Understanding Toxicity and Sentiment Dynamics in Social Media: LLM Analysis of Diverse and Focused Interest Users

Abi Oppenheim[1] and **Federico Albanese**[2,3] and **Esteban Feuerstein**[1,3]
{aoppenheim, falbanese, efeuerst}@dc.uba.ar

[1]Departamento de Computación, Facultad de Ciencias Exactas y Naturales,
Universidad de Buenos Aires, Buenos Aires, Argentina.

[2]Instituto de Cálculo, CONICET Universidad de Buenos Aires, Buenos Aires, Argentina.

[3]Instituto de Ciencias de la Computación, CONICET Universidad de Buenos Aires, Buenos Aires, Argentina.

Abstract

Online platforms host users with diverse interests, ranging from those with focused interests to those engaging in a wide range of topics. This study investigates the behavioral differences between these user types on Reddit, focusing on the level of toxicity in their posts and associated sentiment scores across nine emotional categories. By employing community embeddings to represent users in a high-dimensional space, we measure activity diversity using the GS score. The analysis utilizes a dataset of 16,291,992 posts from 4,926,237 users spanning 2019 to 2021, assessing toxicity and sentiment scores for each post. Results indicate that subreddits characterized by users with specialized interests exhibit heightened toxic behavior compared to those with diverse interest users. Additionally, subreddits populated by users with focused interests display elevated sentiments of sadness, annoyance, and disappointment, while those inhabited by diverse interest users demonstrate increased expressions of curiosity, admiration, and love. These insights contribute to understanding user behavior on online platforms and inform strategies for fostering healthier online communities.

1 Introduction

Social media platforms have become breeding grounds for toxic behaviors, hate speech, and online harassment, posing significant challenges to digital communities. A recent report (Thomas et al., 2021) underscores the widespread prevalence of online threats, with nearly half (48%) of individuals worldwide reporting encounters with such menacing behavior. Marginalized communities, in particular, bear a disproportionate burden of online harassment (Golbeck, 2018; Massanari, 2017), leading to detrimental effects on mental health and civil discourse. Various studies have explored methods for detecting and analyzing toxic texts and emotions, including sentiment analysis on COVID-19 vaccine-related discussions (Melton et al., 2021), examining toxicity in political contexts (Albanese et al., 2023), analyzing toxic language usage on Facebook in relation to political interest (Kim et al., 2021), and studying toxicity in conversations prompted by tweets from news outlets and political candidates(Saveski et al., 2021). Building upon this research, our work integrates these strands with the methodology proposed by Waller and Anderson (Waller and Anderson, 2019). Their study investigates user behavior by distinguishing between generalist and specialist tendencies and utilizes community embeddings to represent online communities. They employ the community2vec algorithm (Martin, 2017) to assign vectors to communities, facilitating the quantification of similarity between communities based on cosine similarity in the vector space. Additionally, they introduce the Generalist-Specialist score (GS-score) to measure a user's activity diversity, where specialists have a low GS-score, contributing to a tight cluster of communities, while generalists have a high GS-score, contributing to diverse, distant communities.

However, up to our knowledge, the relationship between user activity diversity and the level of toxicity in their online interactions remains unexplored. In this study, we hypothesize that communities comprising users with a broader array of interests, are less likely to have harmful behavior compared to communities dominated by those who concentrate on a limited range of topics. To test our hypothesis, we analyze a dataset of 16,291,992 posts from 4,926,237 users on Reddit, spanning the period from 2019 to 2021. By employing community embeddings to represent users in a high-dimensional space, we measure activity diversity using the GS score. We assess the degree of toxicity and sentiment scores across 9 emotional categories for each post. Our findings reveal two key observations.

- **Diversity of Interest and Toxic Behavior**. Our investigation shows that subreddits cultivated by individuals with specialized interests experience more toxic behavior than those curated by individuals with diverse interests.

- **Emotions Across Communities Types**. Subreddits shaped by specific interest users show more sadness, annoyance, and disappointment, while those by diverse interest users display higher levels of admiration, amusement, and affection.

These findings underscore a compelling link between the diversity of user interests and a notable decrease in toxic behavior observed on social media platforms.

2 Methodology

2.1 Dataset

We utilized the Pushshift API, a platform for collecting, analyzing, and archiving social media data (Baumgartner et al., 2020). The Pushshift dataset encompasses submissions and comments posted on Reddit since June 2005, offering extensive query limits and user-friendly access. To ensure scalability and representativeness, we randomly selected 15,000 submissions from each of the top 5,000 most popular subreddits, evenly distributed across 2019 to 2021. Our dataset comprises 16,291,992 posts from 4,926,237 users.

2.2 Generalist-Specialist Score (GS-score)

The GS-score quantifies user activity diversity, distinguishing between specialized and generalized engagement patterns (Waller and Anderson, 2019). Specialized users concentrate activity within specific areas of the community space, while generalist users show broader dispersion.

