Examination of Dialogue Rules on Reddit Using Natural Language Processing

Research Report Submitted in Partial Completion of

PB312: Research Apprenticeship during BSc in Psychological and Behavioural Science

Awarded First-Class

Table of Contents

Theoretical Background

This lab performed an exploratory investigation into the structure of subreddit rules on Reddit to identify salient patterns in the dialogue rules upheld by its communities. Reddit is an online discussion website comprised of individual communities centred around a topic, known as subreddits. Users engage with subreddits by posting content, commenting on posts or voting on posts and comments. Users are held accountable to both a Reddit-wide content policy that specifies general rules of conduct alongside a set of subreddit-specific rules stipulated and enforced by each subreddit's moderator(s). Moderation is also supported by Reddit's moderation algorithm AutoModerator that flags content for review (Seering et al., 2019).

Examining dialogue rules on online communities is important for several reasons. First, despite the prominence of deliberative theories, the conditions necessary for deliberation remain ambiguous especially in online contexts; further, disproportionate focus has been placed on communication processes rather than elements of institutional design, such as rules and norms (Friess & Eilders, 2015). Second, this research sheds light on the domain of netiquette that reflects how social norms are manifested through digitally mediated interactions (Heitmayer & Schimmelpfennig, 2023). Reddit, for example, stipulates its own set of norms in the form of Reddiquette. Understanding the nature of dialogue rules on digital platforms can guide the development of normative frameworks for novel platforms and inform the construction of moderation technologies used to enforce rules (Chandrasekharan et al., 2018). Thus, interdependencies between technological features of online platforms and the norms and rules relevant to its stakeholders positions this research focus as a sociotechnical phenomenon (Chandrasekharan et al., 2019). This is an increasingly important focus given the growing role of online social platforms in facilitating public discourse regarding important social issues that are often accompanied by incivility in the form of

flaming and trolling which hinder the quality of demographic discourse (i.e., messages intended to express anger or provocation towards another user, respectively) (Lampe et al., 2014). Indeed, research has shown that the linguistic content of comments differed across subreddits with policies of ‘safe space’ and ‘free speech’, even for the same user (Gibson, 2019). This suggests that subreddit rules guide differences in the affect, style and topic employed in public discourse, reinforcing the relevance of this lab.

Reddit is a suitable medium for this research as it provides a large naturally occurring sample of rules across a heterogenous range of topics. Further, Reddit’s focus on long-form discussion renders it as a platform that has robust content moderation practices centred around maintaining discussion quality. Extant research has identified a three-level structure to norms on Reddit, whereby norm violations were clustered at macro, meso and micro levels (Chandrasekharan et al., 2018). Macro norms were consistent across the majority of subreddits and included, for example, the prohibition of hate speech, personal attacks and abuse of moderators. Meso norms were enforced across a selection of subreddits (e.g., prohibition of personal reactions); micro norms were highly specific and only enforced in a minority of subreddits (e.g., high-school science theories). Our lab furthers this research by examining the underlying structure of rules as opposed to the patterns in norm violations, illustrating the normative expectations of dialogue in contrast to what is being enforced. This approach can clarify both desired and prohibited behaviour while revealing the broader normative landscape on Reddit.

Motivated Hypotheses

As this lab was exploratory in scope, no confirmatory hypotheses were tested.

Methods

This lab employed natural language processing (NLP) to examine dialogue rules. NLP is a subdomain within computer science focused on applying computational methods to

examine human language and has recently been applied to big data given the ubiquity of digital text (Hirschberg & Manning, 2015). The analysed dataset was extracted using Reddit Application Programming Interface (API) that enabled public access to a subset of Reddit's data. A total of 11,000 subreddits and constituent 48,769 rules were extracted in their English format. 10,000 of the subreddits were extracted based on highest popularity in 2018, and another 1,000 were extracted based on highest popularity when the code was executed in November 2022.

Topic modelling was applied to discern prominent patterns in rules across all subreddits. Topic modelling is a NLP approach that identifies latent topics within a dataset (i.e., corpus) whereby each topic represents a semantic concept (Zhao et al., 2021). This lab utilised latent Dirichlet allocation (LDA) to classify the rules in the corpus. Proposed by Blei, Ng and Jordan (2003), LDA is a topic modelling method that classifies words (i.e., features) within a given piece of text (i.e., document) into one or multiple topics. Each topic represents a distribution over features, and features that occur together are likely classified into the same topic. In this way, LDA reduces the dimensionality of the corpus by grouping semantically similar rules together to enable inference of the broader topics to which they belong. LDA is an unsupervised machine learning algorithm as it does not require predetermination of an expected response variable (James et al., 2021). Thus, this approach was suitable for this exploratory lab as there was no clear expectation of the structure of the underlying topics.

