Language Styles Investigating and Analysing between Subreddits

Yang Yinfang

University College Dublin

Research Question

Social media platforms like Reddit provide a rich source of data for investigating language style and language use can vary significantly across subreddits and can be influenced by factors such as the purpose of the community, the demographics of the users, and the cultural and social context in which the language is being used (Ferrer et al., 2021). This research seeks to answer the question of whether it is possible to infer topics and distinguish between subreddits based on language style. In other words, whether each subreddit has a unique language style. Specifically, lexical diversity, frequency distribution, and lexical similarity and difference between different subreddits will be investigated and discussed here.

Methods

To gather data for my study, I utilized a crawler to scrape raw text data from various subreddits. I selected eight subreddits that would provide a representative analysis of language

style. To ensure a comprehensive sample, subreddits with both greater and lesser differences were chosen.To be more specific, the 8 subreddits were grouped into four topic categories, with each group containing two similar subreddits. These subreddits were:

Art and Design (r/art, r/design),

Gaming and PS4 (r/gaming, r/PS4),

Ireland and England (r/ireland, r/england),

Science and Technology (r/science, r/technology).

Through this selection process, I aim to identify the unique linguistic features that distinguish each subreddit, as well as to explore the similarities and differences between subreddits within each topic category.

To prepare the data for analysis, several pre-processing steps were carried out, including tokenization, converting all text to lower case, removing punctuation and stop words, and stemming.

After this, various approaches were employed to analyze the data. Firstly, the lexical diversity of the different subreddits was examined. Then, the most commonly used words in each subreddit were identified, their frequency distributions were calculated, and graphs were created.

For more advanced analysis, two tagging steps were performed: part of speech tagging and named entity recognition. Although both approaches were performed, I do not conduct a dedicated analysis based on NER result in this study, POS tagging serves as a key component of my analysis. Part of speech tagging was used to identify the nouns, verbs, and adjectives used in each subreddit and compare their differences. Part of speech tagging can provide more detailed and accurate information than just analyzing the frequency distributions of the most common tokens.

The next step is to find the most representative tag in Part-of-Speech (POS) tagging for indicating language style. Two important POS tags, JJ and NN, were chosen for this study because they provide valuable information about the content and structure of the text. Adjectives(JJ) provide

information about the qualities and characteristics of the nouns they modify, while nouns(NN) represent the objects, people, places, and ideas mentioned in the text. Analyzing the distribution and frequency of adjectives can reveal important patterns and trends related to the sentiment, tone, and style of the text. Similarly, analyzing the frequency and distribution of nouns can provide insights into the main themes, topics, and entities discussed in the text.Therefore, analyzing the frequency and distribution of adjectives and nouns can help identify the topics and themes discussed in the text, as well as the language style used. To obtain clearer results, the analysis was expanded to include the 20 most frequent words.

By applying these methods, it aim to gain a deeper understanding of the linguistic characteristics of each subreddit, and ultimately, to answer the research question of whether it is possible to distinguish between subreddits based on their language style.

Results

The lexical diversity result

	1	2	3	4	5	6	7	8
subreddits	**art**	**design**	**gaming**	**PS4**	**ireland**	**england**	**science**	**technology**
Lexical diversity Original	0.15	0.19	0.17	0.14	0.13	0.27	0.17	0.14
Lexical diversity Pre-Processed	0.22	0.27	0.26	0.19	0.19	0.42	0.25	0.19

Table 1 The lexical diversity result of 8 subreddits

The lexical diversity scores suggest that the subreddit england has the highest lexical diversity, with a pre-processed of 0.42 and an original of 0.27. The subreddit ireland has the lowest lexical diversity, with a pre-processed of 0.19 and an original of 0.13. Despite both subreddits being

related to countries, they display markedly different lexical diversity levels. More similar subreddits did not show more similar lexical diversity.