To analyze user interactions across diverse communities, we first generate community embeddings using the community2vec approach (Waller and Anderson, 2021; Martin, 2017). By treating communities as "words" and users submitting in them as "contexts", communities were embedded into a high-dimensional vector space. The proximity of two communities in this space is determined by the frequency of users posting in both communities. We trained the skip-gram model with negative sampling using all pairs (c_i, u_j) of users u_j submitting in community c_i, to generate an initial community embedding.

Let user u_i make w_j contributions to community c_j, and let $\vec{c_j}$ denote the normalized vector representation of c_j in our community embedding. The center of mass of u_i is defined as $\vec{\mu_i} = \sum(w_j * \vec{c_j})$. The diversity of u_i's activity, or GS-score, is the weighted average cosine similarity between u_i's communities and its center of mass:

$$GS(u_i) = \frac{1}{\sum w_j} \sum_j w_j \frac{\vec{c_j} \cdot \vec{\mu_i}}{\|\vec{\mu_i}\|}$$

The GS-score ranges from -1 to 1, where -1 indicates extreme generalization and 1 indicates extreme specialization. For detailed insights, refer to Waller and Anderson (Waller and Anderson, 2019). Once user activity diversity is computed, we extend this analysis to understand community activity diversity. The GS-score $GS(c_i)$ of community c_i is the weighted average across its users: $GS(c_i) = \frac{1}{N} \sum_j w_j \cdot GS(u_j)$, where u_j contributes w_j times to c_i and $N = \sum_j w_j$ total contributions. Prioritizing community metrics enhances robustness and comprehensive evaluation of submission sets within communities.

2.3 Emotion and Toxicity Analysis

For emotion classification, we utilized the pre-trained Language Model (LLM) "SamLowe/roberta-base-go_emotions" (Demszky et al., 2020), fine-tuned on the GoEmotions dataset (Liu et al., 2019), which consists of 58,000 Reddit comments annotated across 27 emotional categories, including Neutral. This model's training on Reddit-specific data makes it well-suited for our analysis and has been utilized in various scientific studies (Kocoń et al., 2023; Mostafazadeh Davani et al., 2022; Vu et al., 2021). Considering the vast range of emotions, we focused on specific subsets relevant to our hypothesis, including admiration, amusement, anger, annoyance, disappointment, gratitude, joy, love, and sadness. This model was applied to over 200 pertinent subreddits within our corpus.

To evaluate toxicity, we employed the compact Detoxify Model developed by Unitary AI (Hanu and Unitary team, 2020), trained on a substantial dataset of Wikipedia comments annotated for various forms of toxic behavior (Lan et al., 2019). This model outputs probabilities indicating the likelihood of input text containing language associated with different toxic categories. Over 1100 relevant subreddits were analyzed using this approach.

3 Results

3.1 Subreddit Distribution Analysis

To examine user behavior across subreddit types, we categorized the subreddits into two groups: the bottom 10th percentile of GS-scores, representing specialized-interest user communities, and the upper 10th percentile, representing diverse-interest user communities, each comprising 20 subreddits. To gain insights into the dataset's distribution, Figure 1 presents a histogram illustrating the distribution of GS-scores among the subreddits. The blue

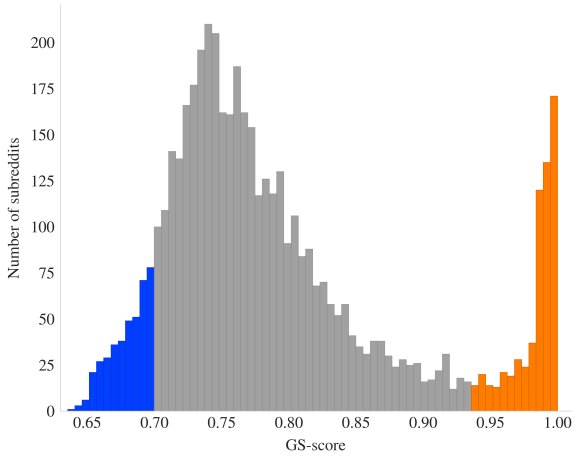

Figure 1: Distribution of GS-scores among subreddits

bars represent subreddits within the lower 10th percentile of GS-scores, while the orange bars represent those within the upper 10th percentile. Notably, the GS-scores predominantly cluster within the range of 0.6 to 1, aligning with prior findings (Waller and Anderson, 2019).

