# Using Sentiment Analysis to measure perception of mental health across different identity groups

**Kylie Van Dyke, Ziqi Zhou, Jiyu Huang**
Golden Data Retrievers
vandy157@umn.edu, zhou2045@umn.edu, huan2803@umn.edu

## Abstract

Large-scale sentiment studies of Reddit's mental-health communities require reliable knowledge of who is speaking, yet demographic cues such as gender, race, and queer identity are rarely explicit and have traditionally been identified by slow, manual coding. We introduce a two-stage pipeline that first filters posts with carefully designed keyword lists and then replaces manual labeling with a lightweight, multi-label DistilBERT classifier. Trained on a gold-standard set of 1,782 hand-annotated posts, the model accurately predicts four binary attributes—mental-health mention, gender identity, racial identity, and queer identity—allowing us to automate what was previously weeks of expert effort. The resulting, fully labeled corpus reveals that sentiment on mental-health subreddits differs markedly across social identities; for example, posts authored by queer-identified users are consistently more negative than those without queer markers, and male-identified posts show slightly more positive sentiment than female-identified ones. By removing the annotation bottleneck, our approach enables far larger, demographically informed analyses of online mental-health discourse.

## 1 Introduction

### 1.1 Problem Definition and Objectives

Mental health issues represent a serious global health concern that disproportionately impacts vulnerable populations. According to the National Institute of Mental Health, approximately 23.1% of adults in the United States have a mental illness (10). Natural language processing (NLP) has been increasingly utilized in the sphere of mental health research for information extraction, sentiment analysis, and emotion detection. Social media platforms like Reddit provide anonymized spaces where individuals discuss personal mental health struggles, seek advice, and provide support. However, the anonymity of these platforms makes it difficult to analyze how different demographics shape users' sentiment or engagement with mental health topics. Our project addresses a significant gap in research analyzing mental health data from social media—the lack of demographic information inclusion in analyses. Current sentiment analyses often apply a generalized perspective without considering demographic factors. This limits our understanding of unique challenges faced by different groups. Our goal is to create a workflow of curating a dataset from Reddit posts and systematically analyze the sentiment towards mental health across different demographic groups to determine the emotional tone (positive, negative or neutral).

### 1.2 Current Approaches and Limitations

The current landscape of mental health analysis using NLP faces several limitations:

- Most research fails to account for demographic variables when analyzing sentiment in mental health discussions
- Manually annotating demographic information is extremely time-consuming and limits sample sizes
- Identity cues appear only sporadically in anonymous posts, making automatic detection challenging
- Existing sentiment analysis models may not capture nuances specific to mental health discourse

Recent studies have demonstrated NLP's ability to detect mental illness and track its progression. For example, Reece et al. (2017) showed that Twitter data could predict the onset and course of mental illness with accuracy comparable to practitioners (4), while Liu et al. (2022) conducted a systematic review of machine learning approaches

for detecting depression on social media (1). Le Glaz et al. (2021) highlighted challenges related to limited labeled datasets in NLP mental health research (7). However, these studies rarely consider how demographic factors influence sentiment patterns. This gap is concerning because research shows significant disparities in mental health experiences across demographic groups. The Trevor Project's 2023 survey found that 41% of LGBTQ+ youth (ages 13-24) had seriously considered suicide (9). Additionally, Olfson et al. (2023) documented substantial racial-ethnic disparities in outpatient mental health care in the United States, with White individuals accessing services at more than twice the rate of Black or Hispanic individuals (11). These disparities highlight the importance of understanding how different demographic groups express mental health concerns online.

### 1.3 Potential Impact and Stakeholders

Our research has potential impact for:

- **Mental health researchers** - Providing insights into how different demographic groups express emotional needs online
- **Healthcare providers** - Revealing patterns that may inform more tailored intervention approaches
- **Social platform developers** - Offering insights into potentially vulnerable user groups
- **Policy makers** - Identifying disparities that could inform mental health resource allocation

By automating demographic identification and analyzing sentiment patterns, our work enables scaling analyses to much larger datasets, potentially revealing insights that would remain hidden in generalized approaches. This could help address the concerning disparities documented by Olfson et al. (2023) (11) by providing a more nuanced understanding of how different demographic groups engage with mental health discourse.

