‘Stonewall was a Riot’: Hacking History Project

By Aparna Uma – University of Leeds, Sophie Russell – Coventry University, Linda Marchant – Coventry University and Phillips O Bature – Coventry University

TABLE OF CONTENTS

Source above: "Stonewall was a shot heard 'round the world and a new beginning in the march toward lesbian/gay liberation" *Stonewall*, June 26, 1969-July 5, 1994 and undated. (June 26, 1969-July 5, 1994; n.d.). MS Lesbian Herstory Archives: Subject Files: Part 6: Spinsters-Youth Folder No.: 13900. Lesbian Herstory Archives. *https://link.gale.com/apps/doc/EMGDBN775730883/DSLAB?u=coventryhh&sid=bookmark-DSLAB&xid=57b6665f&pg=1*

INTRODUCTION

This project explores how violence associated with the Stonewall Riots has been described, framed, and remembered across time, actors, and publication contexts. Using computational text analysis tools in the *Gale Digital Scholar Lab*, we examined both contemporaneous accounts of Stonewall and later retrospective narratives to understand how language, tone, and attribution of responsibility change over time.

Rather than seeking definitive answers, this project asks what computational methods can reveal about patterns in discourse—and what their limits are—when applied to historically and politically complex events like Stonewall.

RESEARCH QUESTIONS

This project examines how the Stonewall Riots and associated acts of violence have been represented across time, actors, and publication contexts. Drawing on both the prebuilt and post-1972 content sets, we focus on the following research questions:

First, how does the tone and language used to describe violence during the Stonewall Riots vary according to the identities of the actors involved, including protesters, police, and institutional authorities?

Second, how do different publications frame and narrate violence associated with Stonewall, and what rhetorical or lexical differences emerge across activist, community-based, and mainstream media sources?

Finally, how do these narratives change over time, particularly as Stonewall shifts from a contemporaneous event to a commemorated historical moment?

TOOL-INFORMED RESEARCH DESIGN

Our research questions were deliberately shaped by the analytical tools available within the *Gale Digital Scholar Lab*, ensuring alignment between interpretive goals and computational methods.

- **Document Clustering:**

  - Used to identify groups of texts with similar linguistic and rhetorical features
  - Supported questions about variation across publications, genres, and historical moments
  - Enabled the designation of cluster names that reflect dominant narrative frames

- **Named Entity Recognition (NER):**
  - Informed analysis of how violence is attributed to specific actors
  - Enabled comparison of references to protesters, police, institutions, and organizations across datasets
- **Ngrams:**
  - Used to surface recurring phrases and patterns in language
  - Supported close attention to how violence is described through repeated lexical constructions
- **Parts of Speech Analysis:**
  - Enabled examination of verbs, adjectives, and modifiers associated with violence
  - Supported analysis of agency, action, and evaluative language
- **Sentiment Analysis:**
  - Provided a comparative measure of tone across publications and time periods
  - Supported longitudinal analysis of affective shifts in narratives of Stonewall
- **Topic Modelling:**
  - Used to identify dominant themes within and across datasets
  - Supported analysis of how narratives evolve as Stonewall shifts from event to commemoration

Together, these tools informed the design of research questions that were computationally tractable while remaining attentive to historical context and critical interpretation.

DATA CURATION

THE INITIAL DATASET

Our project began with the prebuilt Stonewall content set provided in the *Gale Digital Scholar Lab*. The dataset is composed of the following characteristics:

- **Total documents:** 142
- **Source:** Gale primary source collections
- **Chronological coverage:** 1911–1972
- **Authorship:**
    - 14 identified authors
    - Includes both individually authored and collaboratively produced texts
- **Archives represented (5):**
    - *Archives of Sexuality and Gender* (119 documents)
    - *Political Extremism and Radicalism* (11 documents)
    - *Women's Studies Archive* (2 documents)
    - *Archives Unbound* (10 documents)
- **Source libraries (11):**
    - Community-based archives, including the Gay, Lesbian, Bisexual, and Transgender Historical Society and the Lesbian Herstory Archives
    - Major institutional collections, including the New York Public Library
- **Document types:**
    - 37 distinct document types
    - Includes reports, essays, newsletters, articles, letters, newspapers, organizational records, and ephemera such as announcements and advertisements

