields comparable results.

For each mainstream subreddit, we represent each user with a single psycholinguistic feature vector computed by averaging LIWC-22 features across all of their comments within that subreddit. Thus, each user-subreddit pair constitutes one instance in the corresponding classification task. To construct the dataset for each mainstream subreddit, we first collect all users in the positive class who meet the activity threshold for that subreddit. We then randomly sample an equal number of negative-class users from the same subreddit to obtain a balanced dataset. To account for variability introduced by negative-class sampling, we repeat this process five times, resulting in five replicas that share the same positive-class users but differ in their negative-class composition. For each replica, we split users into disjoint training (80%) and test (20%) sets, ensuring that no user appears in both splits. Feature normalization and hyperparameter tuning are performed exclusively on the training data, using grid search with five-fold cross-validation. The final model is then evaluated on the held-out test set. We apply this procedure independently for each of the five replicas and for each mainstream subreddit, yielding multiple evaluations per subreddit that reflect uncertainty due to negative-class sampling.

Since both classes are represented in equal proportions, accuracy provides a clear measure of the classifier's effectiveness without being skewed by class imbalances, making it a suitable choice for assessing how well the classifier distinguishes between conspiracy and non-conspiracy users. Additionally, using accuracy allows for straightforward comparisons across different models and experimental setups, reinforcing the robustness of our findings.

For each classification task, we assess statistical significance by using a permutation test (Ojala and Garriga, 2010) in which training labels are randomly shuffled, and the model is retrained to generate a null distribution of performance (accuracy) scores. The performance of the original model is then compared against this distribution to test the null hypothesis of independence between features and labels. We perform 100 label permutations on the training data. The p-value is computed as $\frac{C+1}{n_{perm}+1}$, where $C$ is the number of permuted models that outperform the original model and $n_{perm}$ is the number of permutations. With 100 permutations, the minimum attainable p-value is 0.0099.

### 3.4 Feature Importance

In the context of Random Forest models, feature importance quantifies the contribution of each input variable to the predictive performance of the model. This metric is crucial for understanding which features are most influential in driving the model's decisions. Traditionally, feature importance in Random Forests is assessed via metrics such as Gini importance or mean decrease in impurity. However, these methods can sometimes be biased or difficult

to interpret in the presence of correlated features. To address these limitations, SHAP (SHapley Additive exPlanations) values offer a more robust and interpretable approach, providing a unified measure of feature importance by considering the contribution of each feature to the prediction of individual instances or a group of instances (Lundberg and Lee, 2017). We first gather all the models produced in the main study and compute the absolute mean SHAP values for each model. This process results in a 110-dimensional feature vector for each model, corresponding to the same psycholinguistic features employed to represent users in the classification experiment. Each value in these vectors reflects the global importance of a given feature, calculated as the mean absolute SHAP value for that feature across all samples. To obtain these SHAP values, we sample 700 positive instances from the dataset. As a result, for each subreddit, we derive a vector representing the importance of each classification feature within that subreddit, effectively mapping it into a 110-dimensional feature space. Finally, we measure similarity between these vectors using cosine similarity, a robust metric in this context. Given the construction of these vectors, cosine similarity is particularly well-suited, as it prevents issues arising from scalar multiples. The clustering in Figure 3 is computed with the UPGMA hierarchical clustering algorithm (Sokal and Michener, 1958).

### 3.5 Additional Subreddits as Positive Class

To assess the robustness of our findings, we extend our analysis to two additional communities on Reddit: `r/AskReddit` and `r/MensRights`. `r/AskReddit` is one of the platform's largest and most active subreddits, where users post open-ended questions intended to spark broad discussions. Its generalist nature attracts a diverse user base, making it a useful benchmark for typical engagement patterns on Reddit. `r/MensRights`, by contrast, is a more ideologically focused community that discusses issues perceived to affect men, including legal bias, custody disputes, and gender norms (De Candia et al., 2022). With over 300 000 subscribers, it is also part of the broader "manosphere", a network of online communities that has been shown to overlap with conspiracy discourse and anti-establishment narratives (Mamié et al., 2021). Similarly to `r/conspiracy`, `r/MensRights` is a longstanding community that has sparked controversy for its content but has not yet been taken down by Reddit moderators. This is the main reason we chose this community for the comparison against `r/conspiracy`. In two separate experiments, we apply the same classification pipeline used for `r/conspiracy`. We analyze the language of users active in `r/AskReddit` and `r/MensRights`, based on their participation in top mainstream subreddits. Our goal is to determine whether the linguistic patterns of these users mirror those observed among conspiracy-engaged individuals—specifically, whether a machine learning model can reliably distinguish them from control users.

### 3.6 Controlling for socio-demographic effects

In designing our experiments, we exclude the potential influence of users' socio-demographic attributes on the results. We aim to ensure that the observed results can be attributed to conspiratorial engagement rather than background characteristics of the individuals. To test the plausibility of this assumption, we construct socio-demographic embeddings for all users active in `r/conspiracy`. This is achieved by extracting the complete set of comments produced by these users on Reddit, totaling more than two billion posts. For each user, we then rely on the subreddit embedding model introduced by Waller and Anderson (2021). These embeddings act as proxies for latent socio-demographic attributes (e.g., age, gender, political orientation), inferred from patterns of community co-participation rather than directly observed traits. We then use these embeddings to measure the difference in socio-demographic distribution across user pools. Our results show no meaningful correlation between these differences and changes in the accuracy of classification, thus providing empirical support for the decision to exclude socio-demographic attributes from the design of our experiments. More details on these analyses can be found in Appendix A.2.

## 4 Results

### 4.1 Binary Classification Experiments

Figure 2 presents the general results of the binary classification experiments across all subreddits. The median accuracy of a Random Forest classifier across subreddits is 0.87 (min = 0.78, max = 0.95), 37 p.p. higher than a random baseline classifier. This performance demonstrates that the classifier can effectively differentiate be-

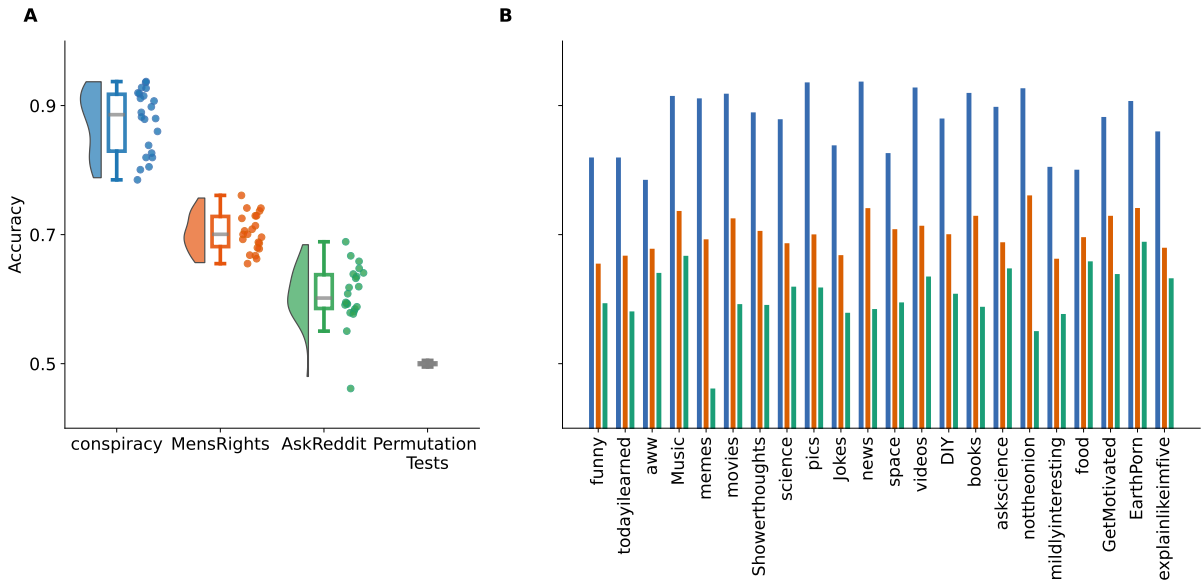



Figure 2: Results of the classification experiments across mainstream subreddits. **A:** Distribution of classification accuracies in distinguishing users who engage with conspiracy communities (in blue) and two control communities (`MensRights` in orange, `AskReddit` in green) achieved by a Random Forest classifier, trained and evaluated separately on each mainstream subreddit; each dot represents a subreddit (Median = 0.87, Min = 0.78, Max = 0.95 for `conspiracy`; Median = 0.69, Min = 0.65, Max = 0.76 for `MensRights`; Median = 0.60, Min = 0.46, Max = 0.68 for `AskReddit`. ). Accuracy from randomized permutation tests is also shown in grey (Median accuracy = 0.5). **B:** Classification accuracy of the Random Forest model broken down by subreddit, ordered by popularity (number of subscribers) in descending order, for the conspiracy community (in blue) and the two control communities (MensRights in orange, AskReddit in green).

tween the two user groups, highlighting that individuals active in conspiracy communities exhibit distinct psycholinguistic patterns compared to users in mainstream subreddits. These results hold for users who posted at least one comment in `r/conspiracy`. Varying the level of activity within the conspiracy community—from 10 to more than 100 comments—yields comparable classification performance (see Appendix A).

