Language, Identity, and Polarization: A Text Analysis of Reddit Communities

Mohamed Fazil Tajudeen
Trinity College Dublin
24357225

Zoe Passiadou
Trinity College Dublin
24361712

Mishel Theckanath
Trinity College Dublin
24342527

Udayanarayana Namagondlu
Trinity College Dublin
24338926

Vishaalini Ramasamy
Trinity College Dublin
24340813

Abstract

This study examines how linguistic markers particularly pronoun usage and psychological language features vary across thematically distinct Reddit communities. Using a dataset collected via the PRAW API, we scraped posts and top-level comments from twelve subreddits categorized into Politics, Mental Health & Life, Discussion & Opinions, and Hobbies Education. Our analysis focused on the frequency and distribution of first-person singular, inclusive and exclusive pronouns, complemented by sentiment and psychological feature extraction, such as tentative/assertive tone, emotional language and politeness markers. Statistical and vector-based similarity metrics, including Chi-square, Fisher's exact test, Cosine similarity and Jaccard similarity, were used to evaluate linguistic divergence between community types. Results indicate that Mental Health subreddits exhibited higher use of first-person pronouns and emotional/polite expressions, suggesting introspective and supportive discourse. In contrast, political communities showed more assertiveness, collective alignment and elevated exclusive pronoun usage, indicative of ideological positioning and group-based framing. These findings contribute to the understanding of identity construction, emotion and social dynamics in online discourse and highlight the role of linguistic style in shaping community norms and polarization.

Keywords :

Language and Identity, Reddit Communities, Linguistic Patterns, Group Identity Markers, Polarization, Social Dynamics, Web Scraping, Natural Language Processing, Online Communities, Pronoun Analysis, Cosine Similarity, Jaccard Similarity, Chi-Square Test, Fisher's Exact Test

1 Introduction

In recent years, online communities have begun to emerge as important sites of social interaction, identity-forming, and ideological expression. It is in these platforms like Reddit, which are structured around thematic subcommunities or ”subreddits,” that we see a rich context for analyses of how language reflects and shapes group dynamics, emotional expression, and social polarization. Language is not only a means of communication but also a means by which individuals and groups construct identities, negotiate belonging, and engage in ideological discourse.

This study examines linguistic markers that differentiate between communities on Reddit through their uses of pronoun and psychological linguistic features. It examines the uses of first-person singular, inclusive, and exclusive pronouns in subreddits of different types that signify personal, collective, and oppositional stances. It further examines other psychological and stylistic features such as emotionality, assertiveness, politeness, and lexical variation with a view of deepening the understanding of the communicative styles of these communities.

Applying a blend of frequency analysis, Chi-square tests, Fisher's exact test, Cosine similarity, and Jaccard similarity, the research quantifies linguistic inclinations that are tested for significance among political, mental health, discussion, and hobby subreddits. When contrasting these linguistic behaviors, the research aims to reveal how online communities vary in the uses of language for identity, emotion, and group identification in a manner that enables broad understandings of digital communication and social structure.

2 Related Work

The discoveries for this paper serve to expand a great deal of past research on linguistic markers of identity, emotion, and polarization in online communities, especially when the studies engage frequency-based analyses, test for statistical significance, and measure distributional similarity.

Linguistic markers of identity

Linguistic features have time and time again been shown effective for identity shaping of groups and persons in text communication. Argamon et al. [2] found that stylistic variation in formal text relies on factors of social identity such as gender, and Pennebaker and Stone [18] demonstrated the way in which language style can vary over an individual's life span in marking developmental and intellectual change. Likewise, through predictions based on syntactic features, Luyckx and Daelemans [15] too established the strong relationship between the application of language and identity. Even Humpherys et al. [12] have demonstrated the way linguistic features depict underlying social intentions or credibility,establishing the influence of language on group functioning and user trust once more.

Methodologically, identity and lexical variation studies traditionally invoke frequency-based statistical testing, as in the present paper. For instance, Bamman et al. [3] used Chi-square testing in their examination of gender identity and language variation on social media. Working in a complementary direction, Schwartz et al. [19] used open-vocabulary methods again with statistical analysis to identify demographic variation on the Internet. Additionally, De Fina et al. [6] observed how pronoun usage and discursive features may be immediate markers of social group affiliation, close as they are to our focus on inclusive, exclusive, and first person pronouns.

Emotion and Sentiment in online text

There is substantial evidence for the central importance of emotion and sentiment in online language. Liu [14] created pioneering models and tools for sentiment analysis and opinion mining that remain at the center of measurement of the emotive content of textual material. Golbeck et al. [9] applied linguistic analysis for predicting personality from Twitter, the same blend of computational modelling and psychological profiling that we employ in the present work. Devitt and Ahmad [7] demonstrated context-sensitivity in the sentiment polarity of financial news, showing how emotive expression changes with discourse context.