INITIAL REFLECTIONS ON THE DATASET

While the Stonewall content set provided a strong and thematically focused foundation for our hackathon work, initial exploration revealed a notable **temporal imbalance**. Although the dataset extends to 1972, the majority of documents were produced **prior to the Stonewall uprising in 1969**, with comparatively few materials reflecting the immediate post-Stonewall period.

This imbalance shaped our decision to build a secondary dataset focusing on post-Stonewall materials extending into the later twentieth century. The initial corpus is particularly well suited for examining the *conditions, discourses, and organizing practices that preceded Stonewall*, but it offers more limited insight into how language, priorities, and strategies shifted in its aftermath. As a result, Stonewall emerges less as a sharp rupture and more as a historical hinge whose consequences are only partially visible within this dataset.

Authorship within the Stonewall content set reflects the collective nature of movement

work, with 14 identified authors and several collaboratively produced texts. Drawn from 5 Gale archives and 11 source libraries (including both community-based and institutional collections), the dataset foregrounds activist perspectives while also reflecting the priorities and constraints of archival collection practices. As a result, certain voices, genres, and historical moments are more visible than others.

The corpus is also highly heterogeneous in form, encompassing 38 document types such as reports, newsletters, letters, newspapers, and organizational ephemera. While this diversity enriches interpretive possibilities, it introduces methodological challenges, particularly for computational analysis sensitive to genre, length, and purpose. This became a particular hurdle during our data analysis.

SECONDARY DATASET

In response to the limitations identified in the prebuilt Stonewall dataset, we constructed a secondary content set manually using the *Gale Digital Scholar Lab*. Titled *New Stonewall Riots Post 1972*, this expanded corpus was designed to foreground post-Stonewall activism and discourse.

Its key characteristics are as follows:

- **Total documents:** 753
- **Source:** Gale documents
- **Chronological coverage:** 1973–2020
- **Authorship:**
    - 293 identified authors
    - Reflects a broader and more diverse range of voices in the post-Stonewall period
- **Archives represented (2):**
    - *Archives of Sexuality and Gender* (731 documents)
    - *The Independent Historical Archive* (22 documents)
- **Source libraries (8):**
    - Lesbian Herstory Archives (427 documents)
    - Lambda Archives of San Diego (121 documents)
    - Gay, Lesbian, Bisexual, and Transgender Historical Society (80 documents)
    - The ArQuives: Canada's LGBTQ2+ Archives (57 documents)
    - Duke University Library (32 documents)
    - Independent Digital News and Media Limited (22 documents)
    - Cornell University Libraries (10 documents)
    - ONE National Gay & Lesbian Archives (4 documents)

SEARCH STRATEGY AND DATASET CONSTRUCTION

The secondary content sets were constructed through an iterative search process using the *Gale Digital Scholar Lab*, informed by both the limitations of the prebuilt Stonewall dataset and emerging questions about post-Stonewall representation over time.

Core Search Parameters

To build the initial post-1972 dataset, we applied the following criteria:

- **Search terms:** "Stonewall" and "riot*"
- **Date range:** After 1972
- **Language:** English

This approach was designed to capture references to Stonewall not only as a historical event but also as an evolving symbol within activist, journalistic, and commemorative discourse.

Anniversary-Focused Searches

To examine how Stonewall is retrospectively framed, we constructed an additional dataset focused on commemorative moments:

- **Search terms:** "Stonewall" and "anniversary"
- **Date range:** After 1972
- **Geographic focus:** UK and US publications

Restricting this dataset to UK and US sources allowed for comparative analysis of how Stonewall's legacy is mobilized across national contexts with shared but distinct LGBTQ+ histories.