We test the robustness of the main results with three additional sets of experiments. First, we perform a randomized permutation test to determine whether the classification accuracy could be attributed to chance. As shown in panel A of Figure 2, the permutation test yields a median accuracy close to that of a random baseline classifier (0.5). The results are statistically significant for every subreddit ($p < 0.001$), confirming that the model's performance reflects meaningful patterns in the data rather than random variation.

Next, we extend our analysis by constructing alternative positive classes using two additional target communities: `r/MensRights` and `r/AskReddit`. This robustness check tests whether the identified psycholinguistic differences are unique to conspiracy communities or whether they also emerge in other ideologically-adjacent or general-interest forums. As shown in panel A of Figure 2, the classifier performs significantly worse in these settings compared to the original experiment using `r/conspiracy` as the positive class (Mann-Whitney test, $p < 0.001$), thus indicating that the linguistic signals associated with conspiracy engagement are more distinctive than those emerging from general-interest or ideologically-adjacent communities. Interestingly, the performance of the classifier is higher on `r/MensRights` than on `r/AskReddit`. The reason for this difference might be due to the similar extremist nature of `r/MensRights`: both conspiracy and manosphere users believe they are privy to hidden truths (e.g., the "red pill" metaphor) (Van Valkenburgh, 2021). These results reinforce our central claim that users who engage with conspiracy content exhibit language patterns that are significantly different from those of mainstream users.

Finally, we compare the performance of separate classifiers to a model trained on the combined data from all subreddits. While this aggregate model is more accurate than a random classifier baseline (accuracy = 0.76), we observe a drop in performance compared to subreddit-specific models. As

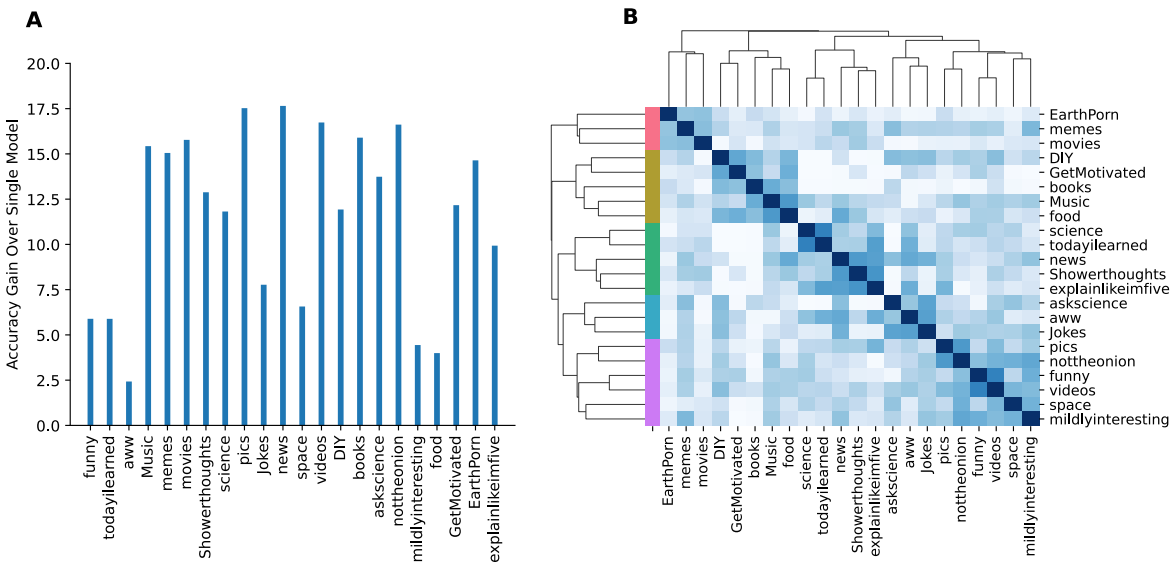



Figure 3: Analysis of linguistic fingerprints across different subreddits. **A:** Difference in accuracy (percentage points) between models trained separately for each subreddit and a single aggregate model trained on all subreddits. **B:** Clustering of mainstream subreddits based on the most important linguistic features used in the classification task.

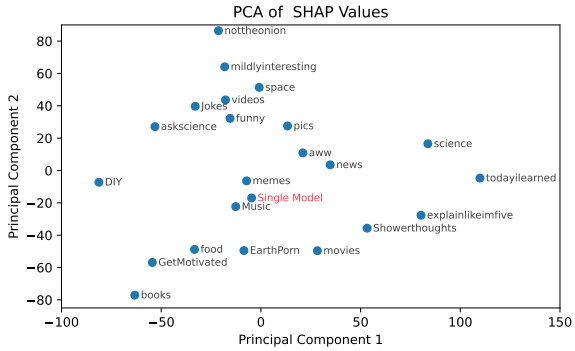



Figure 4: PCA of the SHAP values of the Single Model (Random Forest classifier) trained on the data from all the mainstream subreddits versus the SHAP values of the models trained on each single mainstream subreddit.

shown in Figure 3, panel **A**, the difference in accuracy with respect to the individual models ranges from 2.5 to 17 percentage points. This result suggests that conspiracy-engaged users adapt their language to the specific communities they participate in, rather than maintaining a uniform conspiratorial discourse across contexts. This finding aligns with established research showing that online users adjust to the linguistic norms of individual communities (Danescu-Niculescu-Mizil et al., 2013; Zhong et al., 2017), reinforcing the idea that conspiratorial tendencies are shaped by the social and discursive environments of each subreddit.

### 4.2 Temporal Robustness Analysis

We next examine the stability of classification performance over time prior to the users' first engagement with `r/conspiracy`. Specifically, we evaluate classifier performance on subsets of user comments ordered both chronologically and by cumulative activity within mainstream communities. Temporal analysis shows a small but statistically significant increase in accuracy—2.9 percentage points between the first and last activity windows (Mann-Kendall $p = 0.03$). In contrast, classification performance remains stable across cumulative activity thresholds (Mann-Kendall $p = 0.22$). Overall, these results indicate that the ability to distinguish conspiracy-engaged users is largely consistent regardless of the amount or timing of their prior mainstream activity. This also indicates that users' psycholinguistic patterns are relatively stable before engaging with conspiracy communities, consistent with the hypothesis that initial engagement is driven in part by self-selection rather than solely by exposure within the platform (Imhoff et al., 2022; Spohr, 2017). We leave more details and the plot of these results in the Appendix A.3.

### 4.3 Feature Importance Analysis

To identify the psycholinguistic features that characterize conspiracy-engaged users, we perform a SHAP analysis on subreddit-specific models, computing each feature's contribution to the classification task and comparing feature importance patterns across subreddits. By measuring their similarity, we evaluate whether conspiratorial language manifests consistently across mainstream communities or adapts to the specific norms of each context. As shown in Panel **B** of Figure 3, while some subreddits exhibit more similar feature importance patterns, we do not see clear separations across groups of communities, thus indicating substantial variation across different contexts. Additionally, in Figure 4, we show the results of applying Principal Component Analysis to the SHAP values of the Single Model, trained on data from all mainstream subreddits, as well as the models trained separately on each mainstream community, with all the users' activity thresholds aggregated. The figure illustrates that the Single Model is positioned near the origin ([0, 0]) in the PCA space, suggesting that it effectively represents an average of the community-specific models, which are more widely dispersed around the center. This last analysis reinforces the idea that the language of conspiracy-engaged users is highly context-dependent, adapting to the norms of each community rather than reflecting a fixed linguistic pattern.

To further investigate these results, we qualitatively analyze specific clusters that emerged from our prior analyses. As shown in Figure 3, Panel B, the hierarchical clustering algorithm identifies five distinct clusters. For example, the yellow/gold cluster—comprising the subreddits r/DIY, r/GetMotivated, r/books, r/Music, and r/food—shares emo_anger as a negative predictor. Specifically, a low presence of anger is associated with the non-conspiracy class, likely reflecting a tranquil environment where typical users discuss their hobbies, whereas conspiracy-leaning users display more anger in these spaces. Another notable example is the green cluster, which consists of r/science, r/todayilearned, r/news, r/Showerthoughts, and r/explainlikeimfive. Within this group, mental serves as a primary discriminant predictor, indicating a strong presence of reasoning and cognitive processes. This aligns with the high volume of explanatory language typical of these communities, making them the most likely venues for serious, real-world themes.

Overall, results from our study provide evidence for the existence of a large linguistic difference between conspiracy-engaged users and non-conspiracy-engaged users that can be captured in more mainstream spaces, also compared to other topic-adjacent subreddits. However, they also suggest that there are no universal psycholinguistic fingerprints for conspiracy users discussing mainstream topics, as their manifestation varies across different mainstream and conventional communities and is highly context-dependent. Nevertheless, we find recurring patterns such as negative emotions (Korenčić et al., 2024b) (anxiety, anger) and negative themes (death, illnesses), as shown in Table 1. We report the top-10 most discriminant features for each mainstream subreddit we considered in our experiments in Appendix B.3.

