Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Anonymity and Sexism: Are Less Identifiable Users More Likely to Post Sexist Comments about the Overturn of Roe v. Wade?

Jordan Dawson

Jordan.a.dawson@northumbria.ac.uk

Northumbria University

Lee Shepherd

Northumbria University

Jenny Paterson

Northumbria University

Tom Fitzpatrick

Northumbria University

Andrew Taylor

Richard Brown

Northumbria University

Scott Houghton

Northumbria University

Genavee Brown

Northumbria University

Research Article

Keywords: Sexism, Social Media, Reproductive Rights, Anonymity

Posted Date: May 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9612084/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Literature that explicitly links online misogyny with anonymity often does so through a narrow lens of Hostile Sexism, or by discussing anonymity as a purely binary variable. The current study addresses these issues by using the Roe Vs. Wade debate on Twitter (now *X*), using a nuanced measure of anonymity, and coding for a broader range of sexism. We predicted that all forms of anonymity will be positively associated with sexism. For this purpose, 1000 tweets and their 55,167 replies were sampled across two time points during the Roe Vs. Wade overturn. To ensure robustness, tweets were coded for sexism using three separate methods: ReWire's API, Google's Perspective API, and a custom-trained Large-Language Model (LLM). Sexism was predicted via two measures of anonymity: the binary variable of nominal anonymity and the continuous variable of reputational anonymity. Analysis revealed a consistent effect of both forms of anonymity in that less identifiable users posted more sexist sentiment across all three methods of coding. In conclusion, this study links anonymity with posting sexism online and develops an LLM to detect sexism specifically related to reproductive rights. It also further highlights the difficulties researchers face when it comes to automatically detecting modern forms of sexism.

1. Introduction

A small but growing body of literature suggests the internet is not a safe space for women. Women are subjected to sexism in internet memes (Drakett et al., 2018), across social media platforms (Bartlett et al., 2014), and were even subjected to sexism during the early days of the Internet (Barak, 2005). While both men and women report frequent experiences of harassment on the internet, women typically receive more frequent and severe harassment (Duggan, 2014) such as online stalking (26% of women report being stalked, compared to 7% of men) and sexual harassment (25% of women report being sexually harassed, compared to 13% of men). Gender-based harassment is not only prevalent across time but is exceptionally harmful; it significantly increases levels of depression and anxiety, predicts self-esteem issues in women (Swim et al., 2001) and, in the case of online sexism, encourages women to remain silent or leave the internet altogether (Nadim & Fladmoe, 2021). Further, there is strong evidence that gender-based microaggressions, which the internet provides a unique avenue for, can significantly increase symptoms of depression and anxiety (Feigt et al., 2022).

A large amount of research on online misogyny has focused on the moderation of sexism *after* it has been posted (Katsaros et al., 2022). While considerable progress has been made in this area, with promising findings from natural language processing models that can learn to detect sexism automatically (Pamungkas et al., 2020), this often still requires sexism to be posted - and possibly viewed - before action is taken. In addition, many women believe the moderation of sexism to be ineffective (McLean & Griffiths, 2019) and there are calls for effective interventions to curb online misogyny proactively (Townsend, 2022). Therefore, it is important to identify factors that influence the likelihood of users posting sexist comments online. Based on this, we specifically assessed the role of anonymity on the likelihood of people posting sexist content online.

1.1. Defining Sexism and its Harms

This study primarily draws upon Glick and Fiske's theory of Ambivalent Sexism (Glick & Fiske, 1997, 2011), which suggests sexism is more than just outright hostility towards women (Glick & Fiske, 2011, 2018). While traditional, Hostile, sexism still exists within modern society (e.g., "I'm glad women have fewer rights, 'yall don't deserve them") it is undoubtedly less common and more socially taboo to express openly. Glick and Fiske argue Benevolent Sexism occupies the other side of the "sexist coin", encompassing a seemingly complimentary form of sexism that upholds masculine, heteronormative ideals (e.g., "Women are better mothers than men are fathers, they shouldn't be getting rid of their babies!"). This form of sexism is much more socially acceptable to express but still harms women despite its positive-sounding nature, being used, for example, to combat online feminist movements such as #MeToo (Huang et al., 2014). Further, Benevolent Sexism is often used to reinforce masculine-favoured societal structures such as encouraging women to take up roles as full-time mothers or house-makers (Gaunt, 2013).

Research into misogyny, particularly research on online misogyny, tends to focus on Hostile Sexism exclusively (Jha & Mamidi, 2017) as it is easier to identify (Glick & Fiske, 1997, 2011). However, there is compelling evidence to suggest Benevolent Sexism is more persistent and tenacious than Hostile Sexism (Hammond et al., 2018); ignoring it in such research would be misrepresentative of the sexism women face in the modern world. Therefore, this study hopes to address issues that are present in the past literature, aiming to make use of sentiment analysis techniques that can reportedly detect both Benevolent *and* Hostile Sexism. This should provide a more holistic, complete overview of the effects online anonymity has on sexism which past literature has failed to achieve.

1.2. Defining Anonymity and its Harms:

The anonymity afforded by the internet is often cited as a reason for the increased levels of hate speech we observe online (Mondal et al., 2017). This anonymity makes users feel *disinhibited*: there are fewer consequences for their actions and posting hateful content under an anonymous account is a lot easier than saying it in person (Suler, 2004). Incivility is, indeed, observed to increase within anonymous environments (Halpern & Gibbs, 2013; Rowe, 2015), and the factors that constitute disinhibition online closely mirror the risk factors that increase sexism offline, such as a lack of accountability (Dall'Ara & Maass, 1999; Fitzgerald et al., 1995). One of the most popular explanations for this phenomenon is the Social Identity model of Deindividuation Effects (SIDE model). This model suggests that those in an anonymous, group environment view themselves less as individuals and instead as part of a larger whole (Ellemers & Haslam, 2012; Lea & Spears, 1991). Because of this deindividuation, they are more likely to adopt, endorse and spread ideals – good and bad - held by the larger group. There is considerable evidence to suggest that the internet was initially developed for, and around, masculine users (Ritter, 2014) and so anonymity could allow these sexist, anti-woman norms to resurface.

However, anonymity is not as easy to define as some literature would suggest. For example, anonymity is typically implied to exist on a binary: a person can be either anonymous (unnamed) or non-anonymous

(named). Especially within the information age, anonymity can perhaps be better captured as a spectrum (Nissenbaum, 1999) and defined as "*non-identifiability rather than simply un-name-ability or namelessness"* (Wallace, 1999). This effect is evident across modern social media platforms; many accounts under a pseudonym end up with large followings and strong identities that make them far more recognisable and individual than an account that uses a real name but never posts and purely "lurks" (Chen, 2012). Even without their real name displayed, an un-named account could be less susceptible to the effects of the SIDE Model if it has spent time individualising itself to stand out amongst the crowd. Conversely, a user with their real name displayed but with no post history or profile picture could be argued to be just as anonymous as a stranger in a large crowd; some identifying information is on display but the account has no strong, individual identity.