Geographic Restrictions and Replication Challenges

We also attempted to build post-1972 datasets restricted exclusively to UK and US publications in order to maintain comparability across corpora. In some cases, this required duplicating existing datasets and manually refining them. However, for the original prebuilt Gale Stonewall dataset, replication proved challenging: because the original search terms and parameters were not fully visible, it was not possible to reconstruct the dataset precisely and then remove publications from other geographic regions. This limitation shaped the scope of our comparative analysis and underscored the importance of transparent search documentation.

Iteration, Revision, and Dataset Expansion

Throughout the process, we revisited and updated datasets as methodological issues emerged. In one instance, we realized that only the first 100 results of a search had been saved, requiring us to rerun and expand the query. In later stages, we introduced an additional search parameter – "violence" – and extended the timeframe to post-2020, which yielded a further 100 documents. This refinement enabled us to trace how

Stonewall is invoked in relation to contemporary protest, state power, and public memory.

INITIAL REFLECTIONS ON THE SECONDARY DATASET

Dataset construction did not conclude with initial search and selection but continued as an iterative, reflective process shaped by early analytical results. Preliminary analysis – particularly outputs from named entity recognition (NER) and keyword extraction – revealed recurring terms, organizations, locations, and individuals that were either underrepresented or unevenly distributed across the datasets.

These early findings prompted targeted refinement of both content sets. Keywords and entities surfaced through NER were used to reassess search coverage and identify gaps, leading to the expansion or adjustment of search terms and temporal boundaries. For example, entities associated with later protest movements, commemorative events, or institutional actors suggested additional avenues for post-Stonewall material that were not fully captured by the initial "Stonewall" and "riot*" queries alone.

In contrast to the original Stonewall content set, the post-1972 corpus significantly shifts both scale and structure. The larger number of documents and authors supports analysis of longer-term trends, recurring debates, and evolving organizational strategies, while also introducing new methodological considerations related to corpus size and variability.

The expanded temporal coverage foregrounds the aftermath of Stonewall as an ongoing process rather than a discrete event, making it possible to examine how activist language, priorities, and institutional forms changed across multiple generations. At the same time, the continued reliance on community-based archives maintains a focus on activist-produced materials, reinforcing continuity with the original dataset while addressing its temporal limitations. Together, these two content sets enable comparative analysis across pre- and post-Stonewall periods and provide a more comprehensive foundation for interpreting change over time.

At the same time, the size and diversity of the corpus complicate critical analysis. The increased number of authors, document types, and historical contexts introduces greater variability, making it more difficult to isolate causal relationships or attribute changes to specific events. As a result, findings drawn from this dataset are highly limited by the time we had to complete this project and warrants further research.

DATA CLEANING AND TEXT PREPARATION

Textual cleaning was approached as an iterative process, evolving alongside corpus construction and early analytical results within the *Gale Digital Scholar Lab*. Rather than relying on a single default configuration, we adjusted cleaning parameters to support different analytical goals.

In our initial analysis, we used the platform's default cleaning configuration while retaining

numeric characters in order to capture measurement-related data and assess baseline outputs. This approach provided a preliminary sense of term frequency and structure but also revealed substantial noise that limited interpretive clarity.

To refine the results, we developed a second, more restrictive cleaning configuration for n-gram analysis. This configuration removed ASCII characters, stop words, punctuation, numbers, and special characters, and deprioritized high-frequency labels such as "gay," which appeared so frequently that they obscured more analytically meaningful patterns. These adjustments improved signal clarity, allowing recurring phrases and thematic variations to emerge more distinctly while reinforcing the importance of aligning cleaning decisions with specific analytical tasks.

INTERPRETIVE IMPACT OF DATA CLEANING CHOICES

Text cleaning decisions had a direct impact on how patterns were surfaced and interpreted. Removing high-frequency terms, punctuation, and numerical noise improved the visibility of recurring phrases and thematic structures, but also reduced certain contextual markers that may carry historical or rhetorical significance. In particular, deprioritizing commonly occurring identity labels shifted analytical focus away from frequency and toward variation in language use. As a result, findings reflect interpretive choices embedded in the cleaning process, reinforcing the need to read computational outputs as situated analyses rather than neutral representations of the underlying texts.