## 2 Approach

### 2.1 Methodology Overview

Our project utilized a three-stage approach:

1. **Manual annotation** of a dataset from mental health subreddits for four attributes: mental health of the post writer, gender identity, racial identity, and queer identity.
2. **Sentiment analysis** using two established tools (VADER and a fine-tuned DistilBERT).
3. **Development of a multi-label classifier** to automate demographic labeling.

#### 2.1.1 Dataset Creation and Annotation

We began by manually annotating a dataset of 1,782 Reddit posts drawn from mental health–related subreddits. Each post was labeled with binary indicators for the four identity categories. This manual process established a reliable "gold standard" dataset. We developed a structured annotation framework to consistently identify demographic information based on textual cues present in the posts.

The annotation process involved:

- Filtering posts with keyword lists for potential demographic mentions
- Manual verification and labeling by multiple annotators
- Recording the specific terms that indicated identity (e.g., "I (25 FtoM) am suffering from depression...")

#### 2.1.2 Sentiment Analysis Implementation

To analyze sentiment patterns across demographic groups, we applied two established sentiment analysis models:

1. **VADER** (Valence Aware Dictionary and Sentiment Reasoner) – A lexicon and rule-based sentiment analysis tool specifically designed for social media content
2. **Fine-tuned DistilBERT** – A lighter transformer-based model pre-trained for sentiment classification.

Each post in our gold-standard dataset was processed through these models to obtain sentiment scores and classifications (positive, negative, or neutral). This approach allowed us to compare sentiment outputs across the manually annotated demographic groups.

#### 2.1.3 Multi-Label Classification Model

Recognizing the scalability limitations of manual annotation, we trained a DistilBERT-based multi-label classifier to automate the demographic labeling process. The model was trained exclusively on our human-annotated gold standard to predict the

presence of the four identity traits simultaneously. This is not a fully realized annotator, as it predicts only the presence of one of the traits and is unable to assign specific labels.

The classifier architecture included:

- Text preprocessing and tokenization using the DistilBERT tokenizer
- A sequence classification head adapted for multi-label prediction
- Training for 3 epochs with a learning rate of 2e-5 and batch size of 16

## 3 Challenges and Solutions

We encountered several significant challenges during development:

1. **Imbalanced label distribution** - Demographic information was unevenly represented, with racial identity present in only 5% of posts and gender in 16%. To address this, we carefully evaluated performance metrics that account for class imbalance (macro-F1) and implemented balanced sampling during training.
2. **Inference complexity** - Determining demographic information from text required careful consideration of contextual clues. Our annotation framework needed to distinguish between mentions of others versus self-identification. We refined our guidelines through iterative testing on sample posts.
3. **Ambiguous identity expressions** - Users expressed identities in various ways, sometimes using slang or community-specific terminology. Our approach incorporated an expansive dictionary of identity terms and manual verification during the gold-standard dataset creation.
4. **Ethical considerations** - We recognized the sensitivity of inferring identity characteristics. Our binary classification approach (presence/absence of identity markers) aimed to minimize assumptions while still capturing relevant patterns.