The above provides justification and rationale for many of the decisions taken within the current study. Previous research into anonymity has stopped at merely comparing named and unnamed accounts, but current theories of anonymity highlight such an approach is limited. Further, the research that has looked at anonymity's effect on sexism has done so across platforms (Mondal et al., 2017), which has methodological issues when one considers the SIDE Model. When a user experiences deindividualisation they are perhaps more likely to conform to the group norms of the website they are on, not that of a general internet user. Therefore, to create a fair comparison, anonymous and non-anonymous users must be sampled from the same website and the same group norms.

1.3. Detecting Sexism Online

A large amount of research on sexism online has manually coded for sexist content (e.g., Benton-Greig et al. (2018)). However, with the development of artificial intelligence, it is now possible to assess sexism using automatic sexism detectors. Natural Language Processing Models such as BERT or RoBERTa have shown promising results in detecting misogyny automatically, vastly outperforming traditional methods of automatic coding and showing high accuracy compared to manual coding (Pamungkas et al., 2020). In addition, there are several automatic hate speech-detecting Application Programming Interfaces (APIs), such as Google's Perspective API (Perspective API, 2021) and ReWire API (*ReWire API*, 2021), developed within the past few years. These modern models all work largely the same: they are trained on data coded by researchers and then given testing data also coded by researchers to determine their accuracy. While these models can be time-consuming to train, they can then code new data without supervision.

Though these models have gained a large amount of focus within the literature, much of the work into natural language processing has focused purely on the text of a social media post, ignoring the contextual factors that can improve the detection of hate in both post and pre-production. It has been demonstrated that including a user's metadata, such as their Twitter Status, into natural language processors such as BERT improves their accuracy and ability to detect general hate speech (Casavantes et al., 2023). Despite this, a recent meta-analysis found only a handful of studies have investigated the contextual factors surrounding a hateful social media post (Casavantes et al., 2023), and fewer still have investigated this concerning sexism specifically. Therefore, the current study aims to investigate whether

contextual factors surrounding a post, namely a user's anonymity, can help predict sexism within the text.

1.4. Aims and Hypothesis:

The current study aims to address the gaps in the previous literature highlighted above, determining whether anonymity can predict sexist sentiment in a social media post. This study shall contribute to the existing literature by investigating a broader range of sexism, investigating anonymity as more than just a binary variable, and sampling users from a single website rather than across platforms.

We hypothesize, based on past literature, that anonymity will predict the strength of the sexist sentiment present within a social media post.

2. Methods

All methods and analyses have been preregistered. R code and raw outputs can be accessed on the Open Science Framework here: https://osf.io/7pbn2/?view_only=bfaa91fbec4a4b489dead46a67fc8f25.

2.1. Data and Corpus Collection

2.1.1. Context (Defining and Justifying Roe Vs. Wade)

Roe *vs.* Wade was a landmark decision made by the United States Supreme Court in 1973, guaranteeing access to abortion as a federal right. While states could alter the laws around abortion (such as when, during pregnancy, a woman can access abortion), Roe Vs. Wade ensured there was a minimum level of access. In May 2022 a leak from the Supreme Court revealed this decision was going to be overturned in the ruling on Dobbs v. Jackson Women's Health Organization. A month later in June, the ruling in Dobbs was officially released, and states were given the power to legislate women's access to abortion. This decision could only *reduce* a woman's access to abortion across America.

This topic was chosen to be the subject of analysis for a few important reasons. First, the leak in May and the overturn in June generated a large amount of online discussion on Twitter. Twitter allows users to create accounts with varying levels of anonymity, allowing for within-platform comparisons between anonymous and non-anonymous accounts. Secondly, the debate around abortion fits the study's aims well due to its controversial nature, its direct relation to women's rights, and its unique representation of men and women on both sides of the argument. Finally, the discussion generated a variety of sexist opinions that captured both sides of Ambivalent Sexism (Glick & Fiske, 2011). Some benevolent arguments framed women as ideally chaste, motherly, caring, and the primary child carers who require protection from a harsh world. Other, more hostile, arguments framed women as stupid and unable to make important decisions about their bodies without the consent of their partner. Indeed, both extreme and subtle forms of sexism were present in the debate creating a unique opportunity for data collection.

2.1.2. Data Collection

Tweets were collected as the Roe Vs. Wade overturn was being debated in mid-2022, using the (then free and open-access) official Twitter developer portal that allows for tweet collection. The Python package Twarc2 and Rstudio package RTweet were used to integrate with the developer interface for convenience and simplicity. These packages collect relevant metadata that is made use of in this study; namely profile demographics (e.g., followers) and tweet statistics (e.g., retweets).

Using these packages, tweets were collected if they were *not* a Quote, Retweet, or Reply and included one of the following five hashtags: #BansOffOurBodies, #ProLife, #ProChoice, #RoeVsWade, or #RoeVWade. Data collection took place for a week surrounding two time points, when the overturn was leaked (May 2nd 2022). and when the ruling was officially overturned (June 24th 2022), as this was when the discussion was most prevalent. The hashtags were chosen as they encompass both sides of the debate around Roe *vs*. Wade.

These search terms generated an original corpus of data of over 100,000 distinct tweets. However, a majority of these were repeat tweets or tweets that received no engagement from other users. The top 100 most engaged tweets (determined by averaging replies, quotes, and retweets) from every hashtag (5) across both time points (2) were selected for further analysis. This resulted in 1000 total tweets (100 tweets x 5 hashtags x 2 time points). All replies to these 1000 tweets were collected resulting in a final corpus of 1000 “parent” tweets and 85,850 “reply” tweets. See Table 1 for further descriptive statistics of the “parent-level” tweets.

Table 1

Mean	**Likes**	**Replies**	**Quotes**	**Retweets**
	3202 (13992)	**332 (996)**	**110.70 (360)**	**649 (2898)**
Median	427	55	10	105

Mean (SD) and Median likes, replies, quotes and retweets each of the “parent-level” tweets received.

This particular method of sampling was chosen for a variety of reasons. Researchers were interested in creating a large corpus of data that captures both initial tweets and the conversations around them. However, sampling the conversations around all 100,000 initial tweets would be far too time-consuming. By only sampling the most engaged tweets researchers make the corpus of data more manageable whilst also ensuring only the tweets that were viewed by others and generated discussion were analysed. This should also provide some rudimentary protection from bot accounts.

Among the parent tweets, there were 689 unique users. Among the reply tweets, there were 55,167 unique users. The engagement each tweet received varied greatly, with the most engaged tweet receiving over 2000 replies and the least engaged tweets receiving only 2.

2.1.3. Data Cleaning

Tweets were originally excluded if they were duplicated during the data collection process (for example, if a tweet contained both #BansOffOurBodies and #ProChoice it was sampled twice); before data analysis tweets that were purely emojis, not in English, or contained hyperlinks were excluded. This is because the automatic coding systems intended to be used during the analysis process would not be able to accurately label these tweets.

The final corpus consisted of 62,957 tweets with 55,167 unique users. Analyses were run on the 55,167 tweets with unique users to ensure the independence of observations.

2.2. Measurements

As mentioned previously, much research has been undertaken to classify and identify identity attacks within text. Three measures were selected due to their prior use in research: RoBERTa Natural Language Processing, ReWire API, and Google's Perspective API. Our goal in using a variety of automatic coding processes is to ensure a high level of robustness; if a result is significant across all three methods it suggests a consistent effect. Before performing this automatic coding, tweets were converted to lowercase and punctuation, special characters, mentions, usernames, numbers, and line breaks were removed as is standard practice for automatic sentiment coding.