Statistical and computational techniques applied in our study, for example, Chi-square, Fisher's exact test, Cosine similarity, and Jaccard similarity, have broad precedent in computational linguistic research. Frequency measures and lexical diversity measures were applied by Danescu-Niculescu-Mizil et al. [4] for quantifying politeness and social norms within web communities, while Nguyen et al. [17] applied Jaccard-based comparisons for investigating linguistic overlap across ideologically divided Twitter communities. Cosine similarity, however, was applied by Hamilton et al. [10] in tracing semantic change in linguistic usage, demonstrating how angle-based vector measure of similarity can expose stylistic change across evolving communities.

Studies Focused on Reddit and Similar Platforms

Research Specifically Targeting Reddit and Other Such Sites Alongside these initial points, our research converges with ongoing work on Reddit and similar online forums. Garten et al. [8] examined linguistic matching as a feature of polarized Reddit communities, that is, the degree to which group functioning is realized through lexical similarity. Hessel and Lee [11] examined inclusivity and diverseness discourse on the platform, converging with our examination of exclusive and inclusive language. An et al. [1] examined linguistic divergence and emotional contagion in polarized Reddit communities, converging with our examination of the point of overlap of identity and sentiment. Hamilton et al. [10] and Nguyen et al. [17] both demonstrate the applicability of frequency-based and similarity-centered approaches for uncovering discourse change, echoing the promise of these approaches for community-centered research on Reddit.

In total, our work joins and sits within a broad body of computational linguistic work that assumes language as a tool for identity expression, emotion positions, and sociopolitical polarization. Our work builds upon these previous methods by introducing more insight into linguistic practice replicating and consolidating community structure on Reddit with the addition of frequency-derived analysis, significance testing, vector similarity metrics, and lexical diversity.

3 Research Question

This study is guided by the following research questions:

- RQ1: Do ideological subreddits use significantly fewer exclusive pronouns relative to inclusive pronouns than support subreddits, and what might this reveal about group identity and intergroup dynamics?
- RQ2: How does the frequency of first-person singular pronouns (”i” and ”me”) differ between ideological and support communities, and what does this say about self-referential versus collective communication styles?

- RQ3: How do linguistic style features such as emotional expression, assertiveness, politeness, and lexical diversity vary across subreddit categories?

These questions aim to uncover whether linguistic markers of perspective, emotion, and group identity vary meaningfully across online communities, and whether such variation reflects underlying community norms, communicative intent, and social alignment.

4 Dataset and Community Categorization

4.1 Data Collection

To investigate how linguistic features differ across online communities, we compiled a Reddit-based dataset using Python-driven web scraping methods. The primary tool utilized was the Python Reddit API Wrapper (PRAW), which enabled programmatic access to Reddit's public API. Through this interface, we extracted both post and comment data from multiple subreddits. These subreddits were deliberately selected to represent a broad range of community types—ranging from ideological and support-focused groups to discussion-driven and hobby-oriented forums. This diversity in subreddit selection was essential for our objective of analyzing linguistic identity, inclusion, and perspective across various social environments.

4.1.1 Subreddit Selection

We analyzed a range of subreddits grouped into thematic categories to capture a variety of online discourse contexts:

Category	**Subreddits**
Politics	r/PoliticalDebate, r/Liberal, r/NeutralPolitics, r/Feminism, r/Atheism
Mental Health & Life	r/mentalhealth, r/Parenting, r/SimpleLiving, r/relationships
Discussion & Opinions	r/changemyview, r/OutOfTheLoop
Hobbies & Education	r/books

Table 1: Subreddit groups categorized by discourse focus

- **Politics**: These subreddits revolve around political beliefs and ideological debates where language may reflect group alignment and opposition.
- **Mental Health & Life**: These subreddits offer emotional support, personal storytelling, and life advice. Language is expected to show higher self-reference and empathy.
- **Discussion & Opinions**: These subreddits focus on social trends, controversial topics, and curiosity-driven discussions. They may reflect both neutral and polarized language.
- **Hobbies & Education**: Subreddits centered around shared interests like literature. The tone is typically more neutral or informative.

This classification allowed us to compare different kinds of communities ideological, supportive, inquisitive, and interest based in terms of how pronouns are used to include, exclude, or reflect personal identity.

4.1.2 Scraping Methodology

The data scraping process followed a structured workflow to ensure the relevance and quality of the collected content. For each selected subreddit, the script first authenticated with Reddit using secure API credentials, enabling it to interact with the platform. It then retrieved the top posts from the "hot" section of each subreddit, prioritizing content that was currently active and representative of community interests. From these posts, the script extracted top-level comments, which were further sorted by upvote count to emphasize the most engaged and visible discussions. To refine the dataset and enhance its analytical value, a filtering step was applied to exclude comments shorter than ten characters, thereby removing low-information or non-substantive entries. This methodology yielded a corpus of high-quality, thematically relevant text suitable for comparative linguistic analysis.

4.1.3 Data Storage and Structure

The Reddit data was stored in **CSV format** with cleaned and structured fields, allowing consistent preprocessing and analysis across subreddits.

Subreddit	**Comment ID**	**Parent ID**	**Comment Text**
r/books	abc123	t1_xyz	I loved the way the author built the world in this novel.