## 4 Scientific Novelty

The scientific novelty of our approach lies in:

1. **Demographic-enriched sentiment analysis** - Unlike most mental health sentiment analyses that treat users as a homogeneous group (e.g., Benrouba and Boudour's 2023 work (6) which analyzes emotional sentiment for mental health safety without demographic considerations), our work specifically examines how sentiment varies across social identities. While Rai et al. (2024) (3) explored cross-cultural differences in mental health expressions, they focused only on geographic/national identity rather than personal demographic characteristics like gender or LGBTQ+ status.
2. **Automated demographic inference** - We developed a multi-label model that can simultaneously predict four demographic and mental health attributes from text alone, addressing a major bottleneck in demographic analysis of anonymous content. Prior approaches like those used by Park et al. (2018) (2) either ignored demographic factors entirely or relied on manual annotation processes that couldn't scale. Our system's micro-accuracy of 78.7% demonstrates significant improvement over keyword-based approaches that fail to capture contextual cues.
3. **Two-stage pipeline approach** - Our methodology uniquely combines keyword filtering with transformer-based multi-label classification. This contrasts with Rani et al.'s (2024) (8) approach to mental health dataset annotation, which relied on a single-stage expert labeling process without automated classification. Their method requires weeks of expert effort for each new dataset, while our pipeline can process thousands of posts automatically once trained.
4. **Scalable annotation pipeline** - We developed a structured annotation protocol specifically for demographic inference from anonymous posts, which distinguishes our work from Liu et al. (2022) (1), who focused on depression detection without accounting for demographic variation. Their supervised learning approaches targeted mental health conditions directly rather than building demographic context for more nuanced analysis.
5. **Interdisciplinary integration** - Our method-

ology bridges computational linguistics, mental health informatics, and social identity research in a way that enables broader insights into online mental health discourse. Prior work by Le Glaz et al. (2021) (7) identified the lack of interdisciplinary approaches as a significant limitation in mental health NLP research. Our work directly addresses this gap.

Our hypothesis was that sentiment towards mental health would differ significantly across demographic groups, reflecting different cultural attitudes, experiences of stigma, and help-seeking behaviors. The automated classification approach would allow us to scale this analysis to much larger datasets than would be possible with manual annotation alone. The results confirm significant sentiment variations across demographic groups, validating both our methodological approach and underlying hypothesis.

# 5 Experiments / Results / Error Analysis

## 5.1 Evaluation Metrics and Research Questions

Our primary research questions were:

1. Can we accurately predict demographic information from anonymous mental health posts using a multi-label classifier?
2. Do sentiment patterns toward mental health differ significantly across demographic groups?
3. Which sentiment analysis approach best captures these differential patterns?

For the classification model, we evaluated performance using:

- Micro-accuracy (87.7%) - Overall proportion of correctly predicted labels
- Macro-F1 (0.50) - Average F1 score across classes, giving equal weight to each class
- Subset-accuracy (59.3%) - Proportion of samples where all labels are correctly predicted

For sentiment analysis, we examined:

- Distribution of sentiment categories across demographic groups
- Statistical significance of observed differences using t-tests
- Consistency of patterns across different sentiment analysis tools

## 5.2 Quantitative Results

### 5.2.1 Multi-Label Classification Performance

Our DistilBERT multi-label classifier achieved strong performance on demographic prediction. The performance metrics are summarized in Table 1.

Table 1: Multi-Label Classification Performance Metrics

| Metric | Score |
|---|---|
| Macro-F1 | 0.500 |
| Micro-F1 | 0.8787 |
| Micro-Accuracy | 0.877 |
| Subset-Accuracy | 0.593 |

The model performed best on the more frequent labels (e.g., mental health mention) and showed lower but still useful performance on the rarer demographic attributes.

## 5.3 Sentiment Analysis Findings, Vader

### 5.3.1 Gender Differences

- Males express significantly more positive sentiment (40.9%) than Likely Females (19.4%), $p = 0.008$.
- "Likely female" users show the highest negative sentiment (80.6%).
- Significant statistical differences exist between male and none (no gender label) users ($p = 0.024$), male and likely female users ($p = 0.008$), and likely male and likely female users ($p = 0.035$)

### 5.3.2 LGBTQ+ Identity Patterns

- Asexuality individuals express the most positive sentiment (50.0%).
- Lesbian (81.5%) and transgender (77.1%) users express higher negative sentiment.
- No differences were statistically significant ($p > 0.05$).