2.2.1. Researcher-coded Tweets and RoBERTa

BERT (and its derivatives, such as RoBERTa) is one of the most effective natural language processing models to automatically code for sexism (Casavantes et al., 2023). To train a RoBERTa model, the first hashtag to be sampled (#BansOffOurBodies, during the Leak of Roe Vs. Wade) was hand-coded for Sexism. A total of 1600 tweets were divided across three pairs of researchers who coded the datasets independently (Avg. Kappa = 0.7). Then, the two primary researchers settled disagreements between coders until total agreement (Kappa = 1) was achieved.

Tweets were coded according to the definitions of sexism presented in Section 1.2, with researchers carefully considering the notion of Benevolent Sexism *as well as* Hostile Sexism. For example, users arguing against abortion might say it is "a shame" to see women not appreciate motherhood. These types of sentiment were labelled as sexist, as they imply women are predisposed to being mothers and are in the wrong for rejecting the "opportunity" to give birth. Tweets were coded in a binary fashion: either 1 for sexist or 0 for non-sexist.

From this, and utilising the HuggingFace Library, a RoBERTa model (an updated type of BERT model) was trained on the dataset to detect Sexism according to the researcher's definitions automatically. Researchers made use of the pre-trained model "NLP-LTU/bertweet-large-sexism-detector", which is a RoBERTa model already trained to detect sexism within tweets, and finetuned it on the hand-coded data. The eventual model produced fair results, with an accuracy of .837 and an F1 score of .667. This mostly mirrored the results found in Casavantes et al. (2023), though with a slightly lower F1 score due to the smaller amount of training data available.

As with the API's discussed below, this BERT model produces a score of between 0 and 1 which denotes how confident it is sexism (both Benevolent *and* Hostile) is present within a tweet. A score of 0 means the model is 100% sure there is no sexism within a tweet, and a score of 1 means the model is 100% sure there *is* sexism present.

2.2.2. Google and ReWire API

Google's Perspective API and ReWire's API are more generic language models that can detect various sentiments within text. While they also provide a numerical score between 0 and 1, the number represents how likely the model believes a reader would claim a specific sentiment is present within a given text. A score of 1 means the model believes all readers would view the sentiment as present, and a score of 0 means the model believes all readers would view the sentiment as absent. While both APIs code for a number of different sentiment types, the sentiment of interest for ReWire is "Hate" (which is able to detect sexism) and for Google's Perspective API is "Identity Attack" (which is also able to detect sexism). Specifically, Google's API classes an "Identity Attack" as "*Negative or hateful comments targeting someone because of their identity"*. This does mean other forms of identity attack could be detected using these models, but the topic selection of Roe Vs. Wade should ensure most examples of identity attacks are gender-based.

Both these models have been trained on relevant data in order to detect sexism and ReWire in particular claims to be especially effective at detecting sexism (Kirk et al., 2023). However, it should be noted neither API codes for sexism exclusively within their variables of interest. Similarly, the models are better at coding for Hostile Sexism than Benevolent Sexism, though experts have annotated ReWire's training data and it should outperform Google's model in this regard (Kirk et al., 2023).

2.2.3. Coding for Nominal Anonymity

In line with previous research (Mondal et al., 2017), an open-source list of names was gathered from GitHub (https://github.com/zeraye/names-surnames-list), resulting in 5486 common names based on 2020 American Census data. Further to this, a list of the 1000 most common US English words was gathered (https://gist.github.com/deekayen/4148741). Users were labelled as anonymous (coded as 1) if their Twitter handle did **not** match any of the names present on the list of names. In addition, if a Twitter user's handle contained any of the common English words they were labelled as anonymous. Otherwise, they were labelled as non-anonymous (coded as 0). 16 users had to be removed as their names had not been recorded properly.

Random samples were taken from both the anonymous and non-anonymous groups to further refine the exclusion words and inclusion names. For example, the original list of names included words such as King and Angel. While these are legitimately used as names, they were more often used by accounts that were better described as anonymous and therefore these "names" were removed from the name list. Of the 55,151 users, a total of 32,920 were anonymous and 22,231 were non-anonymous.

Researchers labelled this variable as “nominal anonymity”, as it refers to whether the user has an anonymous name attached to their account. Higher scores on Nominal Anonymity indicate *higher levels* of anonymity.

2.2.4. Coding for Reputational Anonymity

Similarly in line with previous literature and with more advanced theories of anonymity, a continuous variable of “reputational anonymity” was created. In the case of Twitter, an account that has a larger following, a more comprehensive set of followed users and a larger trail of previous tweets was determined to be more identifiable than one without. These variables suggest an account is used frequently and has a stronger individual identity that might be less susceptible to deindividuation.

Higher scores on Reputational Anonymity indicate *more* anonymity, indicating a user had fewer followers or tweets.

2.3. Analytical Procedure

Before analysis, many numerical values were log-transformed due to the highly skewed nature of user demographics. A user’s Followers, Following, and Total Tweets were log-transformed. All API scores were also multiplied by 100 to make data handling and interpretation easier.

Due to high inter-correlations, Identifiability was created as a latent variable consisting of Followers, Following, and Total Tweets (higher scores on these variables indicate greater anonymity). As part of the structural equation modelling (SEM) Identifiability and Anonymity were used to predict sexism (separately using ReWire, Google’s Perspective API, and BERT).

3. Results

Three separate Structural Equation Models were conducted on for each of the APIs used. All models examined the relationships among:

1. The latent variable *Reputational Anonymity* consists of total Followers, total Following, and total Tweets. Higher values indicate higher reputational anonymity.
2. The dichotomous variable *Nominal Anonymity* (0 = non-anonymous, 1 = anonymous)
3. The outcome variable, one of the three methods of sexism detection. While the individual model’s scores mean something slightly different (explained in section 2.2), higher scores indicate a stronger sexist sentiment.

All analyses were conducted using RStudio (ver. 4.3.2) and are available on the OSF.

3.1. Descriptive statistics

Consistent with past literature (Mondal et al., 2017), Table 2 shows the rates of sexism across the sample were fairly low. The custom BERT Model detected the highest sexist sentiment (*M* = 12.72, *SD* =

27.29), followed by ReWire (*M* = 12.16, *SD* = 16.63), and then by Google's API (*M* = 7.32, *SD* = 11.92).

In addition, the models demonstrated significant correlations with each other suggesting they somewhat measure the same construct. The correlation between ReWire and Perspective is notably higher than that of the custom BERT model, however, suggesting there is a disagreement between these models as to what constitutes sexism. This disagreement may stem from the BERT model's focus on Benevolent sexism, which is discussed later:

Table 2

BERT	BERT	ReWire	Perspective
ReWire	.13**		
Perspective	.23**	.52**	

The correlation between the three models of sexist sentiment detection. All correlations were significant at p<.001 (**).

3.2. ReWire

The variable of interest for ReWire's API was "Hate". The SEM predicting Hate demonstrated good fit (Hu & Bentler, 1999), as indicated by the Comparative Fit Index (CFI = 0.997), Tucker-Lewis Index (TLI = 0.994) and the Root Mean Square Error of Approximation (RMSEA = 0.035, 90% CI [0.032, 0.038].