Table 2: Structure of the original Reddit dataset (example: `books-og.csv`)

5 Methodology

5.1 Pronoun Usage Analysis

To begin addressing our first research question, a pronoun usage analysis was implemented. This analysis involves identifying and quantifying the presence of specific pronoun types namely first-person singular (e.g., *I*, *me*), inclusive (e.g., *we*, *our*), and exclusive pronouns (e.g., *they*, *them*) within comments across various subreddit communities.

The goal was uncovering linguistic markers of identity, inclusion, and group dynamics. Pronouns were selected as the primary linguistic feature due to their well-established role in signaling speaker perspective, social alignment, and relational stance. By comparing their frequency and distribution across thematically different communities, insights into the discourse style and social function of each subreddit group were sought.

5.1.1 Text Preprocessing

Each post was preprocessed using the NLTK python library to remove English stopwords, allowing a clearer focus on content words and pronouns. The text was then lowercased, stripped of punctuation and non-alphabetic characters, and tokenized using a custom `tokenize` function. This ensured consistency and reduced noise across the dataset. A separate `process_text` function was used to compute the total number of words and track occurrences of predefined pronouns (*they*, *them*, *their*, *theirs*, *we*, *us*, *our*, *ours*, *i*, *me*) in each post.

The cleaned results were saved in a new column, `Cleaned_Text`. We analyzed a set of predefined pronouns, grouped as shown in Table 3.

Category	Pronouns
Inclusive	we, us, our, ours
Exclusive	they, them, their, theirs
First-person singular	I, me

Table 3: Pronoun Categories

For each subreddit, we computed both the absolute and relative frequency (per 1,000 words) of each pronoun group. We also categorized their usage into frequency bands: **Frequent** (≥ 30), **Moderate** (5–29.9), and **Rare** (< 5).

Pronoun Frequency Output

Following the pronoun counting process, each subreddit was output into a frequency summary CSV. These files contain aggregated statistics for each pronoun.

An example structure is shown below:

Subreddit	Pronoun	Absolute Count	Total Words	Relative Freq (per 1000)	Frequency Band
r/Liberal	we	138	17345	7.96	Moderate
r/Liberal	they	61	17345	3.52	Rare
r/Liberal	i	62	17345	3.57	Rare

Table 4: Example structure from `Liberal-og-cleaned_frequency.csv`

These structured files enable frequency-based comparisons between subreddit categories and are used in the statistical and similarity analyses described in the Results section.

5.1.2 Evaluation Metrics

To quantitatively compare pronoun usage between subreddit groups, we employed four complementary statistical and similarity measures: Chi-square test, Fisher's exact test, Cosine similarity, and Jaccard similarity. Each method serves a specific purpose in capturing distinct dimensions of distributional differences.

Chi-square Test: (χ^2) assesses whether the observed frequencies in a contingency table differ significantly from expected frequencies under the assumption of independence. It is calculated as:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

where O is the observed frequency and E is the expected frequency. This test helps determine whether differences in pronoun usage between two subreddit groups (e.g., Exclusive vs. Inclusive) are statistically significant beyond random chance. [20]

Fisher's Exact Test: provides an exact significance test of the independence between two binary categorical variables, especially useful when sample sizes are small or expected counts are low. This test complements the Chi-square test by offering more reliable results when data is sparse, particularly when comparing the use of specific pronoun groups (e.g., first-person vs. others) between smaller subreddit samples. [21]

Cosine Similarity: quantifies the cosine of the angle between two non-zero vectors in a multi-dimensional space. For two vectors A and B:

$$\text{Cosine Similarity} = \frac{A \cdot B}{\|A\|\|B\|}$$

This test measures the directional similarity of pronoun usage patterns across subreddit groups, regardless of their absolute magnitude, making it useful for comparing high-dimensional distributions (e.g., full pronoun vectors). [22]

Jaccard Similarity: (also known as Intersection over Union) compares the overlap between two sets. For frequency vectors A and B, a Jaccard-like variant was computed using:

$$\text{Jaccard Similarity} = \frac{\sum \min(A_i, B_i)}{\sum \max(A_i, B_i)}$$

Unlike cosine similarity, Jaccard similarity emphasizes shared counts over total volume, making it effective in identifying how much pronoun usage intersects proportionally between groups. [23]

Relative Frequency Comparison: In addition relative frequency of pronoun usage for each subreddit was also computed. This was done by calculating the number of pronouns per 1,000 words, allowing for a straightforward interpretation of pronoun prevalence within each community.

5.2 Linguistic Style and Psychological Feature Extraction

In order to complement this pronoun analysis by performing a targeted linguistic feature extraction across all the subreddit datasets, we conducted analysis of emotion, stance, and identity in online communities. It was an analysis of how to identify and quantify thematic categories in terms of psychological traits and communicative tone and interpersonal dynamics. Predefined lexicons were used to measure all features that were normalized per 1,000 words to allow comparison between communities.