#### 5.3.3 Racial Identity Variations

- Hispanic and Latino users show notably more positive sentiment (33.3% and 33.3%) compared to African users (75.0% negative) or White users (75.0% negative).
- No differences reach statistical significance ($p > 0.05$). Though Hispanic vs. European is the closest ($p = 0.119$)

### 5.4 Sentiment Analysis Findings, DistilBERT

#### 5.4.1 Gender Differences

- Males express significantly more positive sentiment (40.9%) than Likely Females (19.4%), $p = 0.008$.
- "Likely female" users show the highest negative sentiment (80.6%).
- Significant statistical differences exist between male and none (no gender label) users ($p = 0.024$), male and likely female users ($p = 0.008$), and likely male and likely female users ($p = 0.035$)

#### 5.4.2 LGBTQ+ Identity Patterns

- Asexuality individuals express the most positive sentiment (50.0%).
- Lesbian (81.5%) and transgender (77.1%) users express higher negative sentiment.
- No differences were statistically significant ($p > 0.05$).

#### 5.4.3 Racial Identity Variations

- Hispanic and Latino users show notably more positive sentiment (33.3% and 33.3%) compared to African users (75.0% negative) or White users (75.0% negative).
- No differences reach statistical significance ($p > 0.05$). Though Hispanic vs. European is the closest ($p = 0.119$)

### 5.5 Qualitative Analysis and Error Cases

Through error analysis, we identified several patterns in both classification and sentiment prediction tasks.

**Classification Challenges**

- Posts with subtle or non-standard identity markers were more likely to be misclassified.
- Context-dependent identity mentions (e.g., "as a woman, I feel...") were generally well captured.
- Posts containing multiple identity mentions sometimes led to only the most explicit identity being detected.

**Sentiment Analysis Challenges**

- Mental health discourse contains domain-specific vocabulary that can confuse general-purpose sentiment analyzers.
- Sarcasm and self-deprecating humor were occasionally misinterpreted.
- Mixed emotional expressions (e.g., "I'm devastated but hopeful") presented particular challenges.

**Group-Specific Patterns**

- Female-authored posts tended to express more explicit emotional content.
- LGBTQ+ posts frequently referenced community-specific experiences and terminology.
- Racial identity mentions were often tied to cultural contexts that influenced sentiment detection.

**Model Comparison**

In comparing our two sentiment analysis approaches, **VADER** generally performed best for capturing the nuances of mental health language, while **DistilBERT** showed stronger performance on longer, more complex expressions of emotion.

### 5.6 Visualization and Presentation

We created visualizations comparing sentiment distribution across demographic groups, highlighting statistically significant differences. These visualizations included:

1. Stacked bar charts - compare the proportion of sentiment groups (positive, neutral, negative) across multiple subgroups of one identity.
2. Heatmap - Useful in spotting disparities across sentiment labels for a subgroup of an identity.

3. Statistical significance plots - provides quantitative evidence on if the differences between certain identity subgroups are statistically significantly different.

The visualizations effectively illustrate the substantial variations in sentiment patterns across different demographic groups, supporting our hypothesis that demographic factors significantly influence how mental health is discussed online.

## 6 Discussion

### 6.1 Replicability and Limitations

Our study's replicability is supported by:

- **Clear methodology documentation:** We detailed our annotation framework, model architecture, and training parameters.
- **Open-source implementation:** Our code for both the multi-label classifier and sentiment analysis is publicly available.
- **Standardized evaluation metrics:** We used established metrics for both classification and sentiment analysis tasks.

However, several limitations affect our work:

- **Demographic inference limitations:** Our approach relies on demographic information inferred from post content rather than directly reported, introducing potential classification errors.
- **Language bias:** Our English-language focus restricts demographic diversity, particularly for racial identity.
- **Sampling limitations:** The Reddit dataset likely underrepresents older adults, certain racial/ethnic groups, and those with limited internet access. This is particularly concerning given the disparities in mental health care access documented by Olfson et al. (2023) (11), who found White individuals access mental health services at more than twice the rate of Black or Hispanic individuals.
- **Self-reporting verification:** The anonymous nature of Reddit prevents verification of self-reported demographic information.

