Research Notebook

Weekly Entry: Community Language Analysis

This week's work consisted of improving the linguistic analysis of Reddit communities by building a suite of Natural Language Processing (NLP) methods to investigate the ways in which various types of community's express identity, tone and interactional behavior. These analyses were run on a variety of subsets of subreddits such as ideology forums, support-based spaces, and hobby based groups. The larger aim was to understand better how language is employed to talk inclusion, depict emotion, or remove distance between these groups.

Based on other work done on pronoun usage and sentiment analysis, the current analysis built upon five complementary linguistic features. In the language analysis of tentative vs. assertive, how frequently do users use the cautious or confident aspects of their language was examined. The strength of words such as "maybe," "perhaps" and "I think" was classified as tentative and the strength of "definitely," "must" and "clearly" as assertive. It helps the community to distinguish whether the community fosters openness and careful discussion, or strong direct messaging.

The second method evaluated the use of emotional vs. cognitive language. To indicate whether conversations were based more on personal feelings (e.g., happy, sad, worried) or reasoning and critical thought (e.g., think, know, explain), emotional words (e.g., happy) were contrasted with cognitive terms (e.g., think). This method is effective in differentiating support-oriented communities like r/mentalhealth compared to debate oriented communities like r/PoliticalDebate that use cognitive language.

The third concerned politeness and rudeness. A civility balance was built to number polite words (thank you, please, appreciate) against rude expressions (idiot, stupid, shut up). This analysis shows the overall way communication is being done and may show that a community tends toward supportive dialogue or confrontational language.

The fourth method consisted in tracking question usage through question marks and question words (why, what, how). This metric gives us an idea on how curious or interactive a community is (or is not), which is usually a sign of an open and willing to engage community. For example, communities that are good for inviting discussion will include high question frequency, such as r/ChangeMyView or r/OutOfTheLoop.

Finally, lexical variety was counted to determine the number of words in each community. The type-token ratio (unique words count / total word count) was used to do this. The amount of lexical diversity in words, whether low or high, usually signals thoughtful and varied responses or, conversely, repetitiveness or echo chamber behaviour.

A relevant study by Zirikly et al. (2023), "Style Matters: Investigating Linguistic Style in Online Communities" (arXiv:2302.10172) was reviewed in order to support these analyses. The authors show that in natural communities, different communities naturally adopt different linguistic styles to further their social goals such as personal narratives and emotional expression in support groups and reasoned argumentation in ideological forums. This only reinforces what we can find on this project, that this is an analytical approach to mapping how language reflects the structure, the norms and the dynamics of the online communities.

Taken together, these methods offer some useful things to say about how language gets used to make an identity, make someone included or keep a group distinct in the online world. The research looks at multiple linguistic dimensions as it pertains to different community types to provide a more holistic picture of how online group behaviour plays out in the textual communication.