Echoes of Society: Topic and Toxicity Trends in 80 Years of Music

Letizia Dimonopoli
3132775
Zeynep Güler
3180012
Slavena Koleva
3297741
Loren Llagami
3141444
Santiago Luque
3297703
Omar Mukhtar
3182217

Abstract

This paper analyzes 80 years of U.S. song lyrics to trace manospheric language across time and genres. We address challenges like evolving slang using a custom amplified lexicon, keyword-based scoring, and a BERT-based model. Classifiers reveal high manospheric content in hip hop lyrics and themes such as money, drugs and violence. Our approach enables deeper insight into gendered narratives in music and offers tools for cultural, academic, or platform-based analysis.

1 Introduction

Music offers a powerful lens for understanding how social values evolve over time. As cultural attitudes evolve, the language used in music reflects the ideologies of each generation. In this paper, we explore how gender dynamics play out in top-ranked songs in the United States, from the 1940s to the present day.

Although the U.S. has made substantial progress toward greater gender equality (e.g., *Female Labour Force Participation (OECD)* (2017)), significant disparities remain. These inequalities make it crucial to examine how gender structures are reflected or challenged in mainstream music.

We investigate how manospheric language appears and varies across musical genres and themes. We adopt the definition based on the framework proposed in Ribeiro et al. (2020), where manosperity refers to the intersection of toxic masculinity and misogyny, particularly as an **anti-feminist movement** characterized by hostility toward women and often violent rhetoric.

Our findings could offer a valuable resource for future work on misogyny, toxic masculinity and media, and offer practical applications for music streaming platforms, to identify and flag offensive lyrics that undermine or demean women.

2 Data

We source song titles and artists from US Billboard's Year-End charts, and lyrics from Genius.

We apply the following cleaning and preprocessing pipeline: (i) lemmatization and stemming using built-in NLTK methods, (ii) expansion of word contractions via a manually compiled dictionary, (iii) removal of punctuation, hyphenated and repeated words, and repeated two or three-letter n-grams, (iv) flagging of rare words (appearing in fewer than two songs), and POS tagging to identify repeated consecutive tags to reduce disfluencies. Then, all flagged instances are manually reviewed to only filter out true disfluencies.

Moreover, to reduce sparsity in genre labels, we identify 10 overarching genres intended to encompass all 400 sub-genres. We generate word embeddings for the overarching genres and sub-genres using fastText. We assign each sub-genre to the most semantically similar overarching genre based on cosine similarity. Finally, we manually review and refine the resulting mappings to ensure accurate categorization.

3 Experiments

3.1 Manosphere Word List

We compile approximately 2,200 unique terms through a multi-step approach integrating prior research, domain-specific corpora, and automated expansion techniques.

a) Previous Research-Based Extraction

We integrate manosphere-related terms from prior studies. The studies include the lexicon from the *LISTN* paper by de Kock (2024), which analyzes language trends in manosphere subreddits, and the expert-labeled vocabulary from Farrell et al. (2019), which categorizes terms by discourse function and ideological role.

We re-categorize the expert-labeled terms from both sources under a condensed typology we developed:

Category	Description
Aggression & Hostility	Insults, threats, dehumanizing slurs
Masculinity Norms	Emotional suppression, redpill slang, hyper-masculinity
Sexual & Physical Objectification	Terms reducing people to body parts, sexual control
Gender Power Structures	Dominance, patriarchy, gendered hierarchy
Identity-based Discrimination	Homophobia, racism, misogyny, classism

Table 1: Recategorization of expert-labeled manosphericity terms

We test zero-shot classification (`facebook/bart-large-mnli`) on 165 expert-labeled terms using our five-category schema. The model achieves a macro F1-score of 0.29, indicating moderate performance without sentence context. We use macro F1 as it accounts for performance across all classes equally, making it appropriate for our imbalanced label distribution. We apply the zero-shot labeling to the rest of the dataset.

b) Historical Slang Expansion (1940s–2010s)

We scrape historical slang from the *Green's Dictionary of Slang* (2010) to identify archaic terms that express manospheric ideas but are missing from modern lexicons. We collect slang by decade from the 1940s onwards for broad temporal coverage.

c) Word2Vec on 100k Songs

To identify semantically related terms over time, we train Word2Vec models on 100,000 song lyrics, split evenly across three periods. Each model uses lyrics from its own time frame to capture generational shifts in language. From each model, we extract distributional synonyms for 80 curated manospheric seed terms.