## 5 Discussion

This work shows that there are detectable psycholinguistic differences between conspiracy-engaged and non-conspiracy-engaged users that are captured within mainstream communities. Our results align with previous research (Klein et al., 2019), suggesting that users who engage with `r/conspiracy` display distinct linguistic patterns. While previous work identified these differences within communities already shaped by the activity of conspiracy theorists, we instead observe that such linguistic discrepancies emerge across various mainstream communities, with notable variations depending on the specific community. This result underscores both the complexity of this task and the critical role of social context in shaping the language of self-expression of individuals (Danescu-Niculescu-Mizil et al., 2011, 2013). Despite this variability, these linguistic fingerprints hold strong predictive power, enabling machine-learned models to achieve high classification accuracy, as shown in Figure 2.

These findings have important implications for the design of content moderation systems and the methodology of social media analysis. Specifically, the contrast between the robust performance of subreddit-specific models and the degradation of the aggregate model challenges the feasibility of "one-size-fits-all" detection approaches. We observed that applying a single model across all communities resulted in a performance drop of up to 17 percentage points compared to local models. This

Table 1: Examples of LIWC psycholinguistic features that appear with high, moderate, or low frequency as top discriminant features in the models trained on single subreddits. We also report the impact of that feature on the classification output. Positive: High values of the feature drive the classification towards conspiracy. Negative: High values of the feature drive the classification towards non-conspiracy. Mixed: The features have different roles for different models in which it appears.

| Frequency | LIWC Feature | Impact | Subreddit Examples |
|---|---|---|---|
| $\geq$ **11 models** | filler | Positive | explainlikeimfive, EarthPorn, todayilearned |
| | WC (Word Count) | Mixed | askscience, science, mildlyinteresting |
| | sexual | Mixed | askscience, Music, books |
| | swear | Positive | explainlikeimfive, science, DIY |
| **4 to 10 models** | illness | Mixed | news, explainlikeimfive, pics |
| | death | Mixed | funny, mildlyinteresting, pics |
| | emo_anger | Positive | food, DIY, GetMotivated |
| | emo_anx | Mixed | EarthPorn, movies, memes |
| | mental | Mixed | todayilearned, news, space |
| $\leq 3$ **models** | curiosity | Negative | jokes |
| | achieve | Positive | nottheonion |
| | Clout | Positive | books |

indicates that effective detection cannot rely on a universal linguistic "fingerprint"; rather, it requires context-aware strategies that account for how users adapt their linguistic register to the specific norms of the community they inhabit.

Furthermore, the pervasiveness of these signals impacts our understanding of "neutral" online spaces. We detected strong discriminatory markers even in innocuous, low-stakes environments such as *r/aww*, *r/food*, or *r/DIY*. This result challenges the assumption that ideological signaling is strictly compartmentalized within political echo chambers (Garimella et al., 2018; Cinelli et al., 2021) or radicalized forums (Calikus et al., 2025). Instead, our results suggest that mainstream, general-interest communities serve as heterogeneous mixing grounds where distinct linguistic identities remain visible and measurable, regardless of the topic being discussed.

From a methodological standpoint, this study establishes that local feature importance is superior to global feature aggregation for characterizing user groups. As illustrated by our PCA analysis of SHAP values, the single aggregate model essentially averages out community-specific nuances, obscuring the specific linguistic levers that distinguish these users in different contexts. Consequently, we suggest that future research in computational social science should adopt an "ecological" approach, characterizing user behavior relative to the specific discursive norms of the environment rather than seeking fixed, platform-wide behavioral signatures.

## Ethical Implications

The ethical implications of this work extend beyond content moderation and speak to longstanding tensions between freedom of expression, individual rights, and the protection of public discourse (Scanlon, 1972). A key ethical risk raised by our findings is the potential misuse of linguistic profiling at the user level. Although our models are intended for population-level analysis, similar techniques could be repurposed to infer ideological or psychological traits of individual users in mainstream, non-political communities (Colacrai et al., 2024). Such inferences are inherently probabilistic and context-dependent, and misclassification may result in unjustified scrutiny, stigmatization, or harm, particularly if applied without transparency, consent, or avenues for contestation. Moreover, the detectability of linguistic differences in general-interest communities increases the risk of preemptive or identity-based moderation strategies that target inferred user traits rather than observable harmful behavior. These approaches may produce negative effects on legitimate speech and disproportionately impact dissenting or minority viewpoints. For these reasons, the results presented here should not be used to justify early reactive interventions such as content removal or de-platforming. This caution is especially warranted given evidence that conspiracy theories are resilient to direct suppression and that heavy-handed enforcement may reinforce conspiratorial beliefs (Monti et al., 2023).

## Limitations

Our study has some limitations that we acknowledge and discuss. The first limitation is the reliance on the psycholinguistic features provided by dictionary-based computational methods. While these features are widely used in the literature (Tadesse et al., 2019; Faasse et al., 2016; Giachanou et al., 2023), and the dictionary is based on well-established psychological theories, they may still fail to capture the full nuances of online discourse. A dictionary-based approach primarily focuses on individual word frequencies, overlooking more complex linguistic structures such as sentence context, syntactic dependencies, and discourse-level features that could be crucial for distinguishing conspiratorial narratives from general discussions (Garten et al., 2018). Nonetheless, this tool was still able to provide meaningful insights into the language of online users, as the machine-learning models we employ in our studies achieve a very high prediction accuracy. Future work could test whether more powerful transformer-based strategies further increase the performance of the classifiers. Applying an interpretable framework to attention weights could capture linguistic structures and nuances that a dictionary-based method, such as LIWC, cannot measure. We expect that these endeavors would not fundamentally alter our core conclusion: that linguistic signals associated with conspiracy engagement are context-dependent across communities. Another limitation is that while our experimental design is based on a large-scale longitudinal dataset, it does not allow us to establish a causal relationship between the difference in user language and their engagement with the conspiracy community, nor is the directionality of this relation explored in this study. Finally, our approach focuses on users' direct activity, specifically, their commenting behavior, rather than their passive exposure to conspiratorial content, which we are unable to measure given the lack of available data. While comments serve as an observable proxy for user engagement, this limitation may overlook more subtle forms of exposure or interaction, such as lurking, which describes most of the activity of Internet users (Sun et al., 2014). Moreover, our strategy for identifying conspiracy-engaged users relies on self-identification and thus overlooks the cases in which these beliefs are shared in other communities without the user self-identifying as conspiracy-engaged. Nonetheless, despite this constraint, our method successfully identifies a substantial number of users who go on to participate in conspiracy communities, offering valuable insights into the behavioral patterns of conspiracy theorists based on their online self-expression.

## References


Asja Attanasio, Francesco Corso, Gianmarco De Francisci Morales, and Francesco Pierri. 2026. Effects of algorithmic visibility on conspiracy communities: Reddit after epstein's 'suicide'. In *Proceedings of the international AAAI conference on web and social media*, volume 20. (Cited on 1)

Abdul Basit. 2021. Conspiracy theories and violent extremism. *Counter Terrorist Trends and Analyses*, 13(3):1–9. (Cited on 1)

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. (Cited on 2)

Yochai Benkler, Robert Faris, and Hal Roberts. 2018. *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press. (Cited on 1)

Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, 10. (Cited on 3, 15)

Ece Calikus, Gianmarco De Francisci Morales, and Aristides Gionis. 2025. Who is at risk? analyzing the risk of radicalization among reddit users. In *European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 375–392. Springer. (Cited on 9)

Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the national academy of sciences*, 118(9):e2023301118. (Cited on 9)

Ernesto Colacrai, Federico Cinus, Gianmarco De Francisci Morales, and Michele Starnini. 2024. Navigating Multidimensional Ideologies with Reddit's Political Compass: Economic Conflict and Social Affinity. In *The ACM Web Conference*, WWW, pages 2582–2593. ACM. (Cited on 9)

Francesco Corso, Francesco Pierri, and Gianmarco De Francisci Morales. 2025a. Conspiracy theories and where to find them on TikTok. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8346–8362, Vienna, Austria. Association for Computational Linguistics. (Cited on 1, 2)

Francesco Corso, Francesco Pierri, and Gianmarco De Francisci Morales. 2025b. Do androids dream of unseen puppeteers? probing for a conspiracy mindset in large language models. *arXiv preprint arXiv:2511.03699*. (Cited on 1)

Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. 2011. Mark my words! linguistic style accommodation in social media. (Cited on 8)

Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: User lifecycle and linguistic change in online communities. (Cited on 7, 8)

Sara De Candia, Gianmarco De Francisci Morales, Corrado Monti, and Francesco Bonchi. 2022. Social Norms on Reddit: A Demographic Analysis. In *ACM Web Science Conference*, WebSci, pages 139–147. ACM. (Cited on 5)

Ahmad Diab, Rr. Nefriana, and Yu-Ru Lin. 2024. Classifying conspiratorial narratives at scale: False alarms and erroneous connections. *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1):340–353. (Cited on 2)

Karen M. Douglas and Robbie M. Sutton. 2023. What Are Conspiracy Theories? A Definitional Approach to Their Correlates, Consequences, and Communication. *Annual Review of Psychology*, 74(1):271–298. _eprint: https://doi.org/10.1146/annurev-psych-032420-031329. (Cited on 2)