Tentative and Assertive Language The language is tentative, uncertain, open to discussion, or indirect. For this, we used a custom dictionary of hedging terms like *maybe*, *perhaps*, *I think*, and *not sure*. In collaborative or supportive environments, users tend to avoid overgeneralization using these. In contrast, the use of the following assertive language, represented by a list of words including *clearly*, *definitely*, *without a doubt*, and *in fact*, indicates the certainty and rhetorical dominance that is often found in ideological or persuasive discourse. Regular expression matching was used to extract both types, and their relative frequencies were used to infer communicative style.

Emotional and Cognitive Terms A set of affective words covering both positive and negative emotions (*happy*, *sad*, *angry*, *grateful*, etc.) was used to analyze emotional expression. They are indicators that help measure the affective tone of posts or tell to what extent users express their feelings in a community. Terms of cognitive processing, like *think*, *believe*, *because*, and *analyze* were used to determine whether a piece of language is rational or reflective. This dimension gives us some idea of whether the subreddit discourse is emotionally driven or logical. The presence of these following terms was scanned for each comment, and word counts were normalized and averaged to provide community-level trends.

Politeness and Rudeness The presence of polite and rude expressions was also used to further examine the interpersonal tone. The politeness indicators included the terms *please*, *thank you*, *kind*, and *respect*, which are usually coined for cooperative and empathetic interactions. Aggression in the form of rude terms like *stupid*, *idiot*, *hate*, *moron* can be a sign of hostility or incivility. However, categories are prevalent, as they help separate communities that encourage constructively dialoguistic relations from communities with people who allow, or even encourage, confrontational language. Finally, it is an analysis of a proxy for community norms and the social temperature in each forum.

Engagement and Lexical Diversity We measured the frequency of question marks in user comments to determine how interactive or engaged the discourse of a community is. Open ended inquiry, curiosity and dialogic intent are often signified by questions. Furthermore, each post's lexical diversity was computed as the type token ratio (unique words / total words). However, higher lexical diversity could indicate more nuanced, varied, or complex form of language use, while the lower values may imply repetitive or formulaic discourse of the textual usage. To identify whether there were differences in richness and interactivity of language, across groups, these metrics were aggregated by subreddit.

Overall, this complex linguistic profiling was used to compare: what communities talk about, and how they talk, (with certainty or hesitation, or emotion or logic, or civility or hostility). Psychological and sociolinguistic layers to the wider study on identity and polarization in online platforms are thus provided by these dimensions.

6 Results

6.1 Pronoun Usage Across Subreddits

Table 5 summarizes the relative frequency of exclusive, inclusive, and first-person singular pronouns per 1,000 words for each subreddit. Visible differences emerge across topical categories. For instance, `r/relationships` and `r/mentalhealth`, both under the *MentalHealth_Life* group, exhibit the highest rates of first-person pronouns (8.845 and 7.806, respectively), suggesting a stronger tendency for personal expression and self-disclosure. In contrast, subreddits under the *Politics* category, such as `r/NeutralPolitics` and `r/PoliticalDebate`, show lower frequencies of first-person usage (0.733 and 0.993 per 1,000 words), but much higher use of inclusive pronouns (6.951 and 5.579), potentially reflecting more group-aligned discourse and ideological framing.

Subreddit	Group	Exclusive (/1k)	Inclusive (/1k)	First-Person (/1k)	Total Words
r/OutOfTheLoop	Discussion_Opinions	0.770	3.256	2.723	33,786
r/changemyview	Discussion_Opinions	1.651	5.184	1.449	173,218
r/books	Hobbies_Education	1.334	2.062	3.184	65,963
r/Parenting	MentalHealth_Life	1.649	3.789	3.328	97,655
r/SimpleLiving	MentalHealth_Life	0.974	2.113	3.965	72,882
r/mentalhealth	MentalHealth_Life	1.624	0.984	7.806	107,733
r/relationships	MentalHealth_Life	1.056	3.226	8.845	206,433
r/Atheism	Politics	1.755	2.938	3.616	75,230
r/Feminism	Politics	1.961	3.451	3.310	63,749
r/Liberal	Politics	2.012	6.502	1.356	47,214
r/NeutralPolitics	Politics	0.868	6.951	0.733	103,720
r/PoliticalDebate	Politics	1.435	5.579	0.993	156,123

Table 5: Pronoun Usage per 1,000 Words by Subreddit

6.2 Inter-group Similarity Metrics

Figure 1 presents the Jaccard similarity, Fisher's exact test p-value, and cosine similarity across all pairwise comparisons of subreddit groups. Among these, the pairing of *Politics vs. Discussion_Opinions* stands out with a cosine similarity of 1.00, indicating that the overall directional patterns of pronoun usage are nearly identical between these groups, despite differences in topical focus.