3.2 Manually Labeled Manosphericity Scores

Three annotators independently label 600 songs lyrics by assigning a manosphericity score (ranging from 1 to 3), based on the following criteria:

- **1: Minimally manospheric**, which includes lyrics with neutral language or hints of traditional gender roles and/or male-centric views.
- **2: Moderately manospheric**, for which the lyrics present clear gender stereotyping or direct blame towards women.
- **3: Extremely manospheric**, where the lyrics suggest male dominance or control, or feature violent or high misogynistic content.

The resulting label was determined by majority vote. For 5% of the test set, where all annotators gave different scores, the highest score was selected after a discussion to prioritize sensitivity to toxic masculinity content.

3.3 Manosphericity Classifiers

We develop two models to detect manospheric content: a keyword-based scorer and a BERT-based classifier and later combine them into a hybrid approach.

a) WordCounter Model

In the first model, we develop a custom scorer, the WordCounter, using a **category-weighted keyword** and **co-occurrence analysis**. We group keywords by category as defined in Table 1 and weight them by severity. We match the lyrics against the keywords, with log-scaled counts reducing the impact of repetition.

We apply a co-occurrence boost when manospheric terms appear near woman-related words within a context window. If predefined non-manospheric phrases (e.g., "hot girl summer") occur in the same window, we cancel the boost. Final scores combine base counts, co-occurrence context, and category weights, then are normalized across the dataset for comparability.

Finally, we map scores to labels **1**, **2**, and **3** using thresholds optimized through cross-validation for macro F1 and accuracy. Given the small test set and the importance of detecting harmful content, we manually decrease the thresholds to increase sensitivity and reduce overfitting. The full scoring formula is provided in Appendix A.1.

b) Two-Step BERT-based Model

In the second model, we apply a two-step fine-tuning process to a BERT-based classifier.

The initial fine-tuning stage, following Devlin et al. (2018), leverages a binary classification task using 10,000 Reddit posts: 5,000 from known manosphere or extremist subreddits (e.g., *TheRedPill*, *Incels*) labeled as class 1, and 5,000

from general-interest subreddits (e.g., *r/books*, *r/AskReddit*) labeled as class 0. This stage allows the model to develop a general representation of manospheric discourse.

In the second stage, we further fine-tune the model on a smaller, domain-specific dataset comprising 300 of our manually-labeled songs. We assign a value of 1 to songs labeled as moderately or extremely manospheric, while we label all others as 0. Fine-tuning on this dataset helps aligning the model's representations with the specific language style and thematic context of musical lyrics.

Ordinal Evaluation via Isotonic Regression

We originally fine-tune the model for binary classification, while the evaluation dataset includes three expert-labeled classes: *minimally manospheric*, *moderately manospheric*, and *extremely manospheric*. To reflect this ordinal structure, we extract raw logits from the BERT classifier and calibrate them using **Isotonic Regression** by Zadrozny and Elkan (2002), as shown in (6). Using 5-fold stratified cross-validation, we select optimal thresholds $t_1 = 1.15$ and $t_2 = 2.45$ to discretize the calibrated scores into three classes. This method achieves strong performance with macro f1 score of 0.5260 demonstrating its effectiveness in capturing ordinal distinctions from a binary model. (Algorithm and Isotonic Regression formulae can be found in the appendix B).

c) Hybrid Model

We build the hybrid model by combining the `manospherity_score_normalized` from the word counter model described in Section 3.2 with the logit scores from the 2-step fine-tuned BERT model, which we calibrate using isotonic regression. We then feed these two features into an ordinal logistic regression model (Hybrid model) for final classification. To address class imbalance, we resample the training data to ensure a minimum number of samples per class. We tune the hyperparameters using 10-fold cross-validation to maximize the macro F1 score and evaluate the final model on the last 300 samples of the dataset.

Hybrid model outperforms the individual models in terms of both macro F1 score, 0.5374 and accuracy, 0.8833.

3.4 Thematic Labeling

To analyze manospheric content in lyrics, we manually label a subset of 500 songs according to twelve thematic categories expected to have strong links to manospheric content: *heartbreak, romance, money, violence, drugs, sex, spirituality, family, nature, party, politics,* and *time*.