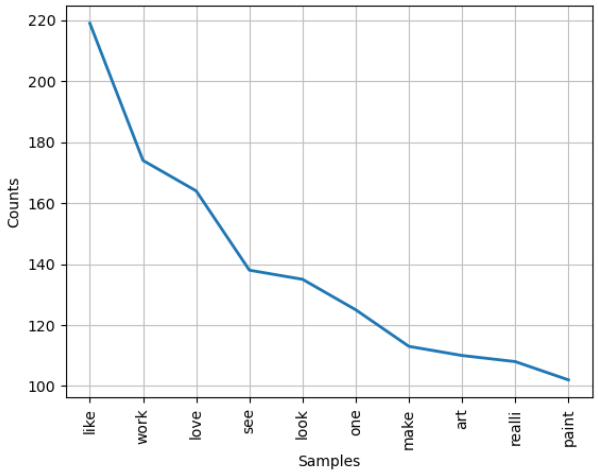
The frequency distributions result

The most common tokens

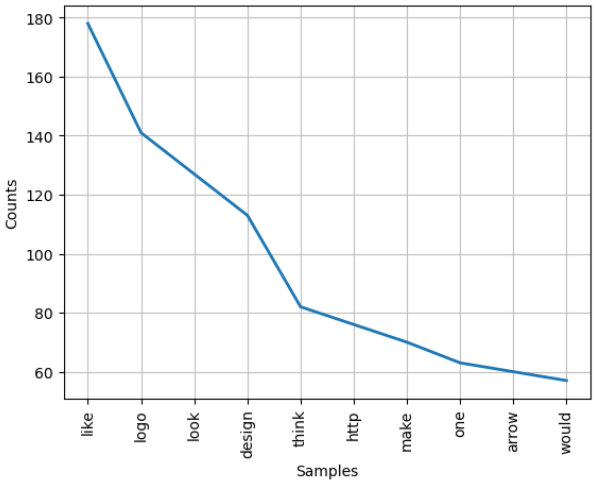
	1	2	3	4	5	6	7	8
subreddits	**art**	**design**	**gaming**	**PS4**	**ireland**	**england**	**science**	**technology**
most common tokens 1	like	like	game	game	american	look	remov	peopl
most common tokens 2	work	logo	like	like	get	good	student	neutral
most common tokens 3	love	look	play	look	us	england	peopl	net
most common tokens 4	see	design	one	play	fuck	like	get	like
most common tokens 5	look	think	get	get	go	great	like	internet
most common tokens 6	one	http	time	one	ireland	come	tax	http
most common tokens 7	make	make	guy	time	like	fan	would	get
most common tokens 8	art	one	go	good	rent	see	http	fuck
most common tokens 9	realli	arrow	peopl	ps4	year	make	one	comcast
most common tokens 10	paint	would	year	onlin	peopl	realli	time	would

Table 2 The frequency distributions result - The most common tokens of 8 subreddits

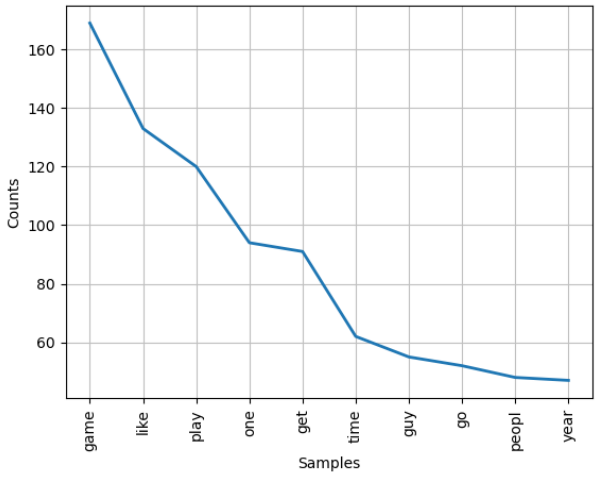
By comparing the frequency distributions and most common tokens, we can gain insights into the language styles and similarities/differences among the various subreddits. Most of the subreddit topic words appear directly in the most common tokens(yellow). There are some common tokens that appear under several subreddits that actually have less meaning(blue), such as like, one, get. They are similar to stop words which don't add much value to the text when analyzing its meaning.