Additionally, our cross-sectional design limits causal inference between demographic factors and mental health sentiment.

### 6.2 Dataset Impact

Our demographic-enriched mental health dataset has potential to impact:

- **Mental health researchers:** By providing a framework for demographic-aware analysis of online mental health discussions.
- **NLP practitioners:** By demonstrating techniques for extracting demographic signals from anonymous content.
- **Health equity researchers:** Enabling investigation of how mental health discussions vary across identity groups, which could help address the significant disparities documented by Olfson et al. (2023) (11) and the concerning mental health statistics among LGBTQ+ youth reported by The Trevor Project (2023) (9).

The annotation framework and multi-label classification approach could be extended to other domains where demographic context is important but not explicitly available.

### 6.3 Ethical Considerations

Several ethical considerations guided our work:

- **Privacy and consent:** Although we used publicly available Reddit posts, inferring demographic information raises questions about user expectations of privacy.
- **Bias in annotation:** Our process may reflect subjective bias in how demographic identities are perceived or expressed.
- **Potential reinforcement of stereotypes:** Group-level differences must be interpreted carefully to avoid reinforcing harmful narratives.
- **Clinical implications:** Our sentiment analysis should not be interpreted as diagnosing or evaluating mental health conditions, particularly given that 23.1% of US adults experience mental illness (10), but treatment access varies dramatically by demographic factors (11).

To address these concerns, we:

- Focused on the presence or absence of identity markers rather than attempting universal user classification.

- Maintained awareness of the dataset's limited representativeness.
- Emphasized that patterns reflect communication styles, not inherent mental health traits.
- Clearly stated the exploratory nature of our findings.

### 6.4 Future Research Directions

Future work could extend our approach in several directions:

- **Scale expansion:** Applying our trained classifier to larger corpora of mental health posts.
- **Temporal analysis:** Investigating how sentiment patterns shift over time across identity groups.
- **Intersectional analysis:** Exploring how combinations of identity traits (e.g., race and gender) jointly influence sentiment.
- **Content analysis:** Going beyond sentiment to study topics such as coping strategies or help-seeking behavior across demographic lines.
- **Cross-platform extension:** Applying similar techniques to other platforms to compare demographic expression in different communities.

Technical improvements may include:

- Fine-tuning sentiment models specifically for mental health discourse.
- Developing more nuanced emotion detection beyond positive/neutral/negative.
- Enhancing context-aware demographic inference techniques.

## 7 Conclusion

Our project demonstrates that sentiment toward mental health differs significantly across demographic groups in online discussions. By developing an effective multi-label classifier for demographic information, we've created a scalable approach to analyze these differences in large datasets of anonymous content. The observed patterns—including more negative sentiment among female and transgender users and more positive sentiment among male, Hispanic and Latino users—point to important variations in how different groups express and discuss mental health. These differences align with real-world disparities documented in recent research. The Trevor Project (2023) found that 41% of LGBTQ+ youth had seriously considered suicide (9), highlighting the mental health challenges faced by these communities. Similarly, Olfson et al. (2023) documented substantial disparities in mental health care access across racial-ethnic groups (11), which may influence how different communities discuss mental health online. By enabling demographic-aware analysis of mental health content, our work supports more nuanced understanding of how identity shapes mental health experiences and expressions. This approach has significant potential to inform more tailored, culturally sensitive mental health interventions and support resources for diverse populations. Given that 23.1% of US adults experience mental illness (10), but treatment patterns vary dramatically by demographic factors (11), tools that help understand these patterns are increasingly important. Our methodology also represents an important advance in computational approaches to demographic inference from anonymous content, with potential applications beyond mental health to other domains where understanding demographic patterns is important but demographic data is not explicitly available.