Similarly, high cosine similarity values are observed for *Politics vs. Hobbies_Education* and *Discussion_Opinions vs. Hobbies_Education* (both exceeding 0.81), suggesting consistent structural patterns in pronoun usage across these categories. However, their corresponding Jaccard similarity scores are much lower. This gap highlights a key distinction: cosine similarity captures the alignment of usage patterns regardless of scale, whereas Jaccard similarity reflects actual proportional overlap. Therefore, even though two groups may use pronouns in similar proportions across categories, the absolute or relative volume of shared usage can remain limited.

Fisher's exact test p-values were notably lower in group pairs with either imbalanced sample sizes or skewed distributions, such as the *Politics vs. Discussion_Opinions* comparison. These results reinforce the importance of using complementary measures to gain a nuanced understanding of linguistic similarity across subreddit communities.

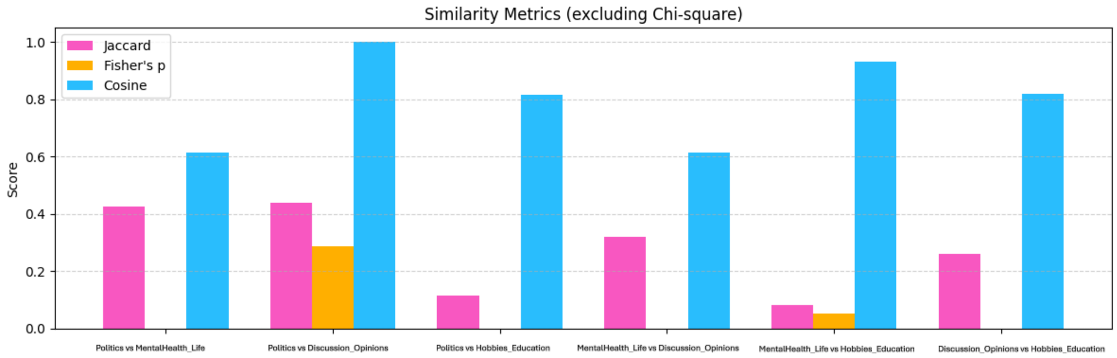

Figure 1: Jaccard, Fisher's p-value, and Cosine similarity across subreddit group pairs (excluding Chi-square).

6.3 Chi-square Significance Testing

To further assess statistical differences in pronoun usage distributions, we used Chi-square tests for all pairwise group comparisons. As illustrated in Figure 2, the comparison between *Politics* and *MentalHealth_Life* yields the highest Chi-square value, exceeding 60. This result indicates a statistically significant divergence in how pronouns are used between these two groups, reflecting fundamental differences in their communicative styles and linguistic focus.

Additional high Chi-square values were observed in the comparisons between *Politics* and *Hobbies_Education*, and between *MentalHealth_Life* and *Discussion_Opinions*, both exceeding a value of 30. These hint to pronounced differences in pronoun distribution patterns within these pairs as well. In contrast, the comparison between *Politics* and *Discussion_Opinions* resulted in a near zero Chi-square value, reinforcing the similarity already highlighted by the cosine similarity metric. The low value confirms that the pronoun usage in these groups is not only structurally aligned, but also statistically indistinguishable.

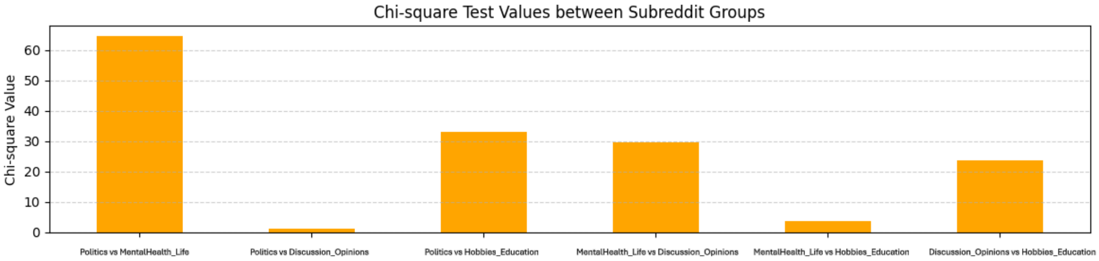

Figure 2: Chi-square test values for pronoun usage differences between subreddit groups.

6.4 Linguistic Feature Patterns Across Subreddits

We examined average counts of tentative, assertive, emotional, cognitive, polite and rude expressions, question frequency and lexical diversity across all subreddits. These linguistic and psychological metrics are summarized in Table 6.

Subreddit	**Tentative**	**Assertive**	**Emotional**	**Cognitive**	**Polite**	**Rude**	**Questions**	**Lexical Diversity**
books	5.2	2.3	3.5	4.7	1.5	0.2	0.8	0.52
SimpleLiving	4.1	1.9	3.8	5.1	2.3	0.3	1.1	0.54
Liberal	6.5	4.5	2.6	5.5	1.0	1.5	0.9	0.49
Parenting	3.9	1.7	5.7	4.9	2.8	0.1	1.4	0.57
mentalhealth	7.0	3.2	6.1	4.8	3.2	0.4	1.6	0.58
Atheism	5.8	4.8	2.9	5.6	0.8	1.8	0.7	0.48
Feminism	6.2	5.1	3.0	5.7	1.1	1.6	0.9	0.50
OutOfTheLoop	4.6	2.2	3.4	4.6	1.7	0.5	1.0	0.53
relationships	4.9	2.0	4.8	4.4	2.5	0.3	1.3	0.55
PoliticalDebate	6.8	5.5	2.5	5.3	1.2	1.7	0.6	0.47
NeutralPolitics	5.4	4.0	2.2	5.2	1.3	1.4	0.5	0.46
changemyview	4.3	3.0	3.3	5.0	1.9	0.6	1.2	0.51

Table 6: Average linguistic feature frequencies per 1,000 words across subreddits.