Results suggest a small, but significant, effect of nominal anonymity and a slightly larger effect of reputational anonymity on the presence of Hate in a Tweet. Results suggest that users with anonymous names tend to post more hateful sentiment (β = 0.79, p < .001, standardised β = 0.02), and posts by those with more reputational anonymity are typically more sexist (β = 0.43, p < .001, standardised β = 0.06). See Fig. 1 for a graphic representation of the results, including details on the latent variable, and Fig. 2 for a word cloud highlighting the most frequently used words in the most sexist 25% of tweets as well as the least sexist 25% of tweets.

Figure 2

Two word clouds highlighting the most frequently used words within the most sexist 25% of tweets (left) and the least sexist 25% of tweets (right), according to ReWire's API

3.3. Google's Perspective

The variable of interest for Google's Perspective API was "Identity Attack". The SEM predicting Identity Attacks demonstrated good fit, as indicated by the Comparative Fit Index (CFI = 0.997), Tucker-Lewis Index (TLI = 0.994) and the Root Mean Square Error of Approximation (RMSEA = 0.035, 90% CI [0.032, 0.038].

Results suggest a small, but significant, effect of nominal anonymity and a slightly larger effect of reputational anonymity on the presence of an Identity Attack in a Tweet. Results suggest that users with anonymous names tend to post more hateful sentiment (β = 0.58, p < .001, standardised β = 0.02), and those with more reputational anonymity are typically more sexist (β = 0.25, p < .001, standardised β = 0.05). See Fig. 3 for a graphic representation of the results, including details on the latent variable, and Fig. 4 for a word cloud highlighting the most frequently used words in the most sexist 25% of tweets as well as the least sexist 25% of tweets.

Figure 4

Two word clouds highlighting the most frequently used words within the most sexist 25% of tweets (left) and the least sexist 25% of tweets (right), according to Google's Perspective API.

3.4. Custom BERT Model

The custom-trained BERT model creates two variables; a dichotomous variable called "Label" determining whether it thinks a tweet is sexist or not and a continuous variable called "Score" in a similar manner as the previous APIs. The variable of interest in this case was "Score". The SEM predicting sexism demonstrated a good fit, as indicated by the Comparative Fit Index (CFI = 0.997), Tucker-Lewis Index (TLI = 0.994) and the Root Mean Square Error of Approximation (RMSEA = 0.035, 90% CI [0.032, 0.038].

Results similarly suggest a small, but significant, effect of nominal anonymity and a slightly larger effect of reputational anonymity on the presence of sexism in a Tweet. Results suggest that users with anonymous names tend to post more sexist sentiment (β = 0.59, p < .001, standardised β = 0.01) and those with more reputational anonymity are typically more sexist (β = 0.69, p < .001, standardised β = 0.06). See Fig. 5 for a graphic representation of the results, including details on the latent variable, and Fig. 6 for a word cloud highlighting the most frequently used words in the most sexist 25% of tweets as well as the least sexist 25% of tweets.

Figure 5

SEM Plot demonstrating the relationship between Anonymity (both Reputational and Nominal) and sexist sentiment (Custom BERT Model)

Figure 6

Two word clouds highlighting the most frequently used words within the most sexist 25% of tweets (left) and the least sexist 25% of tweets (right), according to the custom-trained LLM

4. Discussion

4.1. Overview of Findings:

This study investigated the role of anonymity and whether this influences the presence of sexism sentiment in a tweet, specifically in response to the Roe *vs*. Wade overturn. The present study employed a robust measure of sexism, utilising three separate APIs on a large corpus of data. Across all three measures, and in line with original predictions, a consistent effect was found: anonymous users were more likely to post sexist sentiment. For nominal anonymity, the effect was small (approximately a 1% change between anonymous and non-anonymous users) but for user reputational anonymity, this effect was slightly larger (the least identifiable users were around 10% more sexist). Engaged and identifiable Twitter users that had many followers, posted many tweets and followed more users, were less likely to reply with sexist sentiment compared to a less identifiable account that had fewer tweets, fewer followers, and fewer followed users.

4.2. Nominal and Reputational Anonymity and Sexism:

The current study contributes to previous literature by investigating anonymity as more than just a dichotomous variable, highlighting the merits of such an approach. This is especially true when considering the small effect nominal anonymity had across all three measures of sexist sentiment, being associated with a 1% increase in sexism in the most generous of cases. While this effect was quite consistent across the three language models, further research is required to understand why the effect of namelessness was so small when compared to past literature and other forms of online hate crime (Mondal et al., 2017).

In contrast, reputational anonymity had a much stronger relationship with sexist sentiment. While previous literature does suggest the effect could be stronger than what is displayed here (Fox et al., 2015), these findings clearly demonstrate the merit of investigating anonymity beyond simple namelessness. This makes sense when considering the SIDE Model (Lea & Spears, 1991), as an account with a larger following, a more robust list of followed users, and more posts is more personalized and perhaps less likely to become deindividualized. Similarly, some of the factors that influence the Toxic Disinhibition effect apply less to unnamed but developed accounts. An account isn't really "invisible", for example, if it has a very large following and a long post-history (Lapidot-Lefler & Barak, 2012). Finally, these findings support past research which suggests contextual factors surrounding a tweet have merit in detecting hate speech (Casavantes et al., 2023). If social media platforms are interested in curbing hate speech and implementing better measures of automatic moderation, it would appear useful to include more than just the content of the post.

The small effect sizes detected in the dataset are somewhat unsurprising. There is considerable evidence to suggest Benevolent Sexism is much more subtle and, to many, isn't easily identifiable as Hostile Sexism (Glick & Fiske, 1997, 2011). This could explain the RoBERTa Model's small effect sizes and low correlations, as this was the model researchers specifically trained to detect Benevolent Sexism. It's possible many users did not view their comments as sexist, so less anonymous users felt comfortable sharing their sexist tweet without being disinhibited or deindividualized. Finally, researchers noted that many users discussed Roe vs. Wade as a political matter, and not a feminist one. Many sexist

men feel a level of entitlement when it comes to abortion (Petterson & Sutton, 2018) and so, again, may have felt like they were having a political discussion rather than a sexist one.

4.3. API Discussion

There were some key differences between the three language models used. Notably, ReWire API tended to demonstrate the strongest effects; this is unsurprising, however, given it has been specifically trained on misogynist social media content and was reported to excel at detecting gender-based discrimination (Kirk et al., 2023). The RoBERTa model, on the other hand, typically had the smallest effect sizes and displayed a low correlation with the other models despite being one of the best methods of automatic sexism detection (Casavantes et al., 2023). As mentioned in the previous section, these results demonstrate the difficulty researchers and social media platforms have if they wish to automatically detect all forms of discrimination. Further to this, some topics were present in both sexist and empowering arguments making automatic detection even more difficult. Specifically, the use of contraceptives was present in anti-abortion (putting the onus on contraceptive use solely on women) and pro-abortion (claiming the SCOTUS was coming for contraceptives next) rhetoric.

4.3. Limitations and Future Directions:

While steps were taken to ensure a high level of robustness, some limitations are present within the study that could be developed in further research. The method by which nominal anonymity was determined matches that of previous literature (Mondal et al., 2017), but is not without its flaws. For example, a user with the name “Tracey the Dragonfly” would be labelled as nominally anonymous despite having a level of identifiability present. As researchers wanted to discuss the nature of binary vs continuous anonymity a simpler method was chosen, but this further highlights the non-binary nature of anonymity. Further research might benefit from dismantling nominal anonymity further into a more continuous variable, capturing the different levels of anonymity a username can encompass.