Karen M. Douglas, Joseph E. Uscinski, Robbie M. Sutton, Aleksandra Cichocka, Turkay Nefes, Chee Siang Ang, and Farzin Deravi. 2019. Understanding Conspiracy Theories. *Political Psychology*, 40(S1):3–35. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/pops.12568. (Cited on 1, 2)

Adam M Enders, Joseph Uscinski, Casey Klofstad, and Justin Stoler. 2022. On the relationship between conspiracy theory beliefs, misinformation, and vaccine hesitancy. *Plos one*, 17(10):e0276082. (Cited on 1)

Kate Faasse, Casey J Chatman, and Leslie R Martin. 2016. A comparison of language use in pro-and anti-vaccination comments in response to a high profile facebook post. *Vaccine*, 34(47):5808–5814. (Cited on 10)

Marc Faddoul, Guillaume Chaslot, and Hany Farid. 2020. A Longitudinal Analysis of YouTube's Promotion of Conspiracy Videos. *arXiv preprint*. ArXiv:2003.03318 [cs]. (Cited on 1, 2)

Amos Fong, Jon Roozenbeek, Danielle Goldwert, Steven Rathje, and Sander Van Der Linden. 2021. The language of conspiracy: A psychological analysis of speech used by conspiracy theorists and their followers on twitter. *Group Processes & Intergroup Relations*, 24(4):606–623. (Cited on 2)

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Political Discourse on Social Media: Echo Chambers, Gatekeepers, and the Price of Bipartisanship. In *The ACM Web Conference*, WWW, pages 913–922. (Cited on 9)

Justin Garten, Joe Hoover, Kate M Johnson, Reihane Boghrati, Carol Iskiwitch, and Morteza Dehghani. 2018. Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis: Distributed dictionary representation. *Behavior research methods*, 50:344–361. (Cited on 10)

Anastasia Giachanou, Bilal Ghanem, and Paolo Rosso. 2023. Detection of conspiracy propagators using psycho-linguistic characteristics. *Journal of Information Science*, 49(1):3–17. (Cited on 10)

Hussam Habib, Padmini Srinivasan, and Rishab Nithyanand. 2022. Making a radical misogynist: How online social engagement with the manosphere influences traits of radicalization. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2). (Cited on 2)

Maissae Haddouchi and Abdelaziz Berrado. 2024. A survey and taxonomy of methods interpreting random forest models. *arXiv preprint arXiv:2407.12759*. (Cited on 4)

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. 2025. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326. (Cited on 14)

Roland Imhoff, Tisa Bertlich, and Marius Frenken. 2022. Tearing apart the "evil" twins: A general conspiracy mentality is not the same as specific conspiracy beliefs. *Current Opinion in Psychology*, 46:101349. (Cited on 7)

Colin Klein, Peter Clutton, and Adam G Dunn. 2019. Pathways to conspiracy: The social and linguistic precursors of involvement in reddit's conspiracy theory forum. *PloS one*, 14(11):e0225098. (Cited on 2, 8)

Damir Korenčić, Berta Chulvi, X Bonet Casals, Mariona Taulé, Paolo Rosso, and Francisco Rangel. 2024a. Overview of the oppositional thinking analysis pan task at clef 2024. *Working Notes of CLEF*. (Cited on 1)

Damir Korenčić, Berta Chulvi, Xavier Bonet Casals, Alejandro Toselli, Mariona Taulé, and Paolo Rosso. 2024b. What distinguishes conspiracy from critical narratives? a computational analysis of oppositional discourse. *Expert Systems*, 41(11):e13671. (Cited on 8)

Edoardo Loru, Alessandro Galeazzi, Anita Bonetti, Emanuele Sangiorgio, Niccolò Di Marco, Matteo Cinelli, Max Falkenberg, Andrea Baronchelli, and

Walter Quattrociocchi. 2025. Ideology and polarization set the agenda on social media. *Scientific Reports*, 15(1):35816. (Cited on 1)

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30. (Cited on 5)

Robin Mamié, Manoel Horta Ribeiro, and Robert West. 2021. Are anti-feminist communities gateways to the far right? evidence from reddit and youtube. In *Proceedings of the 13th ACM Web Science Conference 2021*, pages 139–147. (Cited on 5)

Marcel Meuer, Aileen Oeberst, and Roland Imhoff. 2023. How do conspiratorial explanations differ from non-conspiratorial explanations? A content analysis of real-world online articles. *European Journal of Social Psychology*, 53(2):288–306. Number: 2 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ejsp.2903. (Cited on 2)

Alessandro Miani, Thomas Hills, and Adrian Bangerter. 2022. Interconnectedness and (in)coherence as a signature of conspiracy worldviews. *Science Advances*, 8(43):eabq3668. Number: 43 Publisher: American Association for the Advancement of Science. (Cited on 4)

Corrado Monti, Matteo Cinelli, Carlo Valensise, Walter Quattrociocchi, and Michele Starnini. 2023. Online conspiracy communities are more resilient to deplatforming. *PNAS Nexus*, 2(10):pgad324. (Cited on 1, 9)

Teresa K. Naab and Constanze Küchler. 2023. *Content Analysis in the Research Field of Online User Comments*, pages 441–450. Springer Fachmedien Wiesbaden, Wiesbaden. (Cited on 3)

Markus Ojala and Gemma C Garriga. 2010. Permutation tests for studying classifier performance. *Journal of machine learning research*, 11(6). (Cited on 4)

Shruti Phadke, Mattia Samory, and Tanushree Mitra. 2022. Pathways through conspiracy: the evolution of conspiracy radicalization through engagement in online conspiracy discussions. (Cited on 2)

Cesare Rollo, Gianmarco De Francisci Morales, Corrado Monti, and André Panisson. 2022. Communities, gateways, and bridges: Measuring attention flow in the reddit political sphere. In *International Conference on Social Informatics*, pages 3–19. Springer. (Cited on 2, 3)

Mattia Samory and Tanushree Mitra. 2018a. Conspiracies Online: User Discussions in a Conspiracy Community Following Dramatic Events. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1). Number: 1. (Cited on 2)

Mattia Samory and Tanushree Mitra. 2018b. 'the government spies using our webcams': The language of conspiracy theories in online discussions. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW). (Cited on 1)

Thomas Scanlon. 1972. A theory of freedom of expression. *Philosophy & Public Affairs*, pages 204–226. (Cited on 9)

Amila Silva, Pei-Chi Lo, and Ee Peng Lim. 2021. On predicting personal values of social media users using community-specific language features and personal value correlation. (Cited on 3)

R Sokal and C Michener. 1958. A statistical method for evaluating systematic relationships: The university of kansas science bulletin, v. 38. *Sokal104938University of Kansas Science Bulletin1958*, pages 1049–1438. (Cited on 5)

Dominic Spohr. 2017. Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Business information review*, 34(3):150–160. (Cited on 7)

Kate Starbird. 2017. Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 230–239. (Cited on 1)

Na Sun, Patrick Pei-Luen Rau, and Liang Ma. 2014. Understanding lurkers in online communities: A literature review. *Computers in Human Behavior*, 38:110–117. (Cited on 10)

Robbie M Sutton and Karen M Douglas. 2014. 14 examining the monological nature of conspiracy theories. *Power Polit. Paranoia Why People Are Suspicious Their Lead*, 29:254–272. (Cited on 4)

Robbie M Sutton and Karen M Douglas. 2020. Conspiracy theories and the conspiracy mindset: implications for political ideology. *Current Opinion in Behavioral Sciences*, 34:118–122. (Cited on 4)

Viren Swami, Rebecca Coles, Stefan Stieger, Jakob Pietschnig, Adrian Furnham, Sherry Rehim, and Martin Voracek. 2011. Conspiracist ideation in britain and austria: Evidence of a monological belief system and associations between individual psychological differences and real-world and fictitious conspiracy theories. *British Journal of Psychology*, 102(3):443–463. (Cited on 4)

Michael M Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2019. Detection of depression-related posts in reddit social media forum. *Ieee Access*, 7:44883–44893. (Cited on 10)

Timothy R. Tangherlini, Shadi Shahsavari, Behnam Shahbazi, Ehsan Ebrahimzadeh, and Vwani Roychowdhury. 2020. An automated pipeline for the discovery of conspiracy and conspiracy theory narrative frameworks: Bridgegate, Pizzagate and storytelling on the web. *PLOS ONE*, 15(6):e0233879. Publisher: Public Library of Science. (Cited on 1, 2)

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54. (Cited on 3)

Carlo Michele Valensise, Matteo Cinelli, Alessandro Galeazzi, and Walter Quattrociocchi. 2019. Drifts and shifts: characterizing the evolution of users interests on reddit. *arXiv preprint arXiv:1912.09210*. (Cited on 3)

Jan-Willem van Prooijen and Karen M Douglas. 2017. Conspiracy theories as part of history: The role of societal crisis situations. *Memory Studies*, 10(3):323–333. Publisher: SAGE Publications. (Cited on 2)