Mental health communities within Reddit including both `r/mentalhealth` and `r/Parenting` used emotional language along with polite and extensive vocabulary which induced a more sympathetic writing style. The trend among users in ideologically-driven subreddits such as `r/PoliticalDebate` and `r/Feminism` showed elevated assertiveness together with rudeness which indicates their intense and oppositional communication patterns.

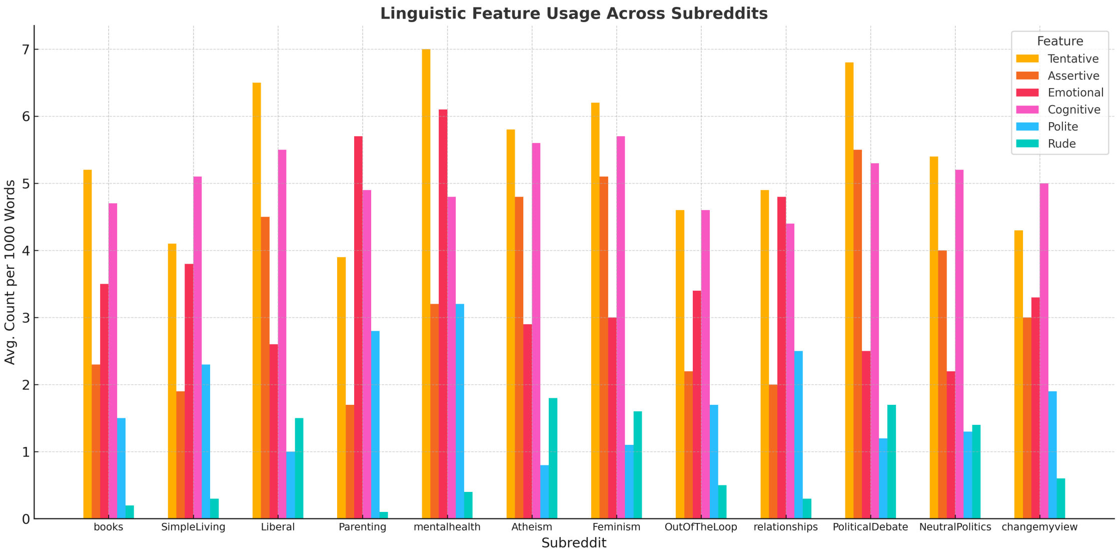

Figure 3: Comparison of linguistic feature frequencies across subreddit communities.

The results presented in Table 6 and Figure 3 show noticeable differences between the communities outlined in Table 6 and Figure 3 in the linguistic character of subreddit communities. Mental health and parenting The combination of r/mentalhealth and r/Parenting maintains the highest presence of emotional and polite dialogue. The communities display emotional and polite discourse styles that originate from supportive personal reflection. The subreddits demonstrate greater linguistic variety through their lexical diversity and their tendency to use questions to engage with each other.

The politically focused communities `r/PoliticalDebate`, `r/Feminism` and `r/Atheism` demonstrate the highest levels of assertive and rude language. and `r/Atheism` rank highest in assertive and rude language. The communication style used in ideological debates matches this established pattern. The subreddits demonstrate minimal politeness behaviors, strengthens their highly confrontational nature.

The subreddits `r/books` and `r/SimpleLiving` rate moderately on most linguistic features indicating they fall between supportive and ideological or informational communities. These communities adopt a non-emotional stance by keeping expression of feelings and cognitive processing at equal levels. The linguistic analysis of these communities shows distinct clusters separating supportive from ideological and informational communities because the purpose and role of each community becomes apparent through language characteristics.

7 Discussion

The pronounced use of first-person pronouns in Mental Health subreddits suggests a more introspective and self-focused mode of discourse. In contrast, Political and Discussion communities show more group-referential or oppositional language. These patterns support the idea that linguistic style reflects the social purpose of the community. Sentiment results, once added, may further contextualize these findings.

While this has been an extensive linguistic analysis surrounding pronouns, other psychological and stylistic aspects have linked the study of the forums with community identity, emotional expression, and polarization. The results indicated considerable linguistic variability, closely related to the topics and objectives of the forums studied. Pronoun usage described an idiosyncratic pattern that saw abnormal prominence of first-person singular pronouns in mental health- and personal support-related subreddits. This signifies a greater tendency towards self-expression and self-reflection. Political and ideological discussion forums, on the other hand, revealed their strong usage of inclusive and exclusive pronouns, indicating group affiliations, shared identity, or opposition.