It is also important to acknowledge the limitations of automatic sentiment coding. This form of coding is, obviously, less accurate than manual coding especially considering the more nuanced nature of sexism. When considering the size of the dataset at hand, however, researchers are unaware of different methods that could be employed. Further, many language models struggle to detect different forms of sexism (Butt et al., 2021) and can neglect Benevolent Sexism in particular (Jha & Mamidi, 2017). Considering ReWire and Perspective don’t specifically target sexism in the first place, this could be a considerable limitation for the two pre-trained APIs. Researchers sought to overcome this with a custom-trained model on hand-coded data to detect sexism specifically, but Benevolent Sexism remains hard to code. Further research could also benefit from including additional training data when creating the custom models. The RoBERTa model in the current study made use of 1600 tweets which were very robustly coded, but the size of the training dataset ended up quite small.

Finally, it’s important to highlight the correlational nature of the study. Therefore, we can infer that anonymity predicts posting sexist sentiment online, but cannot say that it specifically causes online

sexism. Although there are a few experimental studies that show anonymity causes online sexist behaviour (Fox et al., 2015), further, experimental, research is needed to confirm these findings.

4.4. Conclusions and implications:

This research succeeded in determining whether it was possible to predict sexism using measures of anonymity, finding that less identifiable users are more likely to post sexist, hateful content. While there are some limitations in the current study that could be considered for further study, findings are consistent across three different, well-fitting, structural equation models demonstrating a high level of robustness. Further, this study contributes to a small body of literature that highlights the benefits of contextual factors surrounding a tweet, such as the anonymity of the poster, can have.

Declarations

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. The authors also declare that this research received no additional funding to conduct.

Clinical trial number: not applicable.

Author Contribution

The work of the authors has been mapped on to the CREDIT taxonomy: https://credit.niso.org/Jordan Dawson – Conceptualisation, Data curation, Formal analysis, Investigation, Methodology, Writing (original draft). Genavee Brown – Conceptualisation, Investigation, Supervision, Writing (review & editing)Lee Shepherd – Supervision, Writing (review & editing)Jenny Paterson – Supervision, Writing (review & editing)Thomas Fitzpatrick – InvestigationAndrew Taylor – InvestigationRichard Brown – InvestigationScott Houghton – InvestigationAll authors reviewed the manuscript prior to submission.

Data Availability

Data have been anonymised and made available on the OSF: https://osf.io/7pbn2/?view_only=bfaa91fbec4a4b489dead46a67fc8f25

References

1. Barak, A. (2005). Sexual harassment on the Internet. *Social science computer review*, *23*(1), 77–92.
2. Bartlett, J., Norrie, R., Patel, S., Rumpel, R., & Wibberley, S. (2014). Misogyny on twitter.
3. Benton-Greig, P., Gamage, D., & Gavey, N. (2018). Doing and denying sexism: online responses to a New Zealand feminist campaign against sexist advertising. *Feminist media studies*, *18*(3), 349–365.

4. Butt, S., Ashraf, N., Sidorov, G., & Gelbukh, A. F. (2021). *Sexism Identification using BERT and Data Augmentation-EXIST2021*. IberLEF@ SEPLN.
5. Casavantes, M., Aragón, M. E., & González, L. C. (2023). & Montes-y-Gómez, M. Leveraging posts' and authors' metadata to spot several forms of abusive comments in Twitter. *Journal of Intelligent Information Systems*, 1–21.
6. Chen, F. (2012). Passive forum behaviors (lurking): A community perspective. *Embracing Diversity in the Learning Sciences* (pp. 128–135). Routledge.
7. Dall'Ara, E., & Maass, A. (1999). Studying sexual harassment in the laboratory: Are egalitarian women at higher risk? *Sex Roles*, *41*(9–10), 681–704.
8. Drakett, J., Rickett, B., Day, K., & Milnes, K. (2018). Old jokes, new media–Online sexism and constructions of gender in Internet memes. *Feminism & psychology*, *28*(1), 109–127.
9. Duggan, M., Rainie, L., & Smith, A. (2014). *Online Harassment.*. https://www.pewresearch.org/internet/2014/10/22/online-harassment/
10. Ellemers, N., & Haslam, S. A. (2012). Social identity theory. *Handbook of theories of social psychology*, *2*, 379–398.
11. Feigt, N. D., Domenech Rodríguez, M. M., & Vázquez, A. L. (2022). The impact of gender-based microaggressions and internalized sexism on mental health outcomes: A mother–daughter study. *Family relations*, *71*(1), 201–219.
12. Fitzgerald, L. F., Swan, S., & Fischer, K. (1995). Why didn't she just report him? The psychological and legal implications of women's responses to sexual harassment. *Journal of Social Issues*, *51*(1), 117–138.
13. Fox, J., Cruz, C., & Lee, J. Y. (2015). Perpetuating online sexism offline: Anonymity, interactivity, and the effects of sexist hashtags on social media. *Computers in human behavior*, *52*, 436–442.
14. Gaunt, R. (2013). Ambivalent sexism and perceptions of men and women who violate gendered family roles. *Community Work & Family*, *16*(4), 401–416.
15. Glick, P., & Fiske, S. T. (1997). Hostile and benevolent sexism: Measuring ambivalent sexist attitudes toward women. *Psychology of women quarterly*, *21*(1), 119–135.
16. Glick, P., & Fiske, S. T. (2011). Ambivalent sexism revisited. *Psychology of women quarterly*, *35*(3), 530–535.
17. Glick, P., & Fiske, S. T. (2018). The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Social cognition* (pp. 116–160). Routledge.
18. Halpern, D., & Gibbs, J. (2013). Social media as a catalyst for online deliberation? Exploring the affordances of Facebook and YouTube for political expression. *Computers in human behavior*, *29*(3), 1159–1168.
19. Hammond, M. D., Milojev, P., Huang, Y., & Sibley, C. G. (2018). Benevolent sexism and hostile sexism across the ages. *Social Psychological and Personality Science*, *9*(7), 863–874.

20. Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, *6*(1), 1–55.
21. Huang, Y., Osborne, D., Sibley, C. G., & Davies, P. G. (2014). The precious vessel: Ambivalent sexism and opposition to elective and traumatic abortion. *Sex Roles*, *71*, 436–449.
22. Jha, A., & Mamidi, R. (2017). When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. Proceedings of the second workshop on NLP and computational social science.
23. Katsaros, M., Yang, K., & Fratamico, L. (2022). Reconsidering tweets: Intervening during tweet creation decreases offensive content. Proceedings of the International AAAI Conference on Web and Social Media.
24. Kirk, H. R., Yin, W., Vidgen, B., & Röttger, P. (2023). SemEval-2023 task 10: explainable detection of online sexism. *arXiv preprint arXiv:2303.04222*.
25. Lapidot-Lefler, N., & Barak, A. (2012). Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in human behavior*, *28*(2), 434–443.
26. Lea, M., & Spears, R. (1991). Computer-mediated communication, de-individuation and group decision-making. *International journal of man-machine studies*, *34*(2), 283–301.
27. McLean, L., & Griffiths, M. D. (2019). Female gamers' experience of online harassment and social support in online gaming: A qualitative study. *International Journal of Mental Health and Addiction*, *17*(4), 970–994.
28. Mondal, M., Silva, L. A., & Benevenuto, F. (2017). A measurement study of hate speech in social media. Proceedings of the 28th ACM conference on hypertext and social media.
29. Nadim, M., & Fladmoe, A. (2021). Silencing women? Gender and online harassment. *Social science computer review*, *39*(2), 245–258.
30. Nissenbaum, H. (1999). The meaning of anonymity in an information age. *The Information Society*, *15*(2), 141–144.
31. Pamungkas, E. W., Basile, V., & Patti, V. (2020). Misogyny detection in twitter: a multilingual and cross-domain study. *Information processing & management*, *57*(6), 102360.
32. Perspective, A. P. I. (2021). Google. https://www.perspectiveapi.com/
33. Petterson, A., & Sutton, R. M. (2018). Sexist ideology and endorsement of men's control over women's decisions in reproductive health. *Psychology of women quarterly*, *42*(2), 235–247.
34. ReWire (2021). *API*. ReWire. https://rewire.online
35. Ritter, B. A. (2014). Deviant behavior in computer-mediated communication: Development and validation of a measure of cybersexual harassment. *Journal of Computer-Mediated Communication*, *19*(2), 197–214.
36. Rowe, I. (2015). Civility 2.0: A comparative analysis of incivility in online political discussion. *Information communication & society*, *18*(2), 121–138.

37. Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & behavior*, *7*(3), 321–326.
38. Swim, J. K., Hyers, L. L., Cohen, L. L., & Ferguson, M. J. (2001). Everyday sexism: Evidence for its incidence, nature, and psychological impact from three daily diary studies. *Journal of Social Issues*, *57*(1), 31–53.
39. Townsend, K. (2022). September 21, 2022). *Strengthening Accountability for Online Gender-Based Violence – one year later*. World Wide Web Foundation. https://webfoundation.org/research/strengthening-accountability-for-online-gender-based-violence-one-year-later/
40. Wallace, K. A. (1999). Anonymity. *Ethics and Information technology*, *1*(1), 21–31.

Figures

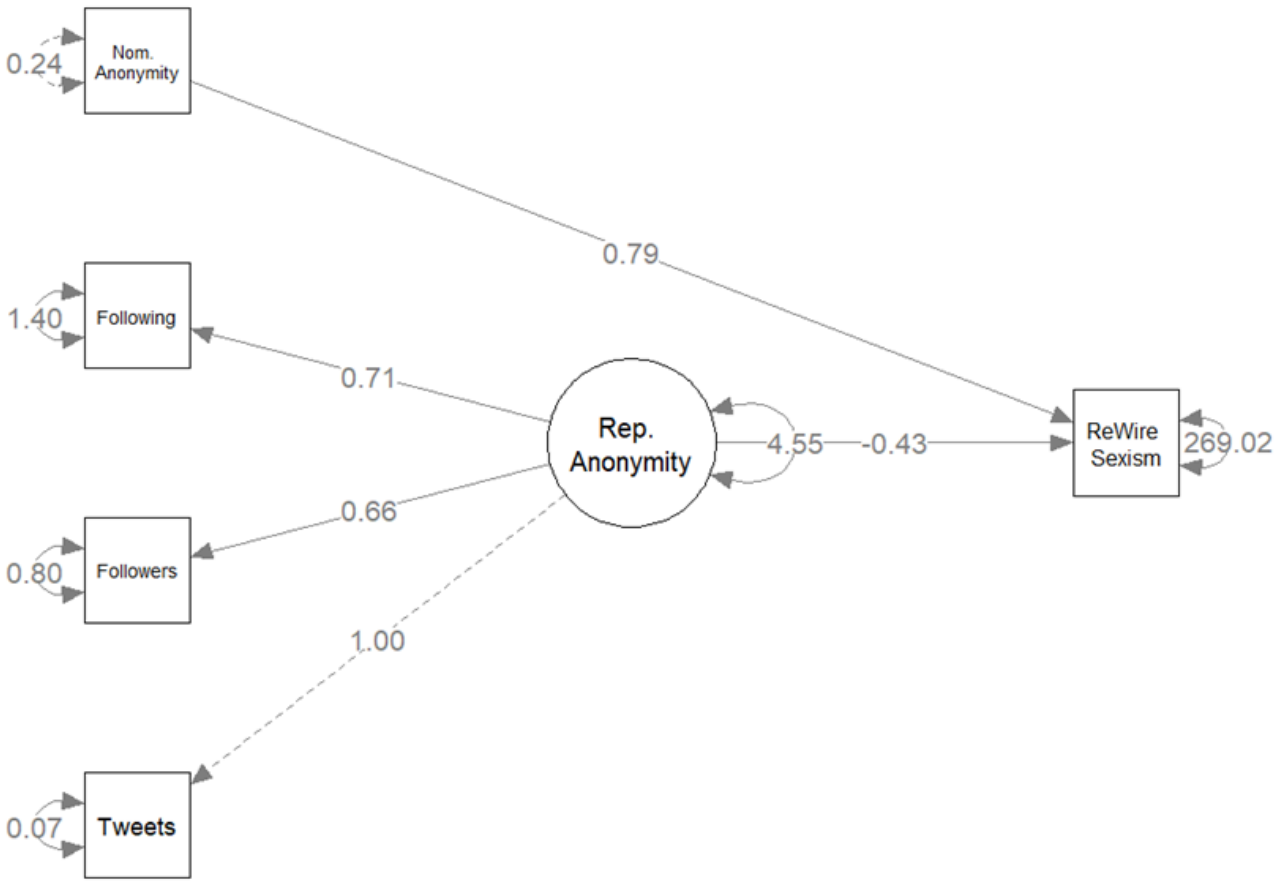


Figure 1

SEM Plot demonstrating the relationship between Anonymity (both Reputational and Nominal) and sexist sentiment (ReWire's API)