Shawn P Van Valkenburgh. 2021. Digesting the red pill: Masculinity and neoliberalism in the manosphere. *Men and masculinities*, 24(1):84–103. (Cited on 6)

Isaac Waller and Ashton Anderson. 2021. Quantifying social organization and political polarization in online platforms. *Nature*, 600(7888):264–268. (Cited on 5, 17)

Changtao Zhong, Hau-wen Chang, Dmytro Karamshuk, Dongwon Lee, and Nishanth Sastry. 2017. Wearing many (social) hats: How different are your different social network personae? (Cited on 7)

# A Effect of engagement level in `r/conspiracy` on linguistic patterns

We conduct a stratified analysis to examine how varying levels of activity in `r/conspiracy` influence model performance. This analysis is motivated by one of our assumptions, which posits a potential link between the level of engagement in a conspiracy community and the degree of linguistic difference exhibited by an online user. To test this, we divide users active on `r/conspiracy` into four groups based on their number of comments: (0, 1], (1, 10), [10, 100), and [100, $+\infty$). For each group, we construct a new dataset composed of the psycholinguistic feature vectors of conspiracy users within that group, along with a balanced random sample of non-conspiracy users. We then replicate the classification pipelines from the main, obtaining results consistent with those from the aggregated analyses. Figure 5 shows the results of the experiments, disaggregated by the activity thresholds defined in Section 3. As noted, model performance is consistent across the different levels of user activity on `r/conspiracy`. These results are supported by significant permutation tests ($p < 0.01$) across all activity classes and experimental conditions.

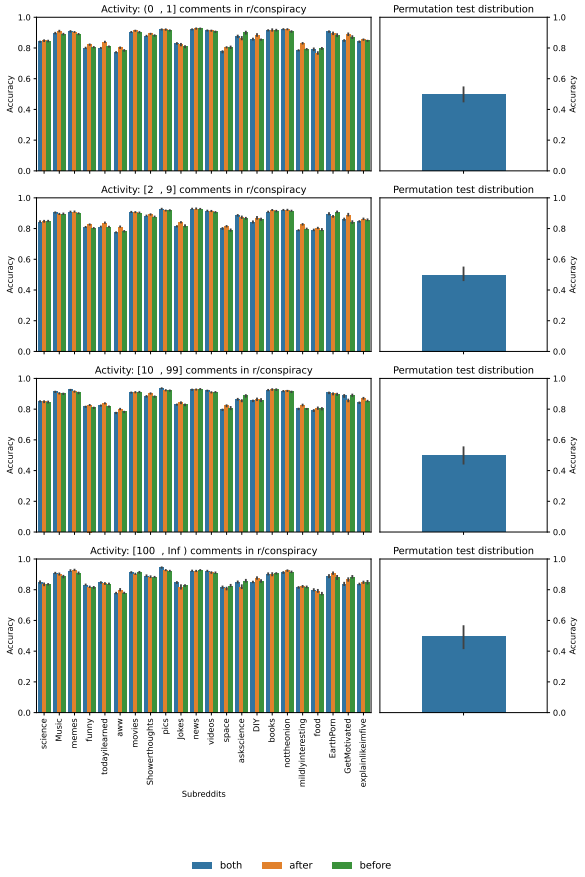



Figure 5: Disaggregated classification performance on the test set, together with the permutation tests performance for the four different classes of activity.

## A.1 Avoiding bias

Figure 6 shows the result of a preliminary experiment we performed to select the non-conspiracy subreddit to which to focus our analyses. We found that if we employed the activity outside of `r/conspiracy` of conspiracy users as a driver for this search, we would incur bias, as we would be selecting subreddits that would not correctly represent the pattern of activity of normal Reddit users, as shown in the figure.

## A.2 Additional Socio Demographics Analysis

As described in Section 3, we are interested in measuring the interplay of socio-demographics and accuracy of the classification. To do so, we represented each user by computing a weighted average of the embeddings of all subreddits in which they were active, where the weights correspond to the relative frequency of their contributions across subreddits. This procedure yielded a vector representation for each user, intended to capture socio-demographic and cultural orientations encoded in subreddit participation patterns.

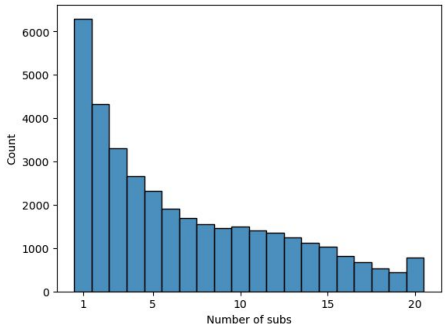



Figure 6: Example of bias in the selection of subreddits using the activity of conspiracy users as a proxy. The users we consider are those with at least 100 comments on `r/conspiracy`.

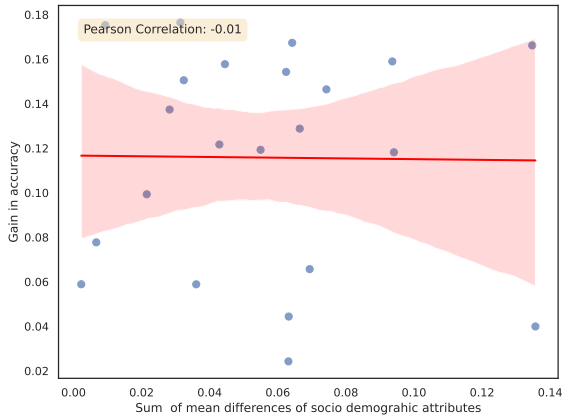



Figure 7: Correlation between the gain in accuracy of the subreddit-specific classifier on the general classifier and the absolute difference between the conspiracy users' attribute distribution and the given subreddit embedding.

To evaluate whether these embeddings introduced systematic bias into our analyses, we compared the distribution of conspiracy-engaged users active within a given mainstream subreddit to the embedding of that same subreddit. Next, we quantified the absolute difference between the mean vector of conspiracy users and the embedding of each subreddit. We then correlated this measure with the improvement in classification accuracy obtained by the subreddit-specific model (described by Figure 3 in the Results section). The results, shown in Figure 7, revealed no meaningful correlation, suggesting that socio-demographic features, as captured by subreddit embeddings, do not significantly account for the performance gains of our classifier. This provides empirical support for the decision to exclude socio-demographic attributes from the core design of our study.

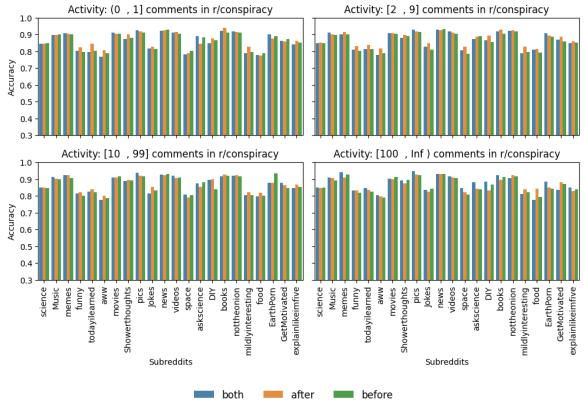



Figure 8: Disaggregated TabPFN experimental results.

### A.3 Temporal Robustness Check

As introduced in Section 4.2, we perform additional robustness checks on the classification results we obtained in our experiments by segmenting the users' activity into quintiles, based on two criteria: the temporal order and the cumulative order. Meaning, we bin the comments of users into five different groups based on the moment in time and in which order they were posted by the user during their activity on Reddit prior to their first comment on `r/conspiracy`.

We apply this framework to each mainstream subreddit in our collection and for each group of users with varying levels of activity on `r/conspiracy`. The datasets generated from these classification setups are balanced, and we employ them to train and evaluate a series of Random Forest classifiers, one for each top subreddit in our collection. We then repeat the same operations we described in the previous analyses, i.e., feature normalization, hyperparameter search, and model training.

### A.4 Top features for each mainstream subreddit

In Appendix B.3 we show the top 10 most discriminant features in the decision tree together with the relative shap value, indicating the contribution of that feature, whether positive or negative, to the classification result. As we describe in Section 4, there is no unique pattern common across all the communities, even though some features are represented many times in these plots.