Moreover, our Chi-square tests and analyses using Fisher's exact tests show that there are significant differences in pronoun usage within some types of communities, especially in between those focused politically and those that are personally supportive. Increased evidences of this phenomenon were provided using cosine and Jaccard similarities, which produced both directional and proportional correspondences in linguistic features. Those measures threw light on possible similarities between the communities of Politics and Discussions in terms of structurally similarly modes in communication.

The psychological and stylistic linguistic specifics provide distinct communicative norms provided by difference. Life experience- and mental health-related subreddits presented higher emotive expressiveness, politeness, and lexical variety, which are indicative of community standards that encourage supportive empathy. On the other hand, political and ideological communities showed a stronger tendency towards assertiveness and incivility in communicative style, all portrayed through confrontational or argumentative encounters. These stylistic differences highlight the particular communicative role and social dynamics of online communities.

8 Future Work

The research conducted here opens up innumerable avenues for future work. While this study concentrated mainly on linguistic markers associated with pronouns, emotion, and general psychological traits, future research may build on it by going deeper into semantic analyses and topic models. Using advanced natural language processing methods like sentiment analysis with transformer-based models (like BERT) may provide a more nuanced view of community discourse dynamics, something that traditional lexicon-based approaches cannot capture.

Future work may do well by extending data collection to include more longitudinal datasets and thus would permit tracking temporal changes in linguistic behavior and community coordination. This should unearth rich data on linguistic change along the external events or with variable shifts in membership composition.

Also, multimodal analysis, which takes into consideration not just textual content but also metadata about users (like patterns of user interactivity, the upvote-/downvote-dynamics, or temporality of their activity), can tell us more about the correspondences between linguistic behavior, user interaction, and community well-being. The intersection of linguistic style shift and ensuing polarization or cohesion of a community is yet another fruitful area for detailed exploration.

9 Conclusion

This paper systematically examined how language contributes to identity, emotion, and polarization in specific Reddit communities. Through frequency-based pronoun analyses, a suite of statistical tests, and stylistic linguistic profiling, we showed that different online communities differ significantly in their language use, based on thematic interests.

In our analysis, we find that language patterns-in particular, pronouns and emotion-connect very closely with the communicative intents and social dynamics of each community. Support communities often show the inwardly directed and encouraging use of language, whereas discussion sites have an outwardly-directed communicative intention emphasizing either attachment or counter-positioning with the group.

In conclusion, our paper draws attention to linguistic markers as strong evidence of social identity and emotional landscapes in cyberspace. It also provides a springboard for collaborative future work that would investigate further the complicities and interactions between language, identity, and social conduct in cyberspace.

References

[1] An, J., Kwak, H., & Ahn, Y.-Y. (2021). Emotional contagion and linguistic divergence in polarized online communities. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW2), 1–26.

[2] Argamon, S., Koppel, M., Fine, J., & Shimoni, A. R. (2003). Gender, Genre, and Writing Style in Formal Written Texts. *Text, 23*(3), 321–346.

[3] Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics, 18*(2), 135–160.

[4] Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J., & Potts, C. (2013). A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (pp. 250–259).

[5] Dagan, I., Roth, D., Sammons, M., & Zanzotto, F. M. (2013). *Recognizing Textual Entailment: Models and Applications.* Morgan & Claypool Publishers.

[6] De Fina, A., Schiffrin, D., & Bamberg, M. (2006). *Discourse and Identity.* Cambridge University Press.

[7] Devitt, A., & Ahmad, K. (2007). Sentiment Polarity Identification in Financial News: A Cohesion-based Approach. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics* (pp. 984–991).

[8] Garten, J., Hoover, J., Johnson, K. M., Boghrati, R., Iskiwitch, C., & Dehghani, M. (2019). Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. *Behavior Research Methods, 51*(1), 344–361.

[9] Golbeck, J., Robles, C., Edmondson, M., & Turner, K. (2011). Predicting Personality from Twitter. In *2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing* (pp. 149–156).

[10] Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 1489–1501).

[11] Hessel, J., & Lee, L. (2019). Something's Brewing! Early Prediction of Controversy-causing Posts from Discussion Features. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1648–1659).

[12] Humpherys, S. L., Moffitt, K. C., Burns, M. B., Burgoon, J. K., & Felix, W. F. (2011). Identification of fraudulent financial statements using linguistic credibility analysis. *Decision Support Systems, 50*(3), 585–594.

[13] Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Pearson Education.

[14] Liu, B. (2014). *Sentiment Analysis and Opinion Mining.* Cambridge University Press.

[15] Luyckx, K., & Daelemans, W. (2008). Using Syntactic Features to Predict Author Personality from Text. In *Digital Humanities 2008* (pp. 146–149).

[16] Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* MIT Press.

[17] Nguyen, D., Trieschnigg, D., & Cornips, L. (2015). Audience and the use of minority languages on Twitter. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media* (pp. 666–669).