DATA ANALYSIS

For the purposes of time, I am highlighting the most developed and promising analytical tools.

NGRAMS

The *Gale Digital Scholar Lab*'s Ngram tool was one of the first tools we used to analyse our content set. We used the Ngram tool to refine our cleaning configuration, as it allowed us to identify common OCR errors in our content set. The Ngram results from our initial content set also helped us to determine the criteria for the second iteration of our textual content set, as the term frequency results were dominated by words completely unrelated to our topic.

The most common terms are broad identity and reference markers – such as "gay," "people," "women," and place-based terms like "New" – indicating that much of the dataset is oriented toward community identity, collective description, and spatial framing rather than singular events.

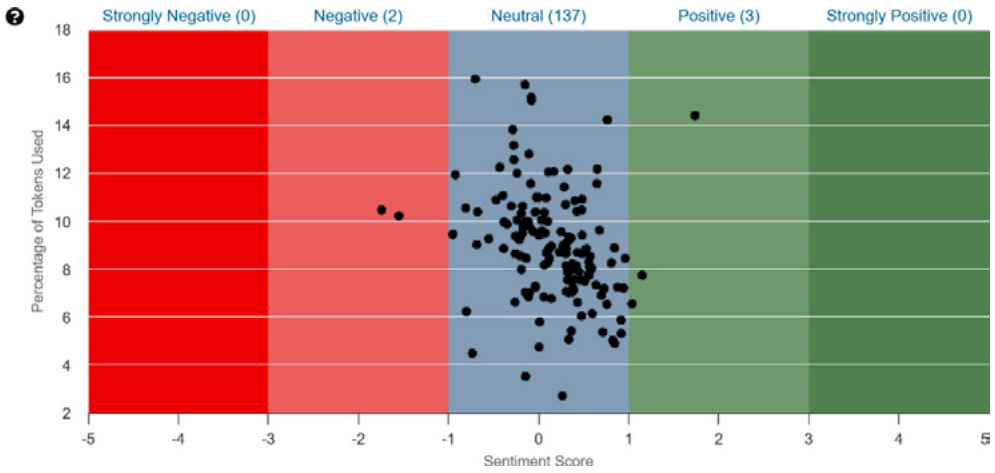
Terms directly associated with violence, including "police," "violence," and "arrested," occur far less frequently but remain significant. Their relative scarcity suggests that explicit descriptions of violence are episodic rather than pervasive, often embedded within broader discussions of rights, identity, and organizing. This pattern supports our earlier observations that violence is frequently framed contextually – as a catalyst or background condition – rather than as the primary narrative focus.

The dominance of high-frequency identity labels also underscores the interpretive importance of data cleaning choices. By deprioritizing terms such as "gay" in later iterations, we were able to surface more contextually specific phrases and action-oriented language, enabling closer attention to how violence is described, attributed, and evaluated across different publications and time periods.