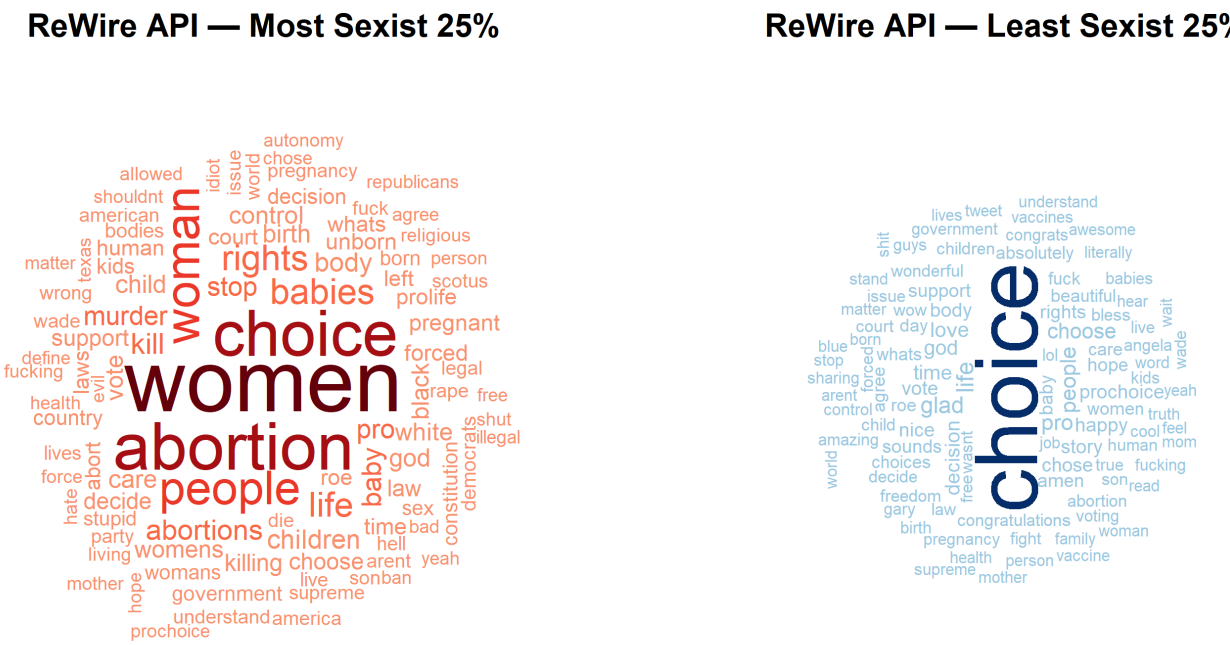


Figure 2

Two word clouds highlighting the most frequently used words within the most sexist 25% of tweets (left) and the least sexist 25% of tweets (right), according to ReWire's API

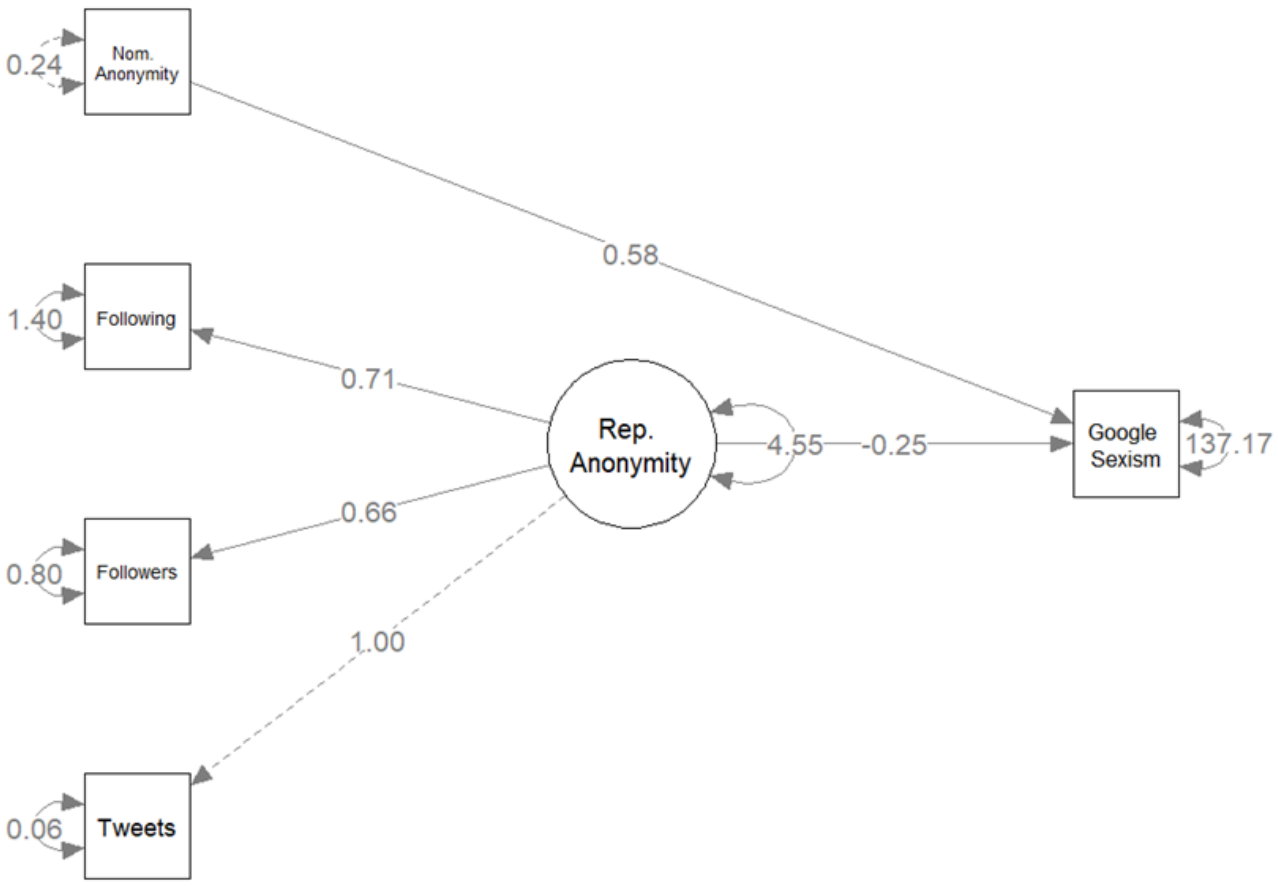


Figure 3

SEM Plot demonstrating the relationship between Anonymity (both Reputational and Nominal) and sexist sentiment (Google's Perspective API)

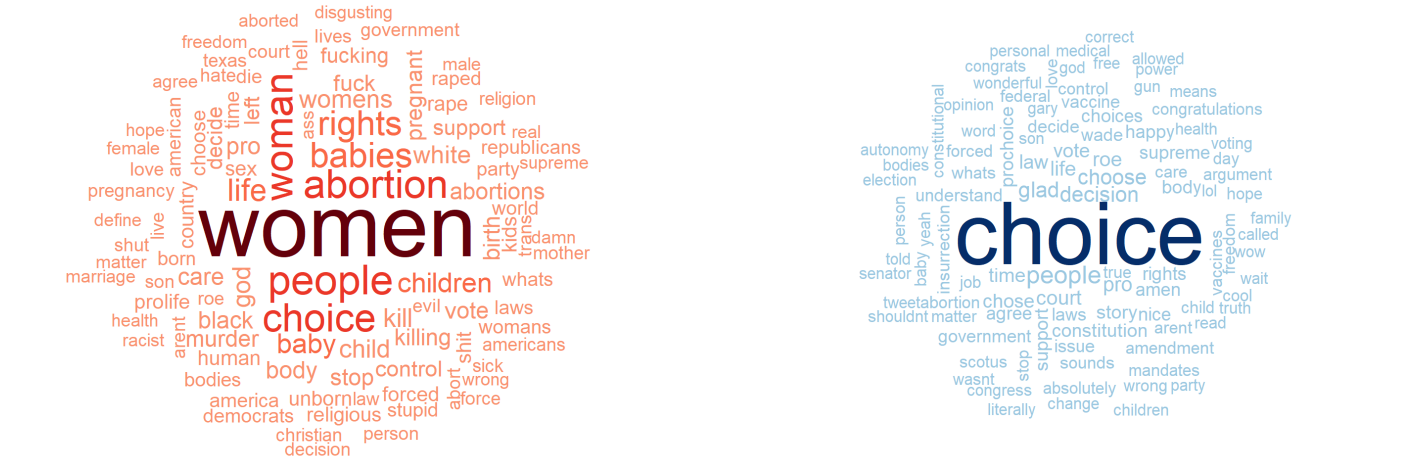


Figure 4

Two word clouds highlighting the most frequently used words within the most sexist 25% of tweets (left) and the least sexist 25% of tweets (right), according to Google's Perspective API.