## B Additional Robustness Checks

We employed TabPFN (Hollmann et al., 2025), a state-of-the-art machine learning architecture, to

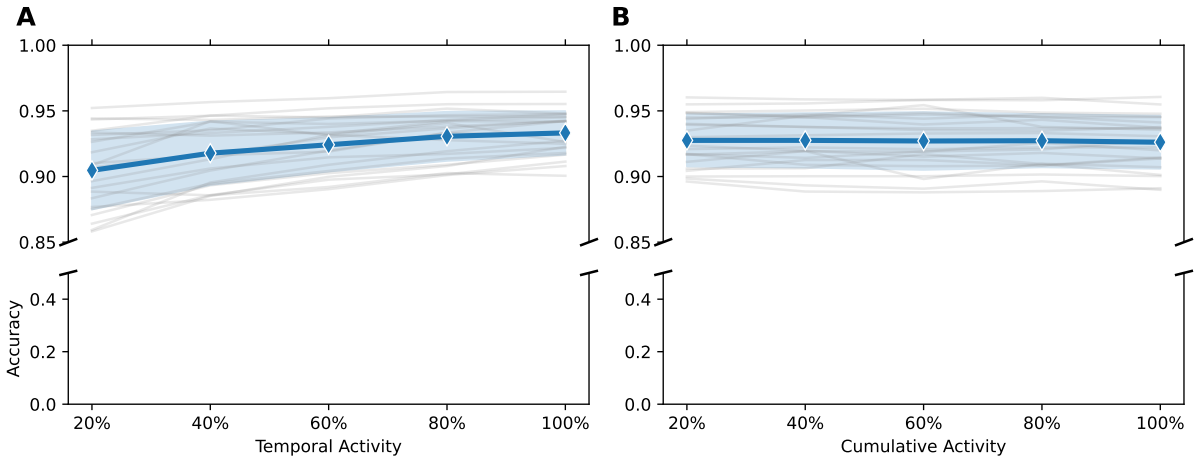



Figure 9: **Results of additional classification experiments on temporal robustness. A**: Classification accuracy across different temporal activity windows (based on comment order). **B**: Classification accuracy across different numerical activity windows (based on cumulative comment volume). In both cases, accuracy remains stable across thresholds, indicating no significant linguistic shift over time.

further validate the robustness of our results. To meet TabPFN's input constraints, we downsampled each dataset to approximately 10 000 instances. We then replicated the classification pipelines from the main experiments using this model. The results closely matched those obtained with the Random Forest classifier, reinforcing the meaningfulness of the psycholinguistic embeddings.

### B.1 TabPFN Results

In Figure 8 we show the results of the experiments, disaggregated by the activity thresholds defined in Section 3. We find the results to be in line with the ones obtained with a Random Forest model in the experiments presented in Section 4. Once again, we show how the volume of activity in `r/conspiracy` only marginally influences the accuracy of the classification.

### B.2 Computational Resources

We ran our experiments on a machine with: a 64-core CPU, 256 GB of RAM, and an NVIDIA A100 GPU. The dataset filtering required a week of execution. The feature extraction required five days of execution. Model training and evaluation required a week of execution. We leave the repository at the following link: https://anonymous.4open.science/r/reddit_ct-EF52/

### B.3 Description of important features

We report here the descriptions of the notable features we showed in Table 1, extracting them from the LIWC22 documentation (Boyd et al., 2022). The following descriptions contain examples from the categories, which can be swears or offensive words. Readers

- **filler**: This category captures conversational filler words. Frequently used examples include words like "I mean", "wow", "sooo", and "youknow".
- **WC (Word Count)**: This is a summary variable that simply represents the total word count of the analyzed text.
- **sexual**: This is an expanded dictionary category that captures sexual language. The most frequently used exemplars include terms like "sex", "gay", "pregnan*", and "dick".
- **swear**: This category captures swear words, with frequent examples including "shit", "fuckin", "fuck", and "damn". In LIWC-22, swear words are conceptualized as part of the overall affect and tone dictionaries (rather than strictly negative emotion), as their usage has evolved and is now just as likely to reflect positive sentiment in informal contexts.
- **illness**: A subcategory of the broader Health dimension, this captures disease names and physical symptoms related to illness. Common examples include "hospital", "cancer*", "sick", and "pain".
- **death**: This category captures language referring to mortality. Frequently used words include "death", "dead", "die", and "kill".
- **emo_anger**: This represents the emotion of anger, restricted strictly to true emotion labels

or words that strongly imply the emotion. Examples include “hate”, “mad”, “angry”, and “frustr*”.

- **emo_anx**: This represents the emotion of anxiety, similarly restricted to clear emotion labels and strong implications of anxiety. Frequently used exemplars include “worry”, “fear”, “afraid”, and “nervous”.
- **mental**: This is a mental health category that typically identifies psychological diagnoses (e.g., “bipolar”, “neurosis”) or related behaviors (e.g., “suicide”, “addiction”). Other frequent examples include “depressed” and “trauma”.
- **curiosity**: Falling under the “Motives” dimension, these words reflect an author’s search for or interest in new knowledge or experiences, which is thought to correlate with the personality trait of openness. Examples include “scien*”, “look for”, “research”, and “wonder”.
- **achieve**: Part of the psychological “Drives” dimension, this category captures language related to achievement. Frequently used examples include “work”, “better”, “best”, and “working”.
- **clout**: This is one of the four summary variables in LIWC-22, representing the language of leadership and social status.