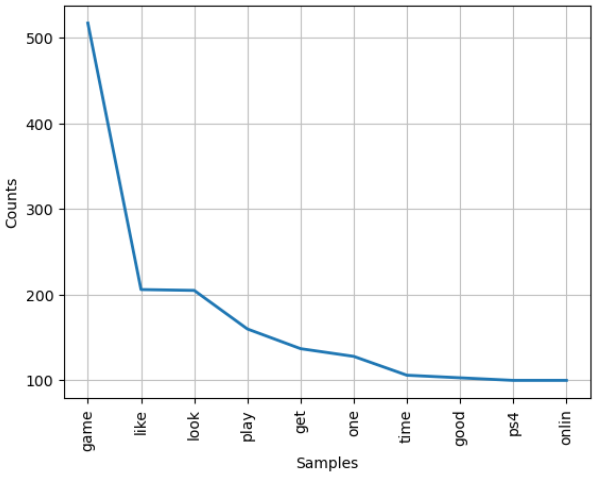
Art frequency distribution



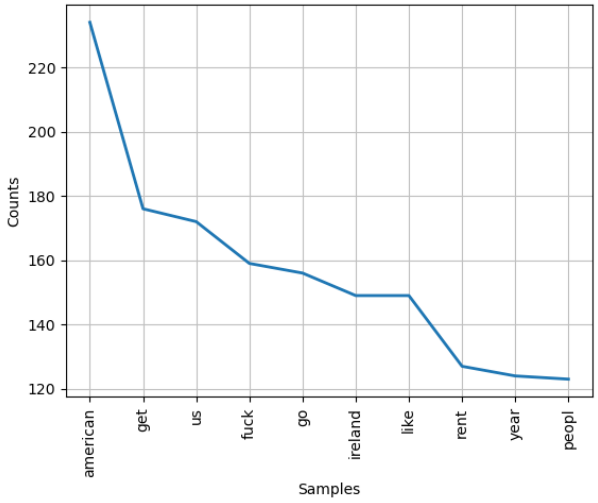
Design frequency distribution



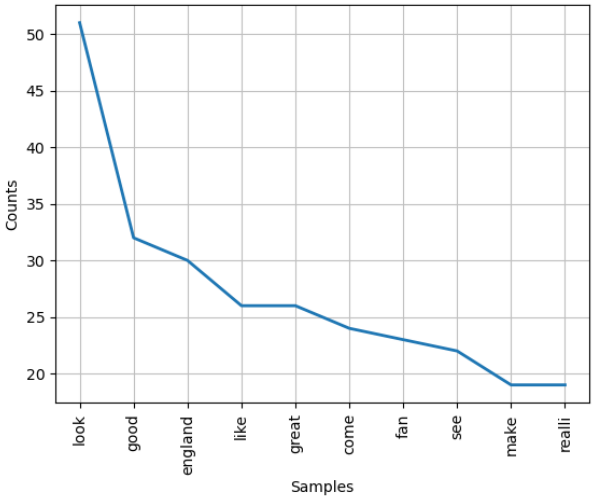
Gaming frequency distribution



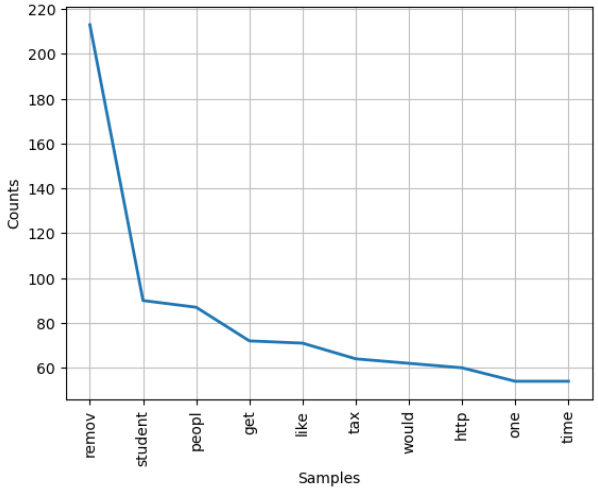
PS4 frequency distribution



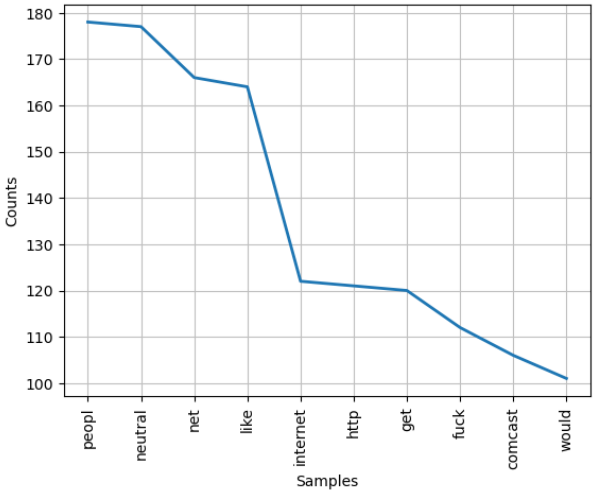
Ireland frequency distribution



England frequency distribution



Science frequency distribution



Technology frequency distribution

Table 3 frequency distribution plots

The plots showed that the most common tokens had varying counts, ranging from 50 to 500. Additionally, we observed that the distribution curves for the 10 common tokens differed across subreddits. Some subreddits showed a more even and average distribution of token frequencies, while others exhibited a more skewed distribution.

Part-of-Speech (POS) tagging result

The most common JJ(adjective) and NN(singular noun)

The subreddits art, design, and Ireland have been chosen here as examples for analysis, with Art and Design being more similar topics. Comparisons and results are analyzed between these three.

Subreddits	Art	Design	Ireland
Most common 20 JJ	amazing, good, great, cool, beautiful, awesome, ', much, incredible, original, many, hard, same, new, sure, other, old, first, perfect, creative.	red, original, great, good, white, nice, right, much, whole, other, *, first, current, new, political, little,], beautiful, bad, logo.	American, ', other, good, Irish, same, much, last, s, few, great, first, bad, new, sad, old, next, many, safe, English
Most common 20 NN	work, art, moon, picture, time, piece, something, job, https, man, way, life, van, t, image, post, thing, painting, paper, martini	logo, design, arrow, https, [, http, color, time, plastic, space, cover,], work, idea, way, something, box, campaign, thing, %	deposit, country, house, t, rent, mortgage, time, year, home, day, money, month, world, dog, shit, https, place, something, landlord, thing

Table 4: The most common JJ(adjective) and NN(singular noun) of art, design, ireland

Discussion

The lexical diversity discussion

According to the results presented in Table 1, the difference in lexical diversity between the pre-processed and original texts suggests that pre-processing the text by removing stop words and other noise can significantly increase the lexical diversity. However, lexical diversity is just one aspect of language style, and it should be analyzed in conjunction with other language features (McCarthy, 2005). Similar topics did not show similar diversity, and we could not infer, by lexical diversity, what the topic was; science, for example, did not show a higher diversity than design. Therefore, we can further analyze the most common tokens in each subreddit to gain a more comprehensive understanding of their language styles.