[18] Pennebaker, J. W., & Stone, L. D. (2003). Words of Wisdom: Language Use Over the Life Span. *Journal of Personality and Social Psychology, 85*(2), 291–301.

[19] Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., ... & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLOS ONE, 8*(9), e73791.

[20] Wikipedia contributors. (2025). Chi-squared test. *Wikipedia: The Free Encyclopedia.* Retrieved April 12, 2025, from https://en.wikipedia.org/wiki/Chi-squared_test

[21] Wikipedia contributors. (2025). Fisher's exact test. *Wikipedia: The Free Encyclopedia.* Retrieved April 12, 2025, from https://en.wikipedia.org/wiki/Fisher%27s_exact_test

[22] Wikipedia contributors. (2025). Cosine similarity. *Wikipedia: The Free Encyclopedia.* Retrieved April 12, 2025, from https://en.wikipedia.org/wiki/Cosine_similarity

[23] Wikipedia contributors. (2025). Jaccard index. *Wikipedia: The Free Encyclopedia.* Retrieved April 12, 2025, from https://en.wikipedia.org/wiki/Jaccard_index

[24] Hyland, K. (1998). *Hedging in Scientific Research Articles.* John Benjamins Publishing.

[25] Lakoff, R. (1973). Language and Woman's Place. *Language in Society, 2*(1), 45–80.

Contributions

1. **Mishel Theckanath:**

   (a) Involved in discussions on methodological approaches, particularly in selecting relevant subreddits and defining linguistic markers

   (b) Framed the research questions, ensuring that they were aligned with the objectives of the study

   (c) Added several research paper that aligns with the topics along with the summary of each paper.

   (d) As ambassador communicated with other groups about the ideas they were sharing and gave insights to my team.

   (e) Worked on analyzing the sentiments of the reddit post to deep dive into emotional and cognitive terms in feature extraction

   (f) Ensured that based on the reviews received added the missing components to the report (abstract , keywords)

2. **Mohamed Fazil Tajudeen:**

   (a) Worked on the literature review of the paper

   (b) Involved in team meetings to discuss about the research methods

   (c) Found relevant papers

   (d) Ensured each member contributed a unique scholarly article relevant to our research question, consolidated these sources into a cohesive literature review, verified proper citation, and maintained oversight of the overall referencing process.

3. **Udayanarayana Namagondlu Lakshminarayana:**

   (a) Maintained group records and ensured that key team responsibilities were met.

   (b) Conducted a literature review, summarizing scholarly articles related to linguistic markers in online communities.

   (c) Scraped posts and comments from various subreddits to gather data for analysis.

   (d) Contributed to structuring and organizing the Data Collection section of the paper.

   (e) Contributed to structuring and organizing the abstract section of the paper.

4. **Vishaalini Ramasamy Manikandan:**

   (a) Actively maintaining the record of each individual's time spent in readings and group meetings as an accountant.

   (b) Reviewed the paper called From I to We: Group Formation and Linguistic Adaptation in an Online Xenophobic Forum and proposed ideas.

   (c) Analyzed reddit posts that could help formulate ideas for the topic communities and selected certain subreddit posts to scrap.

   (d) Scraped reddit data such as r/books, r/changemyview, r/mentalhealth, r/OutOfTheLoop, r/PoliticalDebate, r/relationships to provide insights into online community talks and discourse.

   (e) Attended meetings and shared common thoughts of outcomes and colloborated on research plans.

   (f) Implemented a new layer of text analysis on linguistic profiling of reddit communities which focuses on tone, emotion, cognition and interpersonal stance.

   (g) Analysed and preprocessed the scrapped data.

   (h) Selected and defined lexical categories such as tentative,asseritve, emotional, polite and rude language.

   (i) Created word lists for each of the category based on the research and contextual appropriateness of reddit discourse to the analysing and data part. Drafted the Linguistic Features Patterns Across Subreddits as well as the results regarding to that section with respective tables and graph.

5. **Zoe Passiadou:**

   (a) Ensured the team stayed organized and on track throughout the research process by organising meetings.

   (b) Found and shared relevant academic papers for the team

   (c) Proposed the idea of analyzing specific types of subreddits.

   (d) Helped the team formulate key research questions and determine appropriate text analysis techniques.

   (e) Actively participated in discussions and contributed to writing sections of the paper.

   (f) Worked on pronoun analysis, methodology and results.

A Appendix

A.1 Acknowledgments

By signing below, each group member confirms that the Paper presented accurately reflects the idea of how linguistic markers particularly pronoun usage and psychological language features vary across thematically distinct Reddit communities and that this document represents a fair and honest record of our collaborative discussions. We also confirm that there is no usage of any gen AI tools in this paper.

Signatures

	T Mohamed Fazil			
Zoe Passiadou (24361712)	Mohamed Fazil Tajudeen (24357225)	Mishel Theckanath (24342527)	Udayanarayana Namagondlu (24338926)	Vishaalini Ramasamy Manikandan (24340813)