Model	Macro F1 Score	Accuracy
Hybrid	0.5374	88.33%
WordCounter	0.5340	87.00%
2-Step Fine-tuned	0.5260	87.67%

Table 2: Macro F1 Score and Accuracy Comparison Across Models

We apply an 80-20 train-test split to the labeled dataset. To address severe class imbalance, we oversample underrepresented categories. Using cross-validation across multiple thresholds, we observe that model performance stabilizes when each class includes at least 13 examples, which we adopt as the minimum count.

We evaluate several models to assign thematic labels to song lyrics (see Table 6) and select the one with the highest macro F1 score for further analysis. We include detailed descriptions of the alternative models in Appendix A.3.

Our best-performing model is a fine-tuned version of `all-mpnet-base-v2` SentenceTransformer from Reimers and Gurevych (2020), trained using the `MultipleNegativesRankingLoss`. This loss function pulls semantically related lyric–label pairs closer in the embedding space while pushing unrelated ones apart. We train the model on balanced data using the original, unprocessed lyrics, which preserve sentence structure, punctuation, and stylistic variation. Such features support the model's ability to capture rich sentence- and paragraph-level semantics in a 384-dimensional vector space.

4 Results

Figure 1 illustrates how semantic associations with "woman" have shifted from household-focused terms in the 1940s to more transactional negative terms like "whore" in the 2000s, revealing persistent gender bias over time.

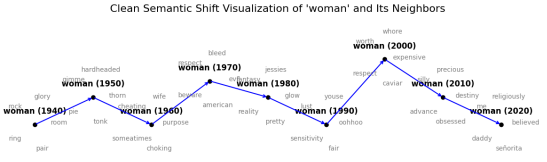

Figure 1: Temporal semantic evolution of "woman"

Temporal analysis (Figure 2) reveals a consistent increase in manospherity-related content starting from the 1990s. While the Hybrid model achieves the highest macro F1 score overall, it fails to detect any manospheric content during the first four decades. In contrast, both the WordCounter and BERT-based models show greater nuance in identifying manospheric language in earlier years, capturing low but present signals. This suggests that while the Hybrid model is more accurate in recent decades, it lacks historical sensitivity. Depending on how conservative or inclusive one wants the model to be, different scoring approaches may be preferable.

WordCounter's conservatism is expected, as it relies on explicit keyword matches, which captures literal expressions in lyrics that often mask meaning through metaphor or slang.

Figure 4 (Appendix) further shows that *aggression & hostility* becomes the dominant manospheric theme after 2000, reinforcing the rise in toxic masculine narratives over time.

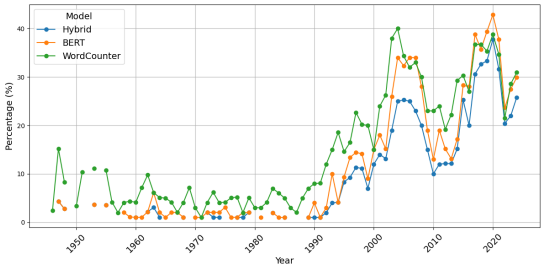

Figure 2: Temporal Distribution of Manospheric Scores Across Models

Thematic analysis in Figure 3 shows that most songs are classified as minimally manospheric. However, themes such as **money, drugs, violence, party** exhibit a higher proportion of songs with moderate to strong manospherity, suggesting that manospheric language tends to cluster around ideas of **power, conflict, and control**. Themes like romance, spirituality, and heartbreak remain largely non-manospheric.

Genre analysis (Figure 5, Appendix) shows that **hip hop** stands out with the highest proportion of songs labeled as moderately or strongly manospheric. **Blues** and **EDM** also show notable presence, while most other genres remain minimally manospheric. This indicates that certain genres provide more prominent platforms for manospheric narratives than others.