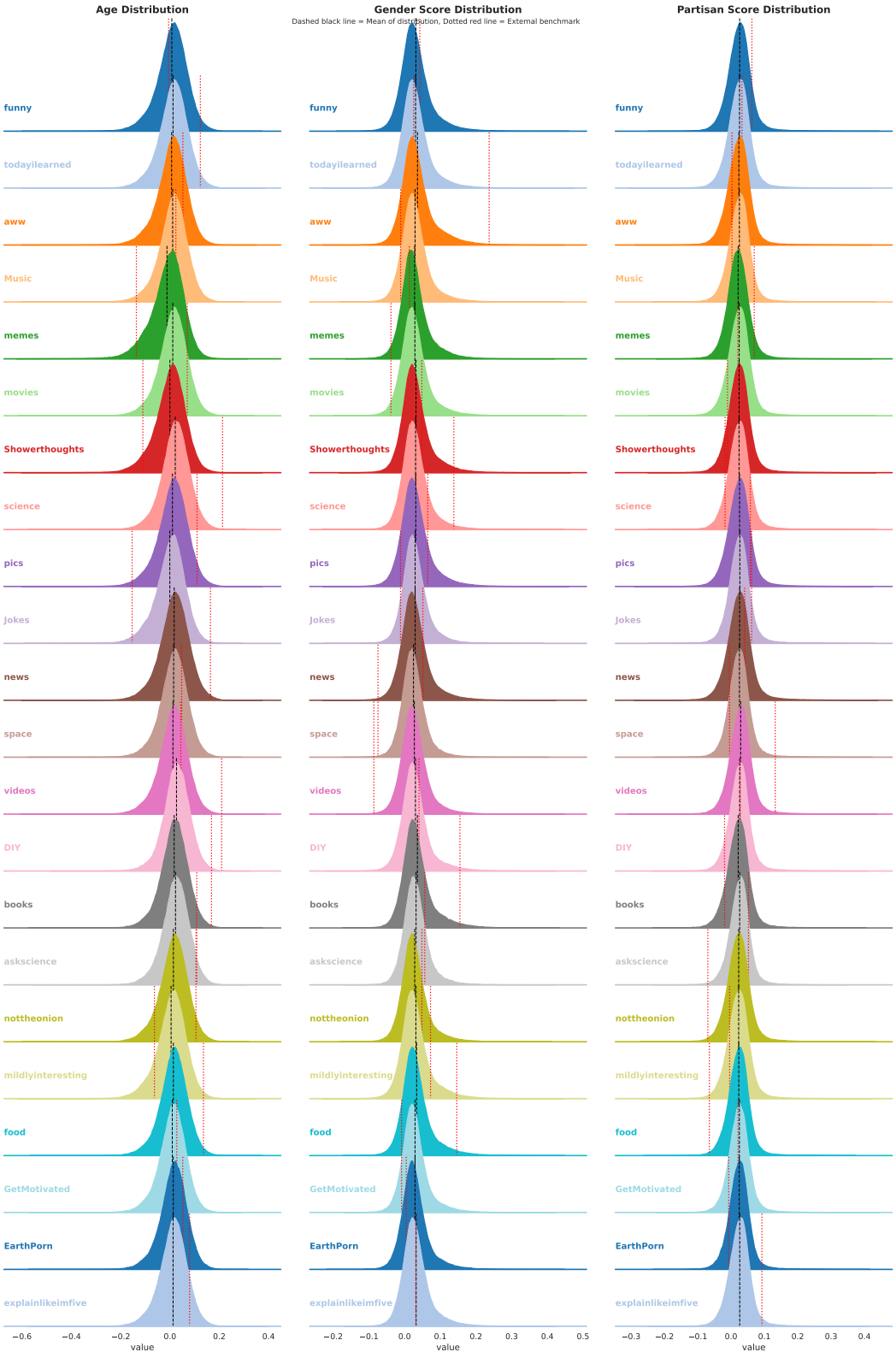


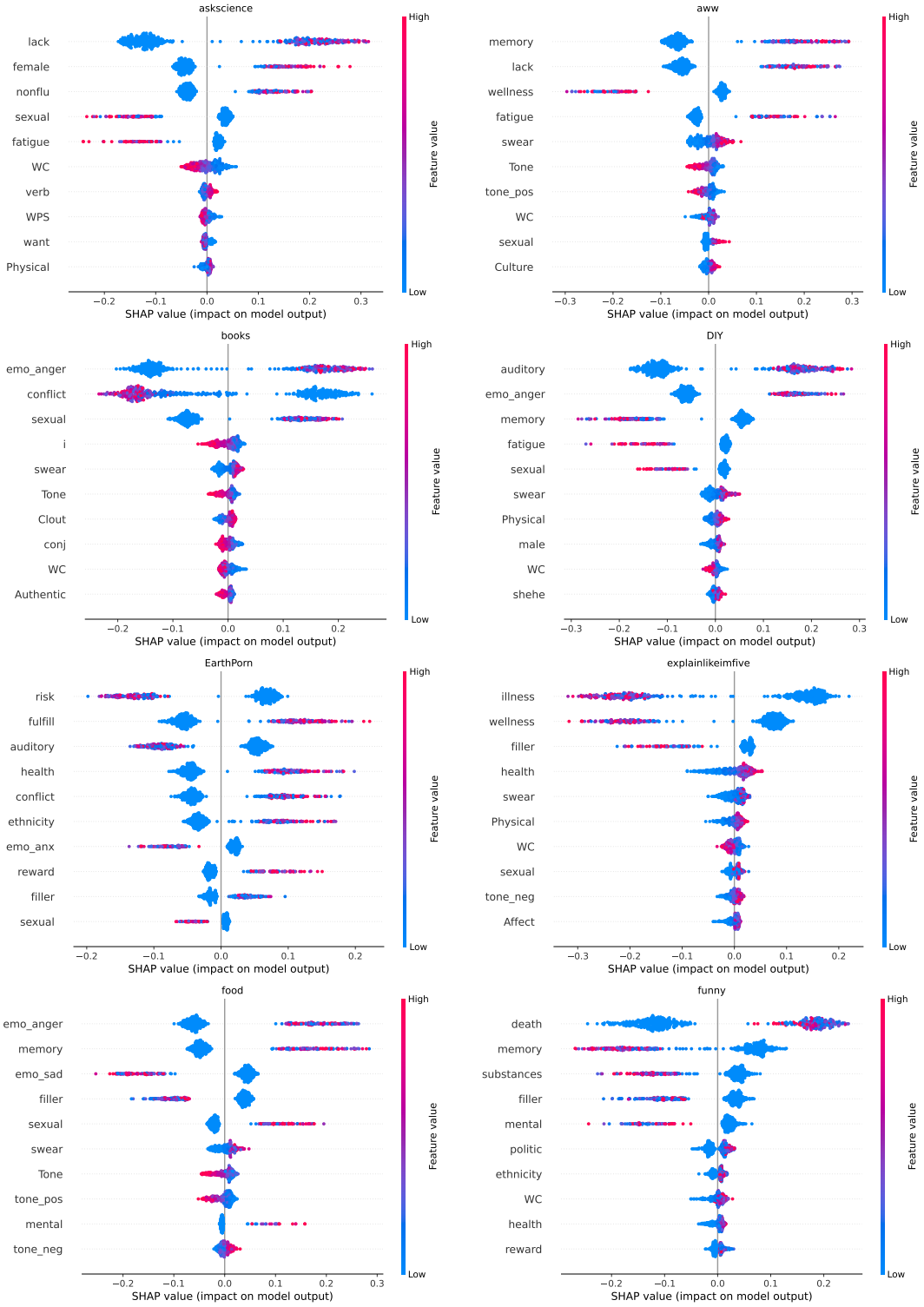

Figure 10: Distribution the socio-demographic scores of conspiracy users active in each mainstream subreddit. Black Dashed Line: Conspiracy mean, Red Dashed Line: Subreddit Score Mean, provided by (Waller and Anderson, 2021)

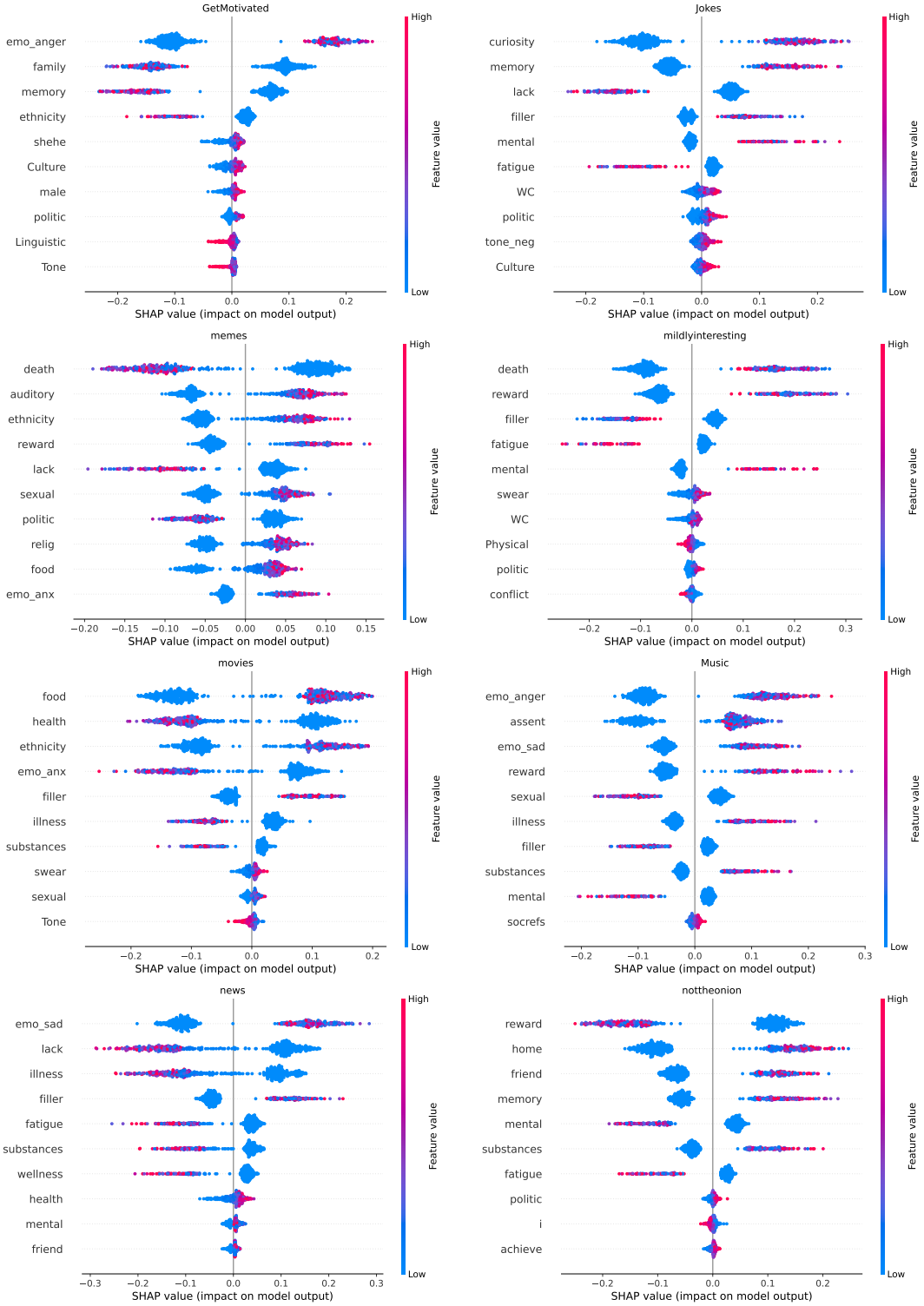
Table 2: Data sizes for each subreddit, in millions of comments. We consider the comments of users who have made at least 20 contributions to the subreddit.