SENTIMENT ANALYSIS

SENTIMENT ANALYSIS: INITIAL STONEWALL CONTENT SET

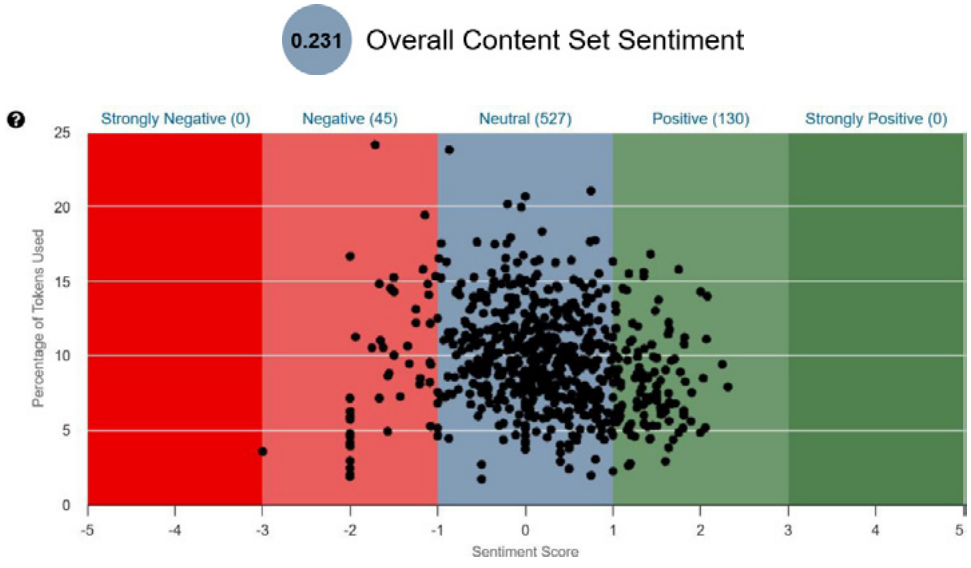
Sentiment analysis was applied to the prebuilt Stonewall content set to identify broad tonal patterns in contemporaneous and near-contemporaneous accounts of the riots and associated violence. Sentiment scores are treated as indicative signals of rhetorical framing rather than direct measures of emotional intent.



Overall, sentiment values cluster around **neutral to mildly negative**, reflecting the dominance of informational and organizational genres such as reports, letters, and newsletters. Strongly positive or negative sentiment is rare, suggesting that violence is typically described in measured, procedural terms rather than through overtly emotive language. When more negative sentiment appears, it is most often associated with references to policing, arrests, and institutional authority, conveying harm or repression through evaluative but restrained language. These patterns indicate that Stonewall is generally framed as part of broader discussions of rights, organizing, and community response rather than as an isolated violent episode.

SENTIMENT ANALYSIS: POST-1972 CONTENT SET

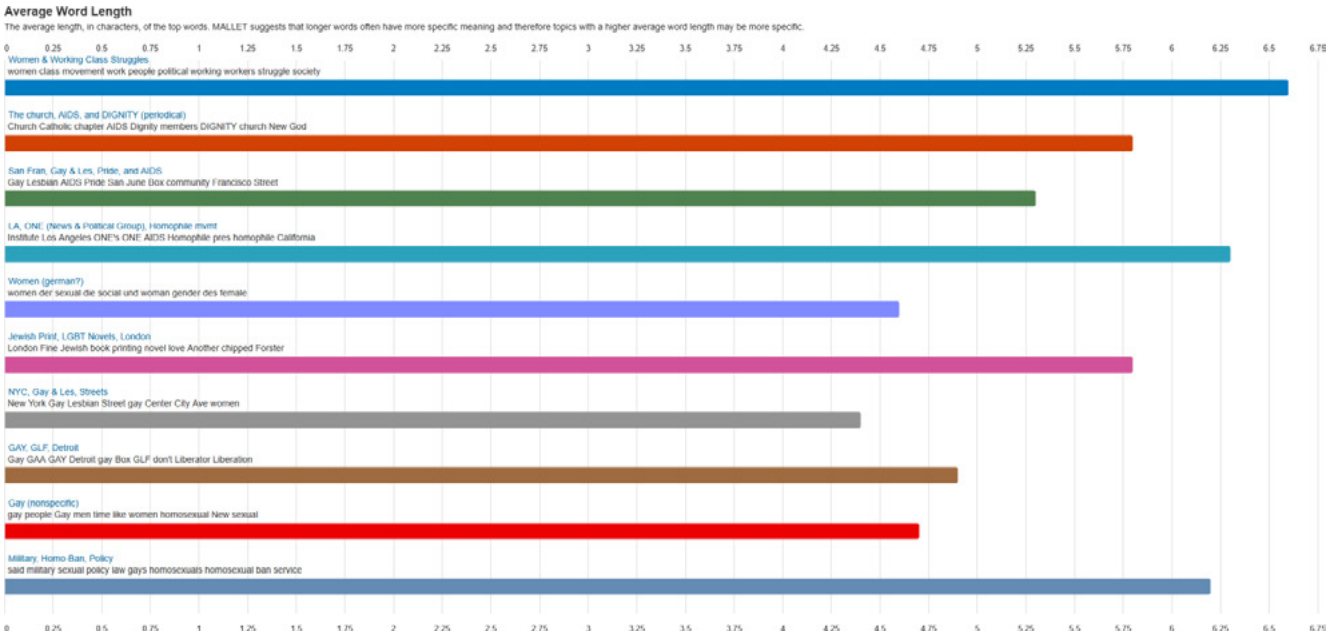
Sentiment analysis of the post-1972 content set reveals a **wider tonal range** than in the initial dataset, with scores extending further into both positive and negative values. This shift reflects changes in genre, publication context, and temporal distance from the events of 1969.