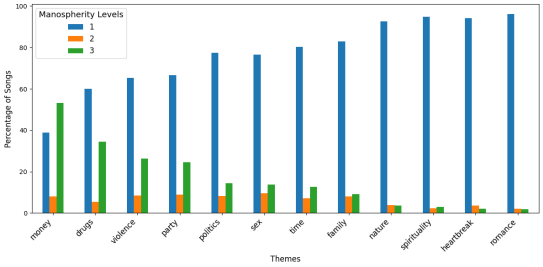

Figure 3: Distribution of Manospherity Levels by Theme (WordCounter Model)

5 Limitations

While our study offers valuable insights, it also comes with several limitations that may affect the interpretation and generalizability of the results. First, the training datasets are relatively small, which may limit the models' ability to generalize beyond the data. Second, the manual labeling process involves subjectivity; many songs include overlapping themes or ambiguous language, making it difficult to assign consistent labels. A further limitation concerns the nature of song lyrics themselves, which often rely on metaphor, slang, and stylistic ambiguity. These characteristics pose challenges for models like SentenceTransformer and BERT, which are primarily trained on literal, prose-based text and optimized for surface-level semantic similarity. As a result, they may struggle to capture the deeper or implicit meanings embedded in poetic or highly stylized lyrics.

6 Conclusion

In conclusion, we trace how manospheric narratives evolve across time, genre, and lyrical themes by considering keyword-based, BERT-based, and hybrid models. Our findings reveal a clear rise in manospheric content beginning in the 1990s, with themes of aggression and hostility becoming especially dominant after 2000. While the Hybrid model achieves the highest overall macro F1 performance, it fails to capture historical manospherity, making the more nuanced WordCounter better suited for longitudinal analysis. We also find that manospheric narratives are most prevalent in genres like hip hop, EDM, and blues, and tend to cluster around themes of power, control, and conflict, such as money, drugs, violence, and party culture. These results offer strong empirical evidence of an increasing trend in manospherity within mainstream music lyrics over the past four decades.

Bibliography

Anglada-Tort, Manuel, Krause, Amanda E., and North, Adrian C. (2021). "Popular music lyrics and musicians' gender over time: A computational approach". In: *Psychology of Music* 49 (3): pp. 426–444.

Rutgers University. *History of Women in the U.S. Congress*. <https://cawp.rutgers.edu/facts/levels-office/congress/history-women-us-congress> [last access: 11/6/2024]

de Kock, Christine (2024). *LISTN: Lexicon induction with socio-temporal nuance*.

Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805. arXiv: 1810.04805.

DeWall, C. N., Pond, Richard S., Campbell, W. K., and Twenge, Jean M. (2011). "Tuning in to Psychological Change: Linguistic Markers of Psychological Traits and Emotions Over Time in Popular U.S. Song Lyrics". In: *Psychology of aesthetics, creativity, and the arts* 5 (3): pp. 200–207.

Farrell, Tracie, Fernandez, Miriam, Novotny, Jakub, and Alani, Harith (2019). "Exploring Misogyny across the Manosphere in Reddit". In: *Proceedings of the 10th ACM Conference on Web Science*. WebSci '19. Boston, Massachusetts, USA: Association for Computing Machinery, pp. 87–96.

(2017). *Female Labour Force Participation (OECD)*. <https://ourworldindata.org/grapher/female-labor-force-participation-oecd?v=1&csvType=full&useColumnShortNames=false> [last access: 11/6/2024]

Green's Dictionary of Slang (2010).

O'Brien, Lucy (2018). "Express Yourself: Reframing women's participation, agency and power in popular music". PhD by Publication. University of Brighton.

Reimers, Nils and Gurevych, Iryna (Nov. 2020). "Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ribeiro, Manoel Horta, Blackburn, Jeremy, Bradlyn, Barry, De Cristofaro, Emiliano, Stringhini, Gianluca, Long, Summer, Stephanie, Greenberg, and Savvas, Zannettou (2020). "The Evolution of the Manosphere Across the Web". In: *ResearchGate*: p. 2.

Our World in Data. *Gender Development Index*. <https://ourworldindata.org/grapher/gender-development-index?v=1&csvType=full&useColumnShortNames=false> [last access: 11/6/2024]

Zadrozny, Bianca and Elkan, Charles (Aug. 2002). "Transforming Classifier Scores into Accurate Multiclass Probability Estimates". In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

A Appendix

A.1 WordCounter calculation and classification

Category Weights. We assign the following weights to emphasize manospheric categories:

Category	Weight α_c
Gender power structures	2.0
Masculinity norms	1.8
Sexual & physical objectification	1.7
Aggression & hostility	1.5
Identity-based discrimination	1.3