The frequency distributions discussion

Based on the most common tokens, as shown in Table 2, we can get a general idea of the topics and language styles of each subreddit and make inferences. For instance, the art subreddit concentrates on art and painting, while the design subreddit focuses on logo design. The gaming and PS4 subreddits are related to video games and have very similar most common tokens. The Ireland subreddit seems to emphasize topics like American influence, rent, and people in Ireland, whereas the England subreddit has a broader mixture of topics. The science subreddit concentrates on topics such as student life, taxes, and people in science, whereas the technology subreddit concentrates on the internet, net neutrality, and internet service providers like Comcast.

I started out by choosing two topics that are similar to each other, and here we can compare them. For instance, we can compare the language styles of r/ireland and r/england since they are both related to countries in the British Isles. We can observe that both subreddits frequently use words such as "like" and "get," but r/ireland also has a high frequency of words such as "fuck" and "rent," which may suggest greater frustration or political discussion compared to r/england. On the other hand, r/art and r/design share some similarities in their use of words like "like" and "one," but r/design has a higher frequency of words related to design such as "logo" and "arrow." This

indicates that the two subreddits have similar language styles in some respects, but also have distinct differences due to their subject matter.

Part-of-Speech (POS) tagging discussion

These JJ words suggest that the language style used in the art subreddit is positive and appreciative of art. These NN words suggest that the subreddit is focused on sharing and discussing art, with a particular emphasis on visual arts such as painting and photography.

From the most common JJ and NN words in the subreddit r/design, we can see that the top JJ words are related to color (red, white) and evaluative words (great, good, nice), while the top NN words are related to design elements (logo, arrow, plastic, space) and digital content (https, http). This indicates that the discussions in this subreddit are focused on the design of logos, graphics, and digital content, with a particular emphasis on the use of color.

When compared with the results from r/art, we can see that there is some overlap in the top JJ words, such as "great", "good", and "original", but r/design also has more specific words related to design elements, such as "logo" and "arrow". The top NN words in r/design are also more focused on design elements and digital content, whereas r/art has a broader range of topics related to different art forms (e.g. painting, drawing, sculpture) and materials (e.g. paper, canvas, clay).

Comparing the results with the previous subreddits analyzed, we can see that the most common adjectives in the "ireland" subreddit are more focused on national and cultural identity, such as "Irish" and "English", whereas in the "design" subreddit, more emphasis was placed on visual aesthetics, such as "red", "original", and "white". Similarly, the most common nouns in the "ireland" subreddit are related to housing and finances, whereas in the "art" subreddit, the most common nouns were related to artistic media and techniques.

In a similar study conducted by Horne et al. (2017), the authors investigated the social signals that drive online discussions in various Reddit communities. They found that different subreddits have distinct social dynamics and cultural norms that influence the content and style of communication.

In conclusion, by examining lexical diversity, frequency distributions, and part-of-speech tagging, we were able to identify unique linguistic features of each subreddit. The comparison between subreddits with similar topics and those with different focuses allowed us to further understand the nuances of language style and its relationship to subject matter. While our analysis has some limitations, it provides a foundation for further research into more sophisticated linguistic features and analysis techniques.

Limitation and future work

The several findings reveal each subreddit has its unique linguistic features. One potential limitation of our analysis is that we do not delve deeper into named entity recognition (NER) beyond identifying and classifying basic entities such as people, places, and organizations. Further NER analysis could reveal more detailed information about the relationships and roles of these entities within the text. Additionally, while our POS tagging analysis provides valuable insights, further examination of verb and adverb usage could provide a more nuanced understanding of the language and syntax in the text. These areas offer promising avenues for future work in improving the accuracy and depth of text analysis.

Word Count: 2137

Bibliography

Ferrer, X., Ross, B., Danescu-Niculescu-Mizil, C., & Kleinberg, J. (2021). Discovering and categorising language biases in reddit. Proceedings of the International AAAI Conference on Web and Social Media, 15.

Horne, B. D., Adalı, S., & Sikdar, S. (2017). Identifying the social signals that drive online discussions: A case study of Reddit communities. Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, 1170-1183.

McCarthy, P. M. (2005). An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD) (Doctoral dissertation, The University of Memphis).