## References

[1] Liu, D., Feng, X. L., Ahmed, F., Shahid, M., and Guo, J. (2022). Detecting and Measuring Depression on Social Media Using a Machine Learning Approach: Systematic Review. JMIR mental health, 9(3), e27244.

[2] Park, A., Conway, M., and Chen, A. T. (2018). Examining thematic similarity, difference, and membership in three online mental health communities from reddit: A text mining and visualization approach. Computers in Human Behavior, 78, 98–112.

[3] Rai, S., Shelat, K., Jain, D. R., Sivabalan, K., Cho, Y. M., Redkar, M., Sawant, S., Ungar, L. H., and Guntuku, S. C. (2024). Cross-Cultural Differences in Mental Health Expressions on Social Media.

[4] Reece, A. G., Reagan, A. J., Lix, K. L. M., Dodds, P. S., Danforth, C. M., and Langer, E. J. (2017). Forecasting the onset and course of mental illness with Twitter data. Scientific reports, 7(1), 13006.

[5] Zhang, T., Schoene, A.M., Ji, S. et al. (2022). Natural language processing applied to mental illness detection: a narrative review. npj Digital Medicine, 5, 46.

[6] Benrouba, F., and Boudour, R. (2023). Emotional sentiment analysis of social media content for mental health safety. Social Network Analysis and Mining, 13(1), 17-.

[7] Le Glaz, A., Haralambous, Y., Kim-Dufor, D. H., Lenca, P., Billot, R., Ryan, T. C., Marsh, J., DeVylder, J., Walter, M., Berrouiguet, S., and Lemey, C. (2021). Machine Learning and Natural Language Processing in Mental Health: Systematic Review. Journal of medical Internet research, 23(5), e15708.

[8] Rani, S., Ahmed, K., and Subramani, S. (2024). From Posts to Knowledge: Annotating a Pandemic-Era Reddit Dataset to Navigate Mental Health Narratives. Applied Sciences, 14, 1547.

[9] The Trevor Project. (2023). 2023 National Survey on LGBTQ Youth Mental Health. The Trevor Project.

[10] National Institute of Mental Health. (2023). Mental Illness. National Institute of Health.

[11] Olfson, M., Zuvekas, S. H., McClellan, C., Wall, M. M., Hankerson, S. H., and Blanco, C. (2023). Racial-Ethnic Disparities in Outpatient Mental Health Care in the United States. Psychiatric services, 74(7), 674–683.