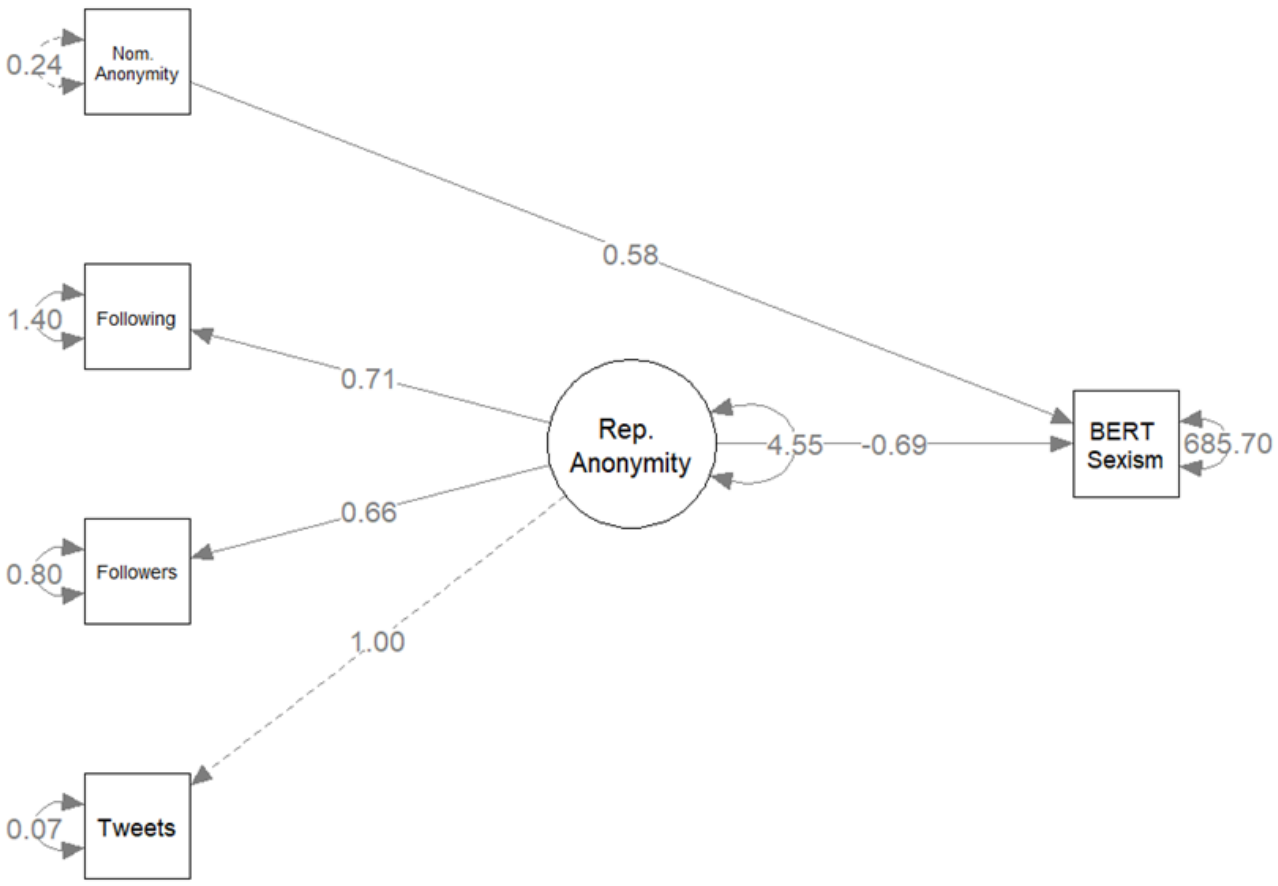


Figure 5

SEM Plot demonstrating the relationship between Anonymity (both Reputational and Nominal) and sexist sentiment (Custom BERT Model)

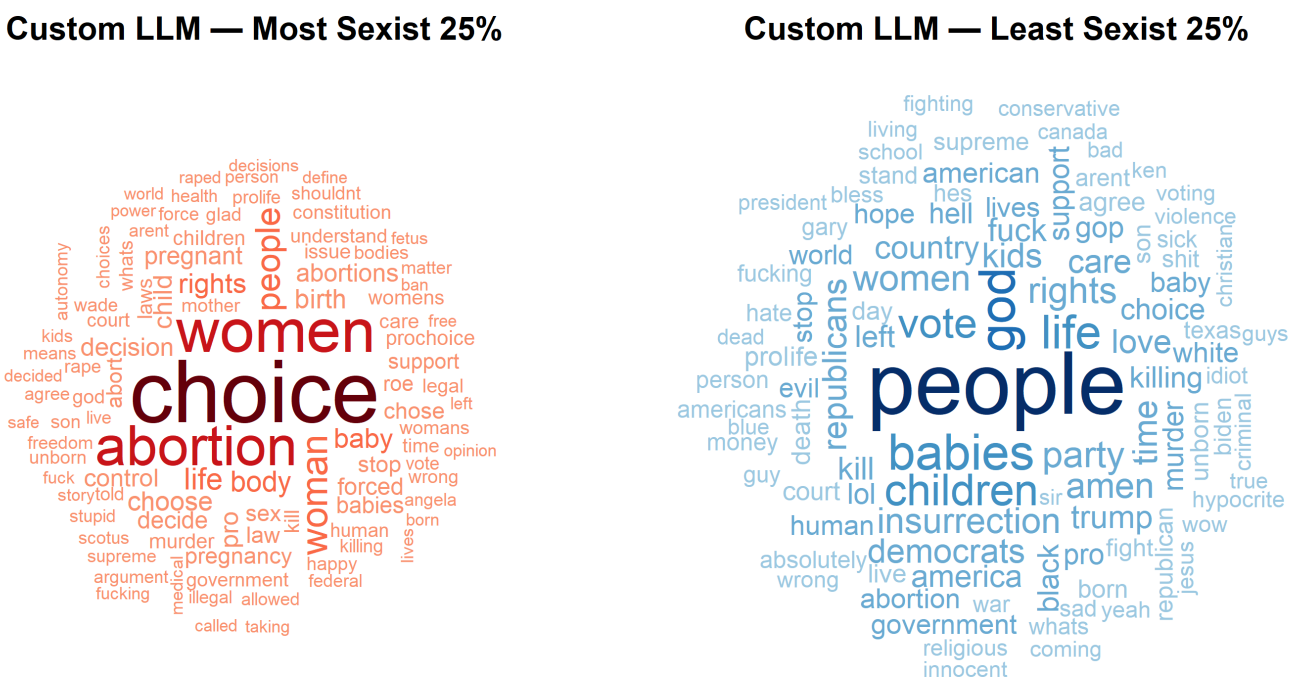


Figure 6

Two word clouds highlighting the most frequently used words within the most sexist 25% of tweets (left) and the least sexist 25% of tweets (right), according to the custom-trained LLM