LDA was employed in R with Benoit et al.'s (2018) quanteda package. Symbols, punctuations, stopwords and numbers prespecified by quanteda were removed during pre-processing to ensure only meaningful features were analysed. Upon manual inspection, a remaining set of 18 high frequency stopwords and symbols were deemed unmeaningful and removed. The features were subsequently converted into a document-feature matrix (DFM) which mapped the frequency of each feature across each document. One necessary condition

of LDA is specifying the parameter *k* which determines the number of topics generated. However, subjective specification of *k* is unideal as it is requires subjective interpretation of model fit and is often time-consuming and tedious (Zhao et al., 2015). This challenge was addressed by employing two hyperparameter tuning techniques to ensure precise selection of *k*. First, ldatuning tested model fit specifying *k* = 2, 3, 4, 5, 10, 20, 30, 40, 50, 75, 100 and 200 as assessed by three metrics conventionally utilised in LDA optimisation (see Murzintcev, 2016). Next, 5-fold cross validation further optimised selection of *k*, testing *k* at the same values. Cross validation is a technique used to assess model performance by randomly dividing the dataset into *n* subsets (folds) where one of which is a validation set and *n* - 1 of which are training sets (James et al., 2021). The model is fitted to the training sets and tested on the validation set until all *n* folds have been tested. This method is advantageous as it tests model performance on novel data instead of the same data used to train the original model which can overestimate model fit. Consistent with Blei, Ng and Jordan (2003), model performance was assessed using perplexity: a conventional metric within text modelling that reflects a model's ability to generalise to new datasets. Perplexity was calculated for the observations in the validation set, with the process iterated five times using a different validation set each time. The perplexity metric for a given model was derived from averaging the perplexity scores across the five folds. The application of two tuning methods ensured that *k* was optimised based on consideration of several metrics. Open-coding was used to manually assign topics in the final LDA model to subreddit rules using the 10 highest frequency features for each topic, consistent with Chandrasekharan et al. (2018).

Co-occurrence network analysis was employed as an exploratory extension to the primary LDA approach to further probe the associations between features. Co-occurrence analysis is a graph-based method which maps features within a document as nodes, and

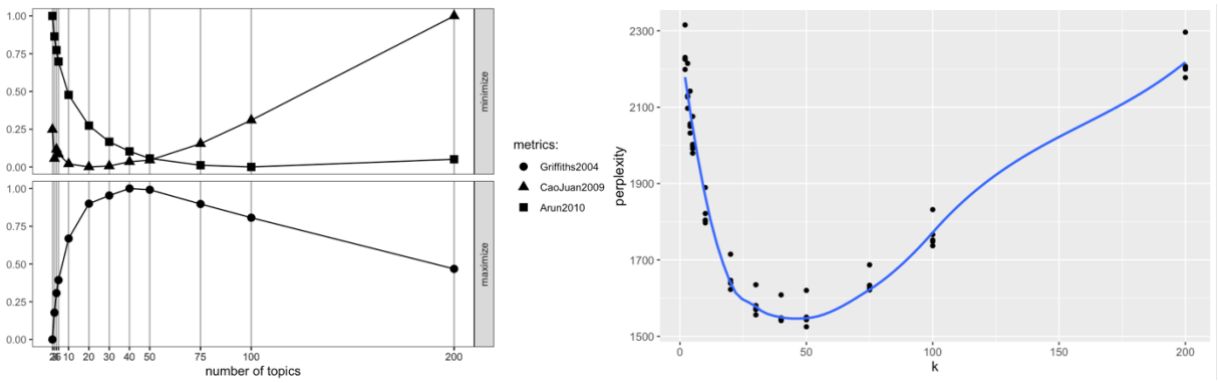
connections between the features as edges; edges are determined by the co-occurrence of *n* features within a window of reference (Millington & Luz, 2021). This lab examined two-feature co-occurrence at a sentence level to identify the most common features co-occurring within the same sentence, as well as the most common features co-occurring with the target feature 'quality'. Quality was chosen as the target as it identified relevant features that, in principle, would be associated with rules that were linked to dialogue quality. Co-occurrence analysis was employed in R. The original corpus was segmented into sentences to facilitate analysis of co-occurrences at a sentence level. The pre-processing stages employed in the LDA model were replicated to transform the corpus into a DFM. Quanteda was used to generate frequency and lambda statistics, whereby lambda assesses the probability of a given co-occurrence relative to chance; standardised lambda was used to control for co-occurrence frequency (Benoit et al., 2018). Next, the statistical significance of the co-occurrences containing the target feature 'quality' was calculated, using log-likelihood.