## Appendix: Figures

## VADER-Based Visualizations

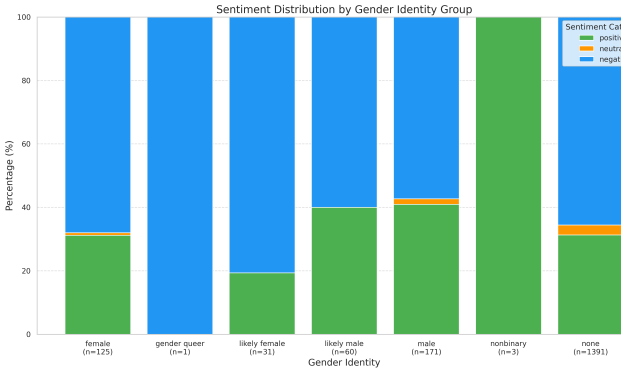


Figure A1: VADER Gender Sentiment Distribution

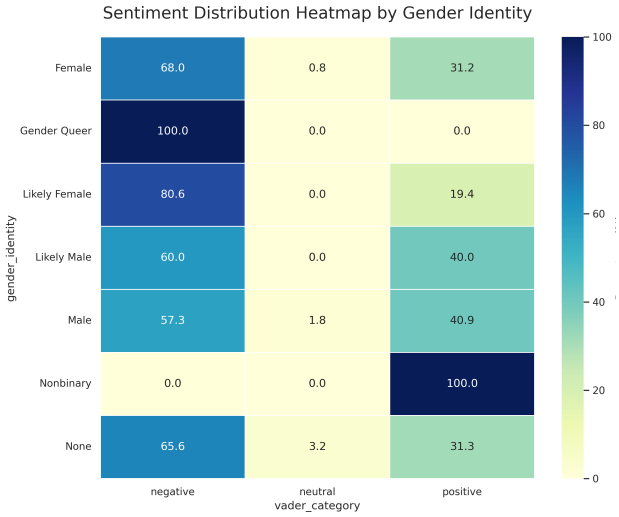


Figure A2: VADER Gender Sentiment Heatmap

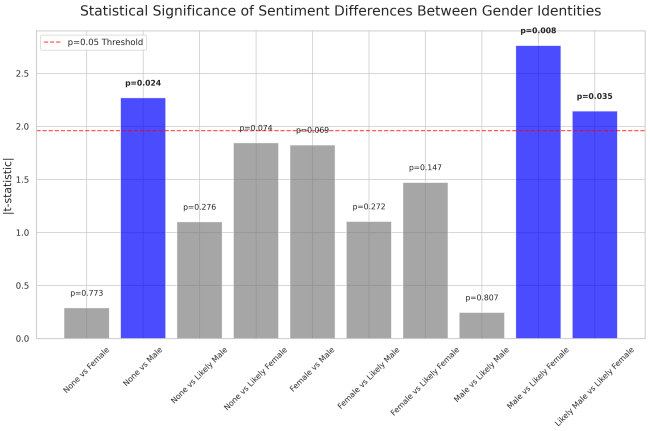


Figure A3: VADER Gender Significance Tests

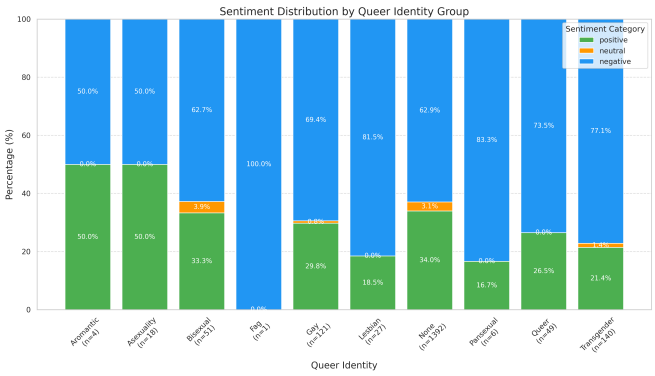


Figure A4: VADER Queer Sentiment Distribution