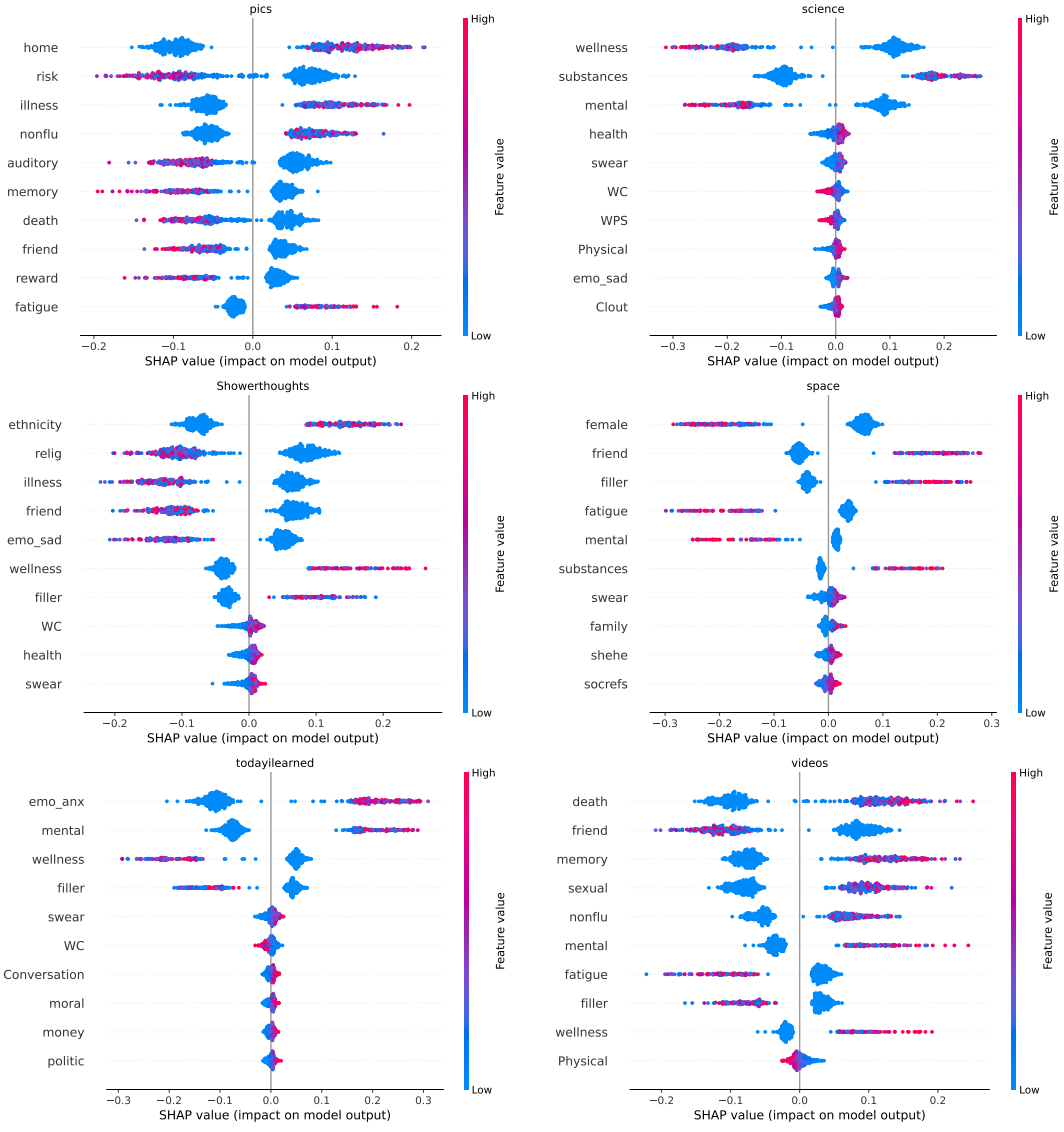
| Subreddit | # comments of CT users | # comments of NOCT users | Total |
|---|---|---|---|
| r/memes | 5.67 | 76.65 | 82.32 |
| r/todayilearned | 17.58 | 39.11 | 46.69 |
| r/Showerthoughts | 6.04 | 23.38 | 29.42 |
| r/nottheonion | 3.99 | 8.74 | 12.73 |
| r/aww | 5.29 | 29.0 | 34.29 |
| r/Music | 3.23 | 12.01 | 15.24 |
| r/movies | 12.41 | 37.76 | 50.17 |
| r/science | 3.67 | 7.94 | 10.61 |
| r/pics | 24.31 | 68.76 | 93.07 |
| r/Jokes | 1.71 | 7.55 | 9.26 |
| r/news | 24.54 | 41.38 | 65.96 |
| r/videos | 14.82 | 32.93 | 47.75 |
| r/space | 1.77 | 4.06 | 5.83 |
| r/askscience | 0.80 | 2.8 | 3.6 |
| r/DIY | 0.88 | 3.51 | 4.39 |
| r/books | 1.71 | 8.35 | 10.06 |
| r/mildlyinteresting | 5.26 | 21.2 | 26.46 |
| r/food | 1.37 | 6.48 | 7.85 |
| r/EarthPorn | 0.6 | 2.76 | 3.36 |
| r/GetMotivated | 0.65 | 2.18 | 2.83 |
| r/explainlikeimfive | 3.52 | 11.23 | 14.75 |
| r/funny | 22.56 | 76.26 | 88.82 |

Table 3: Dataset sizes for each mainstream subreddit in number of users. The datasets are balanced, so the number of conspiracy users is (approximately) equal to the number of non-conspiracy users.

| Subreddit | (0,1] | (1,10) | [10,100) | [100,∞) | Total |
|---|---|---|---|---|---|
| r/funny | 88 362 | 121 808 | 74 412 | 23 250 | 307 832 |
| r/memes | 24 195 | 35 208 | 23 456 | 6072 | 88 931 |
| r/science | 15 564 | 26 124 | 19 187 | 8443 | 69 318 |
| r/Music | 15 849 | 24 982 | 18 147 | 6674 | 65 652 |
| r/todayilearned | 65 644 | 99 640 | 64 532 | 21 908 | 251 724 |
| r/aww | 29 886 | 43 720 | 27 650 | 8719 | 109 975 |
| r/movies | 41 689 | 62 096 | 41 176 | 14 283 | 159 244 |
| r/Showerthoughts | 29 506 | 46 913 | 32 209 | 10 140 | 118 768 |
| r/pics | 94 883 | 137 573 | 85 882 | 26 635 | 344 973 |
| r/Jokes | 8322 | 12 326 | 8688 | 3114 | 32 450 |
| r/news | 69 170 | 110 814 | 75 465 | 27 734 | 283 183 |
| r/videos | 53 849 | 80 752 | 52 473 | 17 093 | 204 167 |
| r/space | 6856 | 11 245 | 8394 | 3416 | 29 911 |
| r/askscience | 3011 | 4662 | 3056 | 1182 | 11 911 |
| r/DIY | 3860 | 5796 | 3872 | 1544 | 15 072 |
| r/books | 8326 | 12 141 | 8507 | 3061 | 32 035 |
| r/nottheonion | 18 078 | 31 548 | 23 808 | 8832 | 82 266 |
| r/mildlyinteresting | 28 254 | 43 890 | 30 176 | 9734 | 112 054 |
| r/food | 6649 | 10 356 | 7048 | 2598 | 26 651 |
| r/EarthPorn | 2158 | 3412 | 2302 | 949 | 8821 |
| r/GetMotivated | 2321 | 3788 | 3059 | 1154 | 10 322 |
| r/explainlikeimfive | 14 155 | 21 875 | 15 700 | 5680 | 57 410 |


askscience
lack
female
nonflu
sexual
fatigue
WC
verb
WPS
want
Physical
SHAP value (impact on model output)
Feature value
High
Low
aww
memory
lack
wellness
fatigue
swear
Tone
tone_pos
WC
sexual
Culture
SHAP value (impact on model output)
Feature value
High
Low
books
emo_anger
conflict
sexual
i
swear
Tone
Clout
conj
WC
Authentic
SHAP value (impact on model output)
Feature value
High
Low
DIY
auditory
emo_anger
memory
fatigue
sexual
swear
Physical
male
WC
shehe
SHAP value (impact on model output)
Feature value
High
Low
EarthPorn
risk
fulfill
auditory
health
conflict
ethnicity
emo_anx
reward
filler
sexual
SHAP value (impact on model output)
Feature value
High
Low
explainlikeimfive
illness
wellness
filler
health
swear
Physical
WC
sexual
tone_neg
Affect
SHAP value (impact on model output)
Feature value
High
Low
food
emo_anger
memory
emo_sad
filler
sexual
swear
Tone
tone_pos
mental
tone_neg
SHAP value (impact on model output)
Feature value
High
Low
funny
death
memory
substances
filler
mental
politic
ethnicity
WC
health
reward
SHAP value (impact on model output)
Feature value
High
Low


GetMotivated
emo_anger
family
memory
ethnicity
shehe
Culture
male
politic
Linguistic
Tone
SHAP value (impact on model output)
Feature value
High
Low
Jokes
curiosity
memory
lack
filler
mental
fatigue
WC
politic
tone_neg
Culture
memes
death
auditory
ethnicity
reward
lack
sexual
politic
relig
food
emo_anx
mildlyinteresting
death
reward
filler
fatigue
mental
swear
WC
Physical
politic
conflict
movies
food
health
ethnicity
emo_anx
filler
illness
substances
swear
sexual
Tone
Music
emo_anger
assent
emo_sad
reward
sexual
illness
filler
substances
mental
socrefs
news
emo_sad
lack
illness
filler
fatigue
substances
wellness
health
mental
friend
nottheonion
reward
home
friend
memory
mental
substances
fatigue
politic
i
achieve


pics
home
risk
illness
nonflu
auditory
memory
death
friend
reward
fatigue
−0.2 −0.1 0.0 0.1 0.2
SHAP value (impact on model output)
High
Feature value
Low
science
wellness
substances
mental
health
swear
WC
WPS
Physical
emo_sad
Clout
−0.3 −0.2 −0.1 0.0 0.1 0.2
SHAP value (impact on model output)
High
Feature value
Low
Showerthoughts
ethnicity
relig
illness
friend
emo_sad
wellness
filler
WC
health
swear
−0.2 −0.1 0.0 0.1 0.2
SHAP value (impact on model output)
High
Feature value
Low
space
female
friend
filler
fatigue
mental
substances
swear
family
shehe
socrefs
−0.3 −0.2 −0.1 0.0 0.1 0.2 0.3
SHAP value (impact on model output)
High
Feature value
Low
todayilearned
emo_anx
mental
wellness
filler
swear
WC
Conversation
moral
money
politic
−0.3 −0.2 −0.1 0.0 0.1 0.2 0.3
SHAP value (impact on model output)
High
Feature value
Low
videos
death
friend
memory
sexual
nonflu
mental
fatigue
filler
wellness
Physical
−0.2 −0.1 0.0 0.1 0.2
SHAP value (impact on model output)
High
Feature value
Low