Data Analyses

Ldatuning and 5-fold cross validation was first employed to ascertain the optimal number of topics, *k,* for LDA modelling. Both tuning techniques found that model performance was optimised when $k \cong 40$ to 50; performance results are visualised in Figure 1.

Figure 1

Performance of LDA Models Tested Using Ldatuning and 5-Fold Cross-Validation



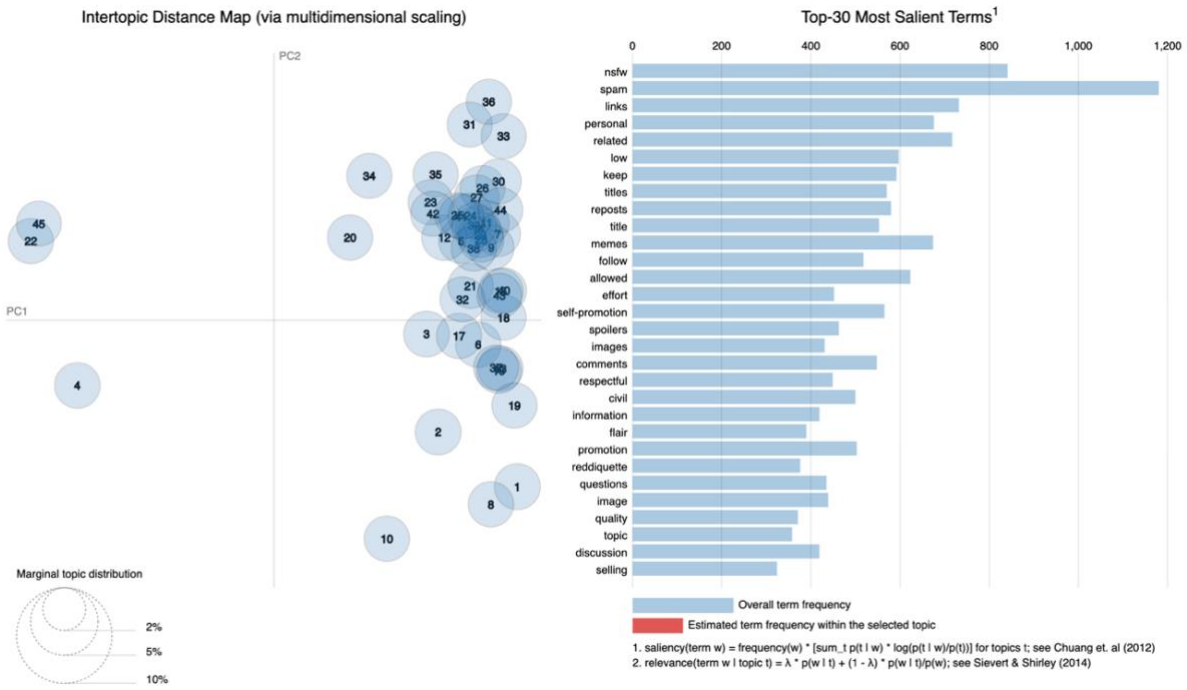
Note. (L) Performance of the Ldatuning package in optimisation of three metrics (see Murzintcev, 2016). (R) Performance of 5-fold cross-validation in optimisation of perplexity metric; lower scores indicate better performance.