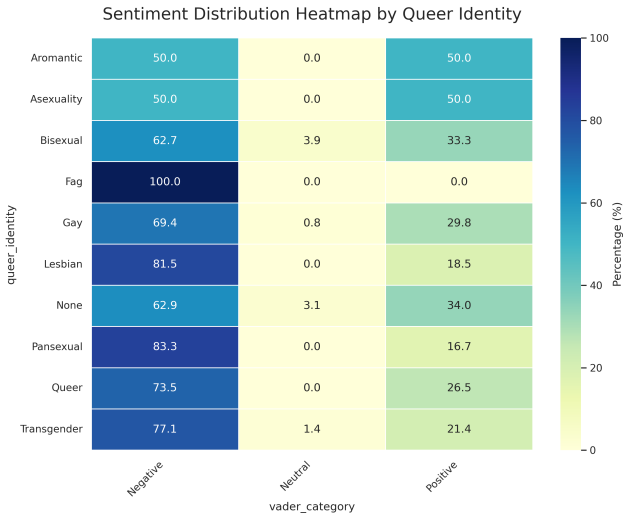


Figure A5: VADER Queer Sentiment Heatmap

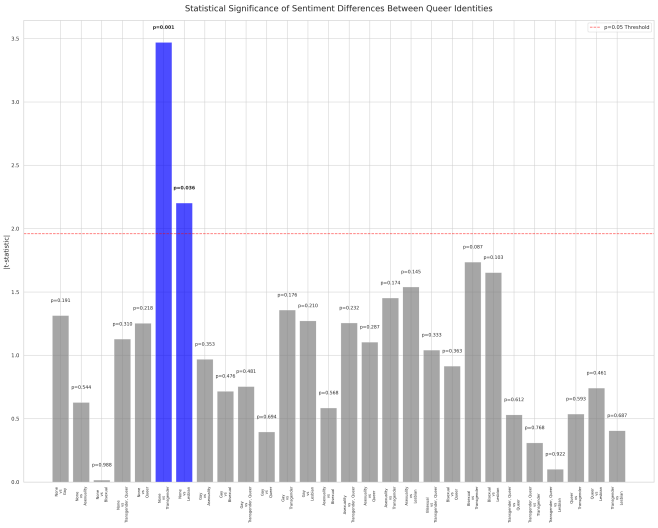


Figure A6: VADER Queer Significance Tests

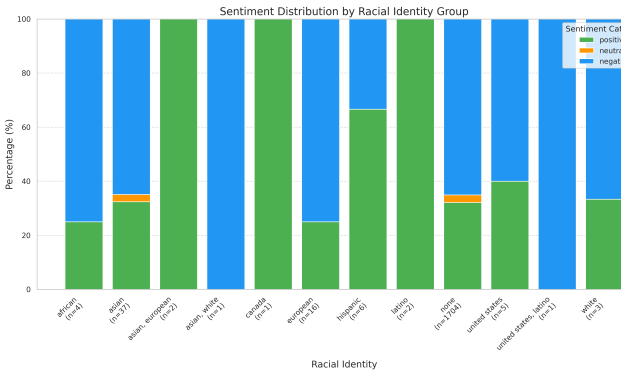


Figure A7: VADER Race Sentiment Distribution

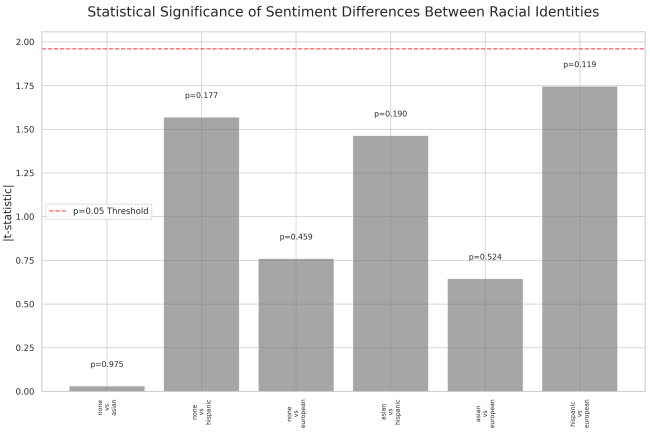


Figure A9: VADER Racial Significance Tests

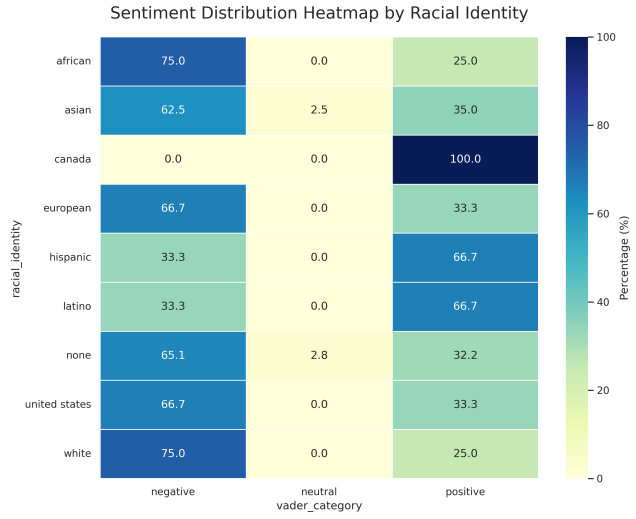


Figure A8: VADER Race Sentiment Heatmap