Negative sentiment appears more frequently in retrospective discussions of violence, policing, and state authority, particularly in journalistic accounts, anniversary coverage, and reflective essays. These texts are more likely to employ explicitly evaluative and moral language, suggesting that distance from the event enables stronger critique. At the same time, the dataset also includes more positive sentiment, often linked to narratives of pride, resilience, and progress that frame Stonewall as a foundational moment within a longer history of activism. Together, these patterns support a longitudinal reading in which Stonewall discourse evolves from restrained documentation toward more emotionally and politically expressive modes of remembrance.

TOPIC MODELLING

Topic modelling was used to surface recurring thematic structures across the corpus and to examine how narratives of Stonewall, violence, and LGBTQ+ activism intersect with broader social, political, and cultural concerns. Topics were interpreted qualitatively based on their dominant terms and document associations, and are understood as probabilistic clusters rather than fixed or mutually exclusive categories.



Several topics foreground **identity, community, and movement formation**. Topics labelled *Gay (nonspecific)*, *NYC, Gay & Les, Streets*, and *GAY, GLF, Detroit* emphasize localized activist geographies and community-based organizing. These topics situate Stonewall within urban spaces and grassroots movements, often framing violence as part of lived experience in public space rather than as a singular historical incident.

Other topics highlight **organizational and institutional contexts**. The *LA, ONE (News & Political Group), Homophile mvmt* topic reflects earlier homophile-era activism and advocacy organizations, while *Military, Homo-Ban, Policy* centres institutional regulation, exclusion, and state power. In these topics, references to violence are more likely to be mediated through policy language, discipline, and bureaucratic enforcement rather than physical confrontation.

Health, religion, and moral discourse emerge strongly in the topic labelled *The church, AIDS, and DIGNITY (periodical)*, which reflects post-Stonewall activism shaped by the AIDS crisis and faith-based responses. Violence in this context is often indirect, framed through neglect, stigma, and institutional harm rather than overt policing.

Several topics foreground **gendered and intersectional perspectives**, including *Women & Working Class Struggles* and *Women (German?)*. These topics suggest that Stonewall-related discourse intersects with feminist, labour, and transnational concerns, complicating narratives that centre exclusively on gay male activism. Violence here appears as both structural and experiential, embedded in class struggle and gendered exclusion.

Finally, topics such as *San Fran, Gay & Les, Pride, and AIDS* and *Jewish Print, LGBT Novels, London* reflect **cultural production and media representation**, including pride events, literary discourse, and international publishing contexts. These topics indicate how Stonewall is rearticulated through commemorative, cultural, and diasporic frames over time.

Taken together, the topic model reveals that Stonewall-related discourse is **diffuse and multi-sited**, with violence appearing unevenly across thematic contexts. Rather than forming a dominant standalone topic, violence is embedded within narratives of identity, policy, health, culture, and memory. This reinforces our broader finding that Stonewall functions less as a singular event in the corpus and more as a recurring reference point through which changing political, social, and cultural concerns are articulated over time.