Thus, a LDA model specifying 45 topics was fitted to the processed DFM. Initial inspection of the frequency plot in Figure 2 revealed that NSFW (i.e., 'not safe for work'), spam, links, personal and related constituted the top 5 most salient features. Further, the intertopic distance map visualises the distance between topics as calculated by Jensen-Shannon divergence: closer distances on the map indicate more common features shared between topics (see Sievert & Shirley, 2014). The map showed that most topics clustered closely together, with only topics 4, 22 and 45 occupying the left plane (see Appendix for features differentiated by topic). This suggested that most of the topics shared common features and thus, likely formulated rules that enforced similar behaviours. Dialogue rules derived from open-coding are shown in Table 1; analysis of dialogue rules revealed the emergence of three clear categories of topics: stipulations of civility, prohibitions of specified content and descriptions of proper posting and dialogue etiquette. Topics within these categories comprised 29 of the 45 total topics generated, while five of the 45 topics did not present a discernible rule and were labelled as miscellaneous: this could be due to computational noise or the absence of contextual knowledge of the subreddits (Chandrasekharan et al., 2018). Topics 1, 9, 18, 19, 25, 26 and 27 focussed on rules including the prohibition of personal attacks, maintaining respect, delivering criticism constructively, maintaining civility and avoiding offensive behaviour, discrimination and trolling. Topics 40 and 45 shared many common features of topic 19 (i.e., civility) and were therefore determined to be repeated topics of topic 19. These topics can all be broadly categorised as stipulations of civility. Additionally, topics 3, 5, 6, 12, 17, 20, 29, 34 and 36 focussed on rules including the prohibition of links, explicit content, advertisements, exchange of goods (i.e.,

buying, selling, trading), spam, inappropriate content, NSFW and piracy and illegal behaviour. Topics 7 and 38 were deemed to be repeated topics of topic 6 (i.e., advertisements). These topics can be categorised as prohibitions of specified content. Third, topics 8, 16, 23, 33, 35, 37, 42 and 44 focussed on rules including proper conduct regarding the attribution of artistic sources, posting relevant content, providing accurate context (i.e., without the intent to mislead), ensuring quality and attribution, discouraging low-quality content in the context of jokes, ensuring adherence to the topic, communicating content in a user-friendly way and ensuring proper format. Topic 39 was deemed be to a repeated topic of topic 8 (i.e., attribution of artistic source). These topics, therefore, appear to broadly represent descriptions of proper posting and dialogue etiquette. Together, the similarities between the topics affirmed the close intertopic distances observed between most topics.

Figure 2

Visualisation of LDA Model Fitted With 45 Topics



Note. Intertopic distance computed via Jensen-Shannon divergence and scaled using Principle

Components (see Sievert & Shirley, 2014); salience reflects the degree to which a feature is informative in differentiating between topics.

Table 1

Dialogue Rules Derived From Open-Coding of 45 Topics

1. Personal attacks	10. Limits in frequency over time	19. Civility	28. Role of moderators	37. Adherence to topic
2. Miscellaneous	11. Reddit terminology	20. Advertisements (repeated)	29. Inappropriate content	38. Advertisements and exchange of goods
3. Links	12. Exchange of goods	21. Action verbs in Reddit usage	30. Miscellaneous (repeated)	39. Attribution of artistic source (repeated)
4. Spoilers	13. Miscellaneous (repeated)	22. Miscellaneous (repeated)	31. Music	40. Civility (repeated)
5. Explicit content	14. Adherence to Reddit content policy	23. Accurate context	32. Threads (repeated)	41. Visual content (repeated)
6. Advertisements	15. Visual content	24. Medical and legal advice	33. Quality and attribution	42. User friendly
7. Advertisements (repeated)	16. Relevance	25. Offensive behaviour	34. NSFW	43. Formatting
8. Attribution of artistic source	17. Spam	26. Discrimination	35. Quality and jokes	44. Miscellaneous (repeated)
9. Respect	18. Constructive criticism	27. Trolling	36. Piracy and illegal behaviour	45. Civility (repeated)

Note. NSFW represents ‘not safe for work’ and is commonly destinated towards content inappropriate to be consumed in public.

Co-occurrence analysis demonstrated that low effort (n = 606), personal information (n = 371), follow Reddiquette (n = 298), low quality (n = 225) and comment karma (n = 161) constituted the five most frequent co-occurring features within a given sentence of the corpus; the co-occurrence statistics are displayed in Table 2. Further, the graphed network of features co-occurring with the target feature ‘quality’ is displayed in Figure 3: the density of the edges demonstrated that ‘quality’ most significantly co-occurred with ‘make’ and ‘low’, which themselves co-occurred with ‘submissions’ and ‘spam’, respectively. These co-occurrences indicated that rules stipulating quality concentrated on ensuring users make quality submissions and produce low spam. ‘Reddiquette’, ‘low-effort’, ‘trolling’, ‘civil’ and ‘memes’ were also prominent within the network.