3.2 Emotion and Toxicity Analysis

For each group, we calculated average emotion and toxicity scores by determining the proportion of users expressing specific emotions relative to the total number of users. To explore variations in emotional expression, we employed the Mann-Whitney U test, suitable for comparing two independent groups when normality assumptions are not met. Our findings indicate significant differences between subreddits frequented by users with diverse and specialized interests, as evidenced by p-values smaller than 0.05. Subreddits with the highest GS-scores exhibited higher toxicity scores compared to those with the lowest GS-scores ($r = .44$). Conversely, subreddits with the lowest GS-scores demonstrated higher scores for emotions such as joy ($r = .48$), anger ($r = .56$), gratitude ($r = .44$), love ($r = .53$), amusement ($r = .58$), and admiration ($r = .4$).Notably, all statistically significant results align with our hypothesis. However, disappointment ($p = .394, r = 0.16$) and sadness ($p = .968, r = .01$) did not show significant differences between the two types of subreddits. Figure 2 compares emotion and toxicity scores between subreddit types.

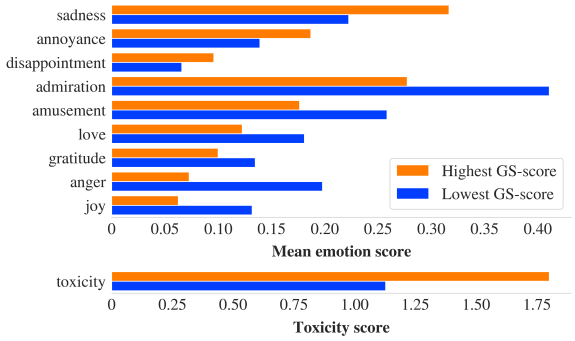

Figure 2: Emotion and toxic scores comparison between subreddits with highest and lowest GS-score

While the lowest GS-score subreddits demonstrated a higher score for anger compared to the other subreddit type, it is essential to differentiate between "angry" and "toxic" content in text analysis. The distinction between anger and toxicity lies in their respective expressions: while anger typically manifests as frustration towards situations, objects, or groups, toxicity often involves directing abusive language or behavior towards individuals.

4 Conclusions

Our study delved into the connection between emotion, toxicity, and activity diversity on Reddit. We found that subreddits characterized by users with diverse interests tend to have lower toxicity and higher positive emotion scores, including joy, gratitude, love, and admiration. Conversely, subreddits shaped by users with more specific interests exhibit higher toxicity and increased negative emotions like sadness, annoyance, and disappointment. These findings underscore the significance of activity diversity in understanding user behavior and its impact on online community dynamics. Our research enhances comprehension of the factors influencing user conduct and offers insights for fostering healthier online interactions. Future studies could investigate causal links between activity diversity, emotion, and toxicity, as well as explore interventions to encourage positive engagement and mitigate toxicity on online platforms.

References

Federico Albanese, Esteban Feuerstein, Gabriel Kessler, and Juan Manuel Ortiz de Zárate. 2023. Aprendizaje automático para el análisis cross-plataforma de la comunicación política: Gobierno y oposición argentinos en facebook, instagram y twitter. *Cuadernos.info*, (55):256–280.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):830–839.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions.

Jennifer Golbeck. 2018. *Online Harassment: A Research Challenge for HCI*, pages 1–2. Springer International Publishing, Cham.

Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

Jin Woo Kim, Andrew Guess, Brendan Nyhan, and Jason Reifler. 2021. The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity. *Journal of Communication*, 71(6):922–946.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, Bartłomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Miłkowski, Marcin Oleksy, Maciej Piasecki, Łukasz Radliński, Konrad Wojtasik, Stanisław Woźniak, and Przemysław Kazienko. 2023. Chatgpt: Jack of all trades, master of none. *Information Fusion*, 99:101861.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach.

Trevor Martin. 2017. community2vec: Vector representations of online communities encode semantic relationships. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 27–31, Vancouver, Canada. Association for Computational Linguistics.

Adrienne Massanari. 2017. #gamergate and the fappening: How reddit's algorithm, governance, and culture support toxic technocultures. *New Media Soc.*, 19(3):329–346.

Chad A. Melton, Olufunto A. Olusanya, Nariman Ammar, and Arash Shaban-Nejad. 2021. Public sentiment analysis and topic modeling regarding covid-19 vaccines on the reddit social media platform: A call to action for strengthening vaccine confidence. *Journal of Infection and Public Health*, 14(10):1505–1512. Special Issue on COVID-19 – Vaccine, Variants and New Waves.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Martin Saveski, Brandon Roy, and Deb Roy. 2021. The structure of toxic conversations on twitter. In *Proceedings of the Web Conference 2021*, WWW '21. ACM.

Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, Damon McCoy, Sarah Meiklejohn, Thomas Ristenpart, and Gianluca Stringhini. 2021. Sok: Hate, harassment, and the changing landscape of online abuse. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 247–267.

Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou, and Daniel Cer. 2021. SPoT: Better frozen model adaptation through soft prompt transfer.

Isaac Waller and Ashton Anderson. 2019. Generalists and specialists: Using community embeddings to quantify activity diversity in online platforms. In *The World Wide Web Conference*, WWW '19, page 1954–1964, New York, NY, USA. Association for Computing Machinery.

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