LIMITATIONS OF TOPIC MODELLING

Topic modelling provides a useful overview of dominant thematic patterns but has important interpretive limitations. Topics are probabilistic groupings based on word co-occurrence rather than stable or conceptually discrete categories, and their labels reflect researcher interpretation rather than intrinsic meaning. As a result, topics may combine multiple rhetorical frames or historical moments within a single grouping.

Additionally, the presence or absence of violence-related language within a topic does not indicate its relative importance, as such references may be diffuse, implicit, or embedded within broader discussions of identity, policy, or culture. Genre diversity and corpus scale further shape topic coherence, particularly in activist and organizational texts that blend personal, political, and institutional language. For these reasons, topic modelling results were interpreted as exploratory signals and were triangulated with clustering, n-gram analysis, and close reading rather than treated as definitive representations of the corpus.

FINAL REFLECTIONS AND METHODOLOGICAL CONSIDERATIONS

Several broader methodological considerations shaped both the scope and interpretation of this project. Issues of overrepresentation emerged across datasets, particularly where certain archives, genres, or recurring terms dominated results and risked obscuring less frequent but potentially significant perspectives. These imbalances reflect both archival survival and search-driven collection practices rather than the full historical record.

Data quality also influenced analysis. OCR inconsistencies, especially in older documents, affected entity recognition, keyword extraction, and sentiment scoring, introducing noise that could not be fully corrected within the timeframe of the hackathon. Time constraints further limited opportunities for deeper corpus cleaning, extensive validation, or sustained close reading, reinforcing the exploratory nature of our findings.

These constraints raise essential interpretive questions: What can we trust, and what are the results actually telling us? Computational outputs reveal patterns of similarity, frequency, and association, but they do not explain causation or historical meaning on their own. Charts and graphs were useful for identifying trends and anomalies, but their value lies in prompting further inquiry rather than providing definitive conclusions. Visualization should therefore be read as a heuristic tool rather than as evidence in isolation.

Methodological mechanics also require careful interpretation. Document clustering groups texts based on shared linguistic characteristics rather than thematic content, meaning that editorials, meeting minutes, book reviews, or even gothic novels may cluster together due to stylistic similarity rather than shared subject matter. Similarly, sentiment analysis does not account for in-group language practices, including the reclaimed or context-dependent use of slurs, which may be misclassified as negative despite their affirmative or political function.

Taken together, these reflections underscore the importance of treating computational analysis as a **critical, interpretive practice** rather than a purely technical one. The methods employed here are most valuable when used in dialogue with historical knowledge, archival awareness, and close reading—allowing patterns to inform questions rather than stand in for answers.

CONCLUSION

This project used the *Gale Digital Scholar Lab* to explore how violence associated with the Stonewall Riots has been described across actors, publications, and time. By combining prebuilt and manually constructed content sets with computational methods such as clustering, topic modelling, n-grams, and sentiment analysis, we examined both contemporaneous accounts and later retrospective narratives of Stonewall.

Taken together, our findings suggest that Stonewall is rarely framed as an isolated violent event. Instead, violence is embedded within broader discourses of identity, activism, institutional power, and collective memory, with tonal and rhetorical shifts becoming more pronounced in post-1972 publications. The contrast between datasets highlights how historical distance enables more explicit critique, commemoration, and affective expression.

Equally important, this project foregrounds the interpretive and methodological labour involved in digital humanities research. Dataset construction, cleaning decisions, tool limitations, and time constraints all shape what can be observed and claimed. Rather than producing definitive answers, this work demonstrates how computational analysis can surface patterns that invite further historical inquiry, critical reading, and methodological reflection.

Please note: Hacking History Event Projects are produced in a single day, during the hackathon, and may not provide fully formed or finalized research. They are published on the Gale Research Showcase to show examples of what can be done in a single day within the digital humanities, using Gale Digital Scholar Lab *and any other DH tools the participants choose to use. They are published to provide inspiration to others in terms of the wide range of approaches, research questions and types of analysis possible within the digital humanities.*