Table 2

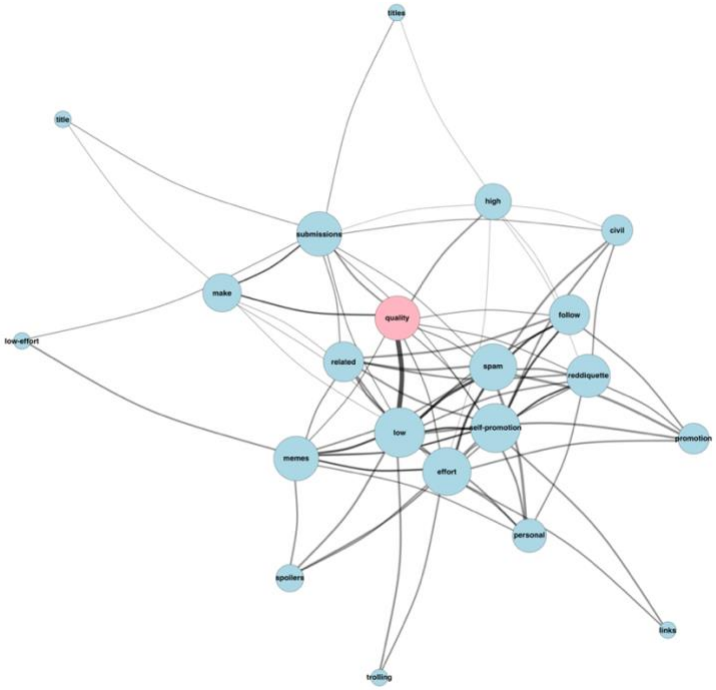
Co-Occurrence Statistics

Co-Occurrence	Count	λ	*z*
1. personal information	371	5.52	62.82
2. low effort	606	8.09	61.32
3. follow reddiquette	298	5.53	59.33
4. low quality	225	4.89	51.54
5. comment karma	161	6.28	49.10
6. provide context	152	7.14	47.61
7. new users	151	5.98	47.20
8. accounts days	146	6.84	46.64
9. users accounts	142	5.76	46.50
10. mark nudity	134	6.32	46.17

Note. Co-occurrence was calculated using λ (lambda). The top 10 co-occurrences are shown, arranged in descending order of *z* (standardised λ).

Figure 3

Undirected Network Graph of Co-Occurrence *Analysis*



Note. Nodes represent features and edges represent co-occurrences with the target feature ‘quality’. Density of edges represent the significance of the co-occurrence via log-likelihood.

Discussion

The findings provide evidence that rules across subreddits share significant similarities and can be broadly described by three categories: stipulations of civil behaviour, prohibitions of specified content and descriptions of proper posting and dialogue etiquette. Given the diversity in topic and audience across the 11,000 subreddits, it appears that these categories are important for dialogue across Reddit. Based on interpretation of feature specificity, this report contends that civility rules encompass common-sense principles of civility assumed in daily life (e.g., respect, prohibition of offensive attacks); by contrast, content and etiquette rules are stipulated with greater specificity and might not always be apparent to new users (e.g., attribution of artistic sources). This contrast in specificity suggests that subreddit rules reflect both medium-universal and medium-specific norms, whereby a common-sense understanding of civility underpins a set of more context-dependent understanding of dialogue on Reddit. Further, significant focus is placed on ensuring posts and comments are of quality, do not include personal information and follow Reddiquette: this deference to Reddiquette implies that subreddits share many common norms that are encapsulated within Reddiquette.

The research supports and extends existing work in several ways. First, the consistency of rules around civility, appropriate content and proper etiquette suggests that that these norms largely invariably underpin dialogue independent of discussion topic. This consistency affirms Chandrasekharan et al.'s (2018) finding that general uncivil behaviour was classified as a macro norm and was universally treated as norm violations across subreddits. Second, the findings support Chandrasekharan et al.'s (2018) observation that norms display a nested structure and are adopted from both a general social context and from Reddit-specific etiquette alongside broader internet etiquette. This lab found that norms adopted from the general social context are stipulated in general terms and mainly focussed

on civility norms, while norms adopted from Reddit and internet-specific etiquette are conveyed with more specificity and concentrated around content and etiquette norms. Third, this lab adds nuance to Heitmayer and Schimmelpfennig (2023) by demonstrating that content rules are overwhelmingly negatively determined (e.g., prohibition of links, explicit content, spam), compared to civility and etiquette rules. Fourth, the research highlights the significance of the normative guidelines stipulated by Reddiquette, even as subreddits are entitled to their own subreddit-specific rules. This further reinforces the position that many norms are context-independent and points to the importance of a robust set of centralised norms that sub-communities tend to rely upon.