Table 3: Manospherity categories and their respective weights

Manospherity Score. We define a manospherity score for each song based on the frequency and context of toxic terms across predefined categories. For each category c, we compute a base score:

$$\text{Base}_c = \sum_{t \in T_c} \log(1 + \text{count}(t)) + \min(M, \text{Boost}_c) \tag{1}$$

where T_c is the set of toxic terms in category c, and Boost_c is the number of co-occurrences with woman-related words within a context window of size W, excluding uplifting phrases.
The final weighted score is:

$$\text{Score}_c = \alpha_c \cdot \text{Base}_c \tag{2}$$

and the total manospherity score is:

$$\text{Total Score} = \sum_c \text{Score}_c \tag{3}$$

Normalization. To ensure comparability across songs, we normalize the total score to a $[0, 1]$ range:

$$\text{Score}_{\text{norm}} = \frac{\text{Total Score} - \min}{\max - \min} \tag{4}$$

where $\min$ and $\max$ are the minimum and maximum scores observed in the dataset.

Label Assignment. We convert normalized manospherity scores into ordinal labels through the following mapping:

$$\text{Label}(s) = \begin{cases} 1 & \text{if } s \leq P_{85} \\ 2 & \text{if } P_{85} < s \leq P_{90} \\ 3 & \text{otherwise} \end{cases} \tag{5}$$

where s is the normalized score, and P_{85}, P_{90} are the 85th and 90th percentiles, respectively. These thresholds approximate the top 10–15% most manospheric songs.

A.2 2-step finetuned Bert and Hybrid model

Algorithm 1. PAV algorithm for estimating posterior probabilities from uncalibrated model predictions.

1: **Input:** training set (f_i, y_i) sorted according to f_i
2: Initialize $\hat{m}_{i,i} = y_i$, $w_{i,i} = 1$
3: **while** there exists i such that $\hat{m}_{k,i-1} \geq \hat{m}_{i,l}$ **do**
4: Set $w_{k,l} = w_{k,i-1} + w_{i,l}$
5: Set $\hat{m}_{k,l} = \frac{w_{k,i-1}\hat{m}_{k,i-1} + w_{i,l}\hat{m}_{i,l}}{w_{k,l}}$
6: Replace $\hat{m}_{k,i-1}$ and $\hat{m}_{i,l}$ with $\hat{m}_{k,l}$
7: **end while**
8: **Output:** the stepwise constant function:
9: $\hat{m}(f) = \hat{m}_{i,j}$ for $f_i < f \leq f_j$

Table 4: **PAV Algorithm** for estimating posterior probabilities from uncalibrated model predictions.

Metric	Class	Hybrid	WordCounter	2-Step
Precision	1	0.9410	0.9630	0.9440
	2	0.1818	0.1140	0.1923
	3	0.8571	0.4400	0.8333
Recall	1	0.9623	0.9150	0.9547
	2	0.2667	0.1600	0.3333
	3	0.3000	0.6670	0.2500
F1-score	1	0.9515	0.9390	0.9493
	2	0.2162	0.1330	0.2439
	3	0.4444	0.5300	0.3846
Support	1	265	542	265
	2	15	25	15
	3	20	33	20
Accuracy	–	**0.8833**	0.8700	0.8767
Macro Avg	Precision	0.6600	0.5060	0.6566
	Recall	0.5096	0.5810	0.5127
	F1-score	0.5374	0.5340	0.5260
Weighted Avg	Precision	0.8974	0.8990	0.8991
	Recall	0.8833	0.8700	0.8767
	F1-score	0.8809	0.8820	0.8764

Table 5: Detailed Classification Report Comparison (Hybrid, WordCounter, and 2-Step Fine-tuned)

A.2.1 Isotonic Regression Formula

$$\min_{\hat{y}_1 \leq \hat{y}_2 \leq \cdots \leq \hat{y}_n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{6}$$

A.3 Model Variants for Thematic Labeling

We test a variety of models for thematic label classification. The first two apply topic modeling: one uses a BERT-based embedding approach, and the other relies on Latent Dirichlet Allocation (LDA), assigning to each song the label of its most probable topic. We also train six SentenceTransformer-based models on both balanced and unbalanced datasets, using lyrics with varying levels of preprocessing. For these models, we assign labels based on cosine similarity between encoded lyrics and candidate labels. In addition, we test a SentenceTransformer model paired with an SVM classifier, as well as two TF-IDF-based models using XGBoost and Naive Bayes, respectively.

Model Abbreviations.