These findings contribute towards a framework of norms that are likely useful for online communities to consider. Specifically, norms surrounding civility, content and posting and dialogue etiquette are three domains that can provide a systematic framework to the development of rules on online social platforms. This framework can be especially useful for new communities where norms may not be clearly distinguished. Moreover, platforms may benefit from differentiating rules that are mainly derived from a general social context compared to those that are unique to the platform and are not common-sense principles. For instance, the prohibition of advertisements and the proper attribution of artistic sources was found to be a consistent rule on Reddit, while these rules may not apply to other online communities. Thus, users switching between platforms will likely be required to adapt their behaviour surrounding content and etiquette but not civility. From a sociotechnical perspective, platform architecture can consider the ways in which platform-specific norms are presented, especially for new users unfamiliar with the platform's norms and may violate norms due to lack of awareness rather than malice.

The research contains some limitations and illuminates directions for future research. First, LDA is limited by its underlying 'bag of words' approach, which disregards

grammatical structure including word order that may be important in understanding the relationships between features and providing context for topic inference (Wallach, 2006). Second, the analysis of subreddit rules examines expected behaviour but does not capture the extent users conform to the rules. Relatedly, subreddit rules are injunctive as they prescribe desired conduct but do not account for descriptive norms (i.e., how other users behave): Heitmayer and Schimmelpfennig (2023) contend that both are important in shaping netiquette and behaviour. Third, Reddit's dialogue rules possibly disproportionately represent the discourse norms held by individuals from WERID (Western, Educated, Rich, Industrialised, Democratic) populations (Henrich, Heine & Norenzayan, 2010). If so, this disposition towards *cultural standardisation* may imply that the insights are limited in their cross-cultural generalisability (Marcoccia, 2012). Similarly, as Reddit is situated within a unique architecture and likely attracts a specific user demographic, future research can examine whether the identified rules are equally upheld on other online communities, including communities that provide more representative cross-cultural samples. Finally, future work can consider the temporal aspect of norms; for instance, dynamic topic modelling can be used to map how norms change over time (Chandrasekharan et al., 2018). This is an important extension as community rules are shown to evolve over time as a function of rule violations, enforcement and creation (Sternberg, 2012, as cited in Seering et al., 2019).

Concluding Remarks

This lab investigated the structure of subreddit rules across a total of 11,000 subreddits and constituent 48,769 rules. Application of NLP methods found that rules clustered towards three categories: stipulations of civility, prohibitions of specified content and descriptions of proper posting and dialogue etiquette. The lab found that civility rules were general while content and etiquette rules were specified with greater detail, affirming Chandrasekharan et al.'s (2018) position that Reddit norms demonstrate a nested structure.

Further, low quality and low effort content was heavily discouraged and Reddiquette established important norms across subreddits. The report concluded with a discussion of employing a normative framework to underpin novel platforms, considered the role of platform architecture in rule enforcement and outlined several directions for future research.

References

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774-774.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993-1022.

Chandrasekharan, E., Gandhi, C., Mustelier, M. W., & Gilbert, E. (2019). Crossmod: A cross-community learning-based system to assist reddit moderators. *Proceedings of the ACM on human-computer interaction*, *3*(CSCW), 1-30.

Chandrasekharan, E., Samory, M., Jhaver, S., Charvat, H., Bruckman, A., Lampe, C., ... & Gilbert, E. (2018). The Internet's hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1-25.

Friess, D., & Eilders, C. (2015). A systematic review of online deliberation research. *Policy & Internet*, *7*(3), 319-339.

Gibson, A. (2019). Free speech and safe spaces: How moderation policies shape online discussion spaces. *Social Media+ Society*, *5*(1), 2056305119832588.

Heitmayer, M., & Schimmelpfennig, R. (2023). Netiquette as Digital Social Norms. *International Journal of Human–Computer Interaction*, 1-21.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and brain sciences*, *33*(2-3), 61-83.

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, *349*(6245), 261-266.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: with applications in R (Second Edition)*. Springer.

Lampe, C., Zube, P., Lee, J., Park, C. H., & Johnston, E. (2014). Crowdsourcing civility: A natural experiment examining the effects of distributed moderation in online forums. *Government Information Quarterly, 31*(2), 317-326.

Marcoccia, M. (2012). The internet, intercultural communication and cultural variation. *Language and Intercultural Communication, 12*(4), 353-368.

Millington, T., & Luz, S. (2021). Analysis and classification of word co-occurrence networks from Alzheimer's patients and controls. *Frontiers in Computer Science*, *3*, 649508.

Murzintcev, N. (2016). Ldatuning: Tuning of the latent dirichlet allocation models parameters. *R package version 0.2-0, See: https://CRAN. R-project. org/package= ldatuning*.

Seering, J., Wang, T., Yoon, J., & Kaufman, G. (2019). Moderator engagement and community development in the age of algorithms. *New Media & Society*, *21*(7), 1417-1443.

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63-70).

Sternberg, J. (2012). *Misbehavior in cyber places: The regulation of online conduct in virtual communities on the Internet*. Rowman & Littlefield.

Wallach, H. M. (2006). Topic modeling: beyond bag-of-words. *Proceedings of the 23rd international conference on Machine learning* (pp. 977-984).

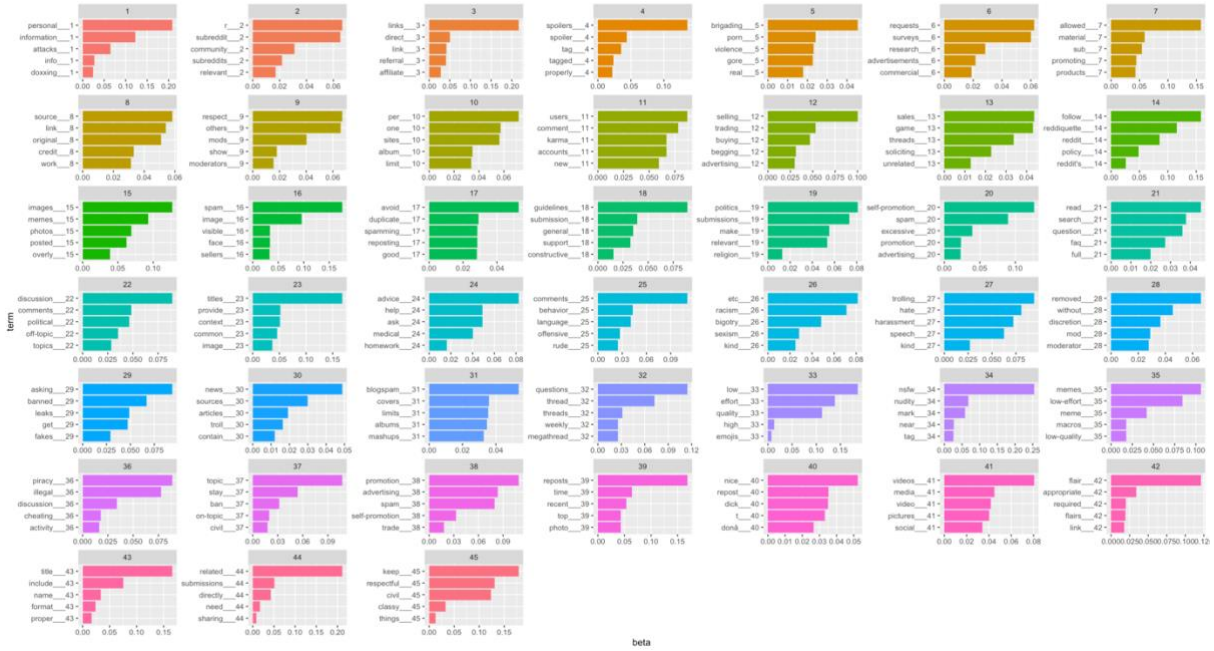
Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., & Zou, W. (2015, December). A heuristic approach to determine an appropriate number of topics in topic modeling. *BMC Bioinformatics* (Vol. 16, pp. 1-10). BioMed Central.

Zhao, H., Phung, D., Huynh, V., Jin, Y., Du, L., & Buntine, W. (2021). Topic modelling meets deep neural networks: A survey. *arXiv preprint arXiv:2103.00498*.

Appendix

Visualisation of Top 5 Most Common Features by Topic

Appendix



Note. Beta represents the probability that a term is association with the given topic.