- **BERT_topic**: Topic modeling using BERT embeddings.
- **LDA_topic**: Traditional Latent Dirichlet Allocation topic model.
- **ST_Unbalanced_OG**: SentenceTransformer on unbalanced data using original lyrics.
- **ST_Unbalanced_C1/C2**: SentenceTransformer on unbalanced data using two cleaned lyric versions.
- **ST_Balanced_OG**: SentenceTransformer on balanced data using original lyrics.
- **ST_Balanced_C1/C2**: SentenceTransformer on balanced data using two cleaned lyric versions.
- **ST_SVM**: SentenceTransformer embeddings followed by an SVM classifier.
- **TFIDF_XGB**: TF-IDF features with an XGBoost classifier.
- **TFIDF_NB**: TF-IDF features with a Naive Bayes classifier.

Model	Macro F1 Score
BERT_topic	0.256
LDA_topic	0.035
ST_Unbalanced_OG	0.352
ST_Unbalanced_C1	0.425
ST_Unbalanced_C2	0.328
ST_Balanced_OG	0.476
ST_Balanced_C1	0.415
ST_Balanced_C2	0.354
ST_SVM	0.285
TFIDF_XGB	0.196
TFIDF_NB	0.354

Table 6: Macro F1 Scores for Different Models

A.4 Results

The results of the WordCounter model additionally show the top-occurring category for each song, based on the raw (non-weighted) frequency of matched keywords. This allows us to aggregate manospherity trends by genre and over time.

Temporal analysis (see Figure 4) reveals a marked increase in the presence of manospherity-related categories from the 1990s onward, particularly in the domain of *aggression & hostility*, which dominated the post-2000 period.

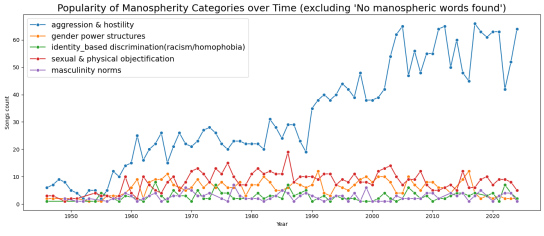

Figure 4: Average normalized manospherity score by year

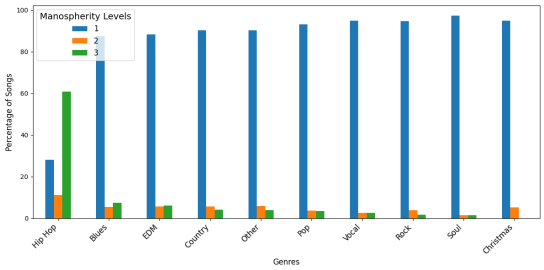

Figure 5: Distribution of Manospherity Levels by Genre (WordCounter Model)

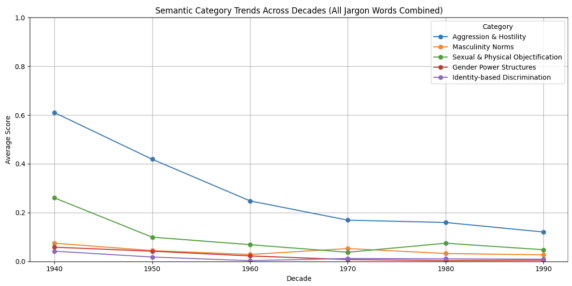

Figure 6: Comparison with Google Book Corpus

A.5 Comparison with Google Book Corpus

To examine how manosphere-related language evolved over time, historical word embeddings trained on the Google Books corpus were aligned with a modern Word2Vec model trained on Reddit manosphere data. Using Procrustes alignment, the closest historical synonyms were retrieved for a set of modern manosphere slang terms across six decades (1940s–1990s). These historical equivalents were then scored using Empath categories, which were aggregated under five thematic frames: Aggression & Hostility, Masculinity Norms, Sexual, Physical Objectification, Gender Power Structures, and Identity-based Discrimination. The analysis showed that although many modern manosphere slang terms do not appear directly in the historical vocabulary, their semantic equivalents were consistently found in earlier decades. In literary and academic writing, different lexical choices have been used to express similar ideas over time. The 1940s, corresponding with the post-war period, showed the highest levels of semantic framing aligned with manospheric themes, particularly in the domains of hostility and objectification. After this period, a downward trend was observed across all categories. This finding suggests that in academic and literary writing, the underlying sentiments and context have long been present but used less manospheric and toxic masculinity terms over time as suggested by the decreasing trend.