From Tumblr to Twitter: Evidence of NSFW Community Migration via Bio Keyword Prediction

Alessio Tosolini
McGill University
Montreal, Québec, Canada
alessio.tosolini@mail.mcgill.ca

Jason Jeffrey Jones
Stony Brook University
Stony Brook, New York, USA
jason.j.jones@stonybrook.edu

Abstract

Tumblr was one of the most popular microblogging platforms throughout the 2010s, boasting loose regulations around adult content that fostered the formation and growth of various fandom communities. After Tumblr's December 2018 ban on adult content, these communities were driven to migrate to another platform. This study provides evidence for a migration of users from Tumblr to Twitter following the ban using bio keyword prediction. We collect 400,000 Twitter bios for every year between 2017 and 2022 and train a logistic regression model to predict the presence of 'porn' and 'NSFW' given the other tokens in the bio and evaluate the models on 100,000 held-out bios. Through the learned model weights, we find evidence for 'porn' being primarily predicted by keywords advertising nudity for sexual consumption, while 'NSFW' correlates with accounts of individuals that present nudity as just one part of their online presence, mirroring how NSFW existed on Tumblr. Using the models' testing accuracy as a proxy for predictability, we find that both 'porn' and 'NSFW' originally existed in small niches on Twitter which grew in scope over the years. Taken together, we conclude that there was a mass influx of users from Tumblr to Twitter following the porn ban that (re-)formed community around the keyword 'NSFW'. Finally, we make available BioKPD (Biographical Keyword Prediction Dataset), a dataset of model weights and testing accuracies for n=1630 frequently used identity signifiers.

Keywords

social media platform migration, identity signifier prediction, NSFW, porn, Tumblr, Twitter, computational social science

ACM Reference Format:
Alessio Tosolini and Jason Jeffrey Jones. 2026. From Tumblr to Twitter: Evidence of NSFW Community Migration via Bio Keyword Prediction. In *18th ACM Web Science Conference (WebSci '26), May 26–29, 2026, Braunschweig, Germany*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3795766.3799778

1 Introduction

On many Web platforms, individuals compose self-descriptions and thus choose *identity signifiers*—symbols that represent aspects of the self. In this paper, we use Twitter profile biographies as a source for self-descriptive text and introduce a method to predict the presence or absence of signifiers of interest.

WebSci '26, Braunschweig, Germany

ACM ISBN 979-8-4007-2504-3/26/05
https://doi.org/10.1145/3795766.3799778

We demonstrate the method by exploring the use of `NSFW` and `porn` within bios. We find evidence that `NSFW` grew in prevalence and scope on the Twitter platform following the introduction of speech restrictions on Tumblr.

1.1 Background

In recent years, the relationship between the online presence of individuals and sexually explicit content has changed drastically thanks to the shifting landscape of social media apps. Tumblr, a microblogging social media site with relaxed guidelines, was known for providing marginalized groups with a platform to share all types of content and develop online communities based on identity [33]. However, a December 2018 ban on adult content on Tumblr voided one of the primary online landscapes of NSFW content [25]. Marginalized sexual communities and subcultures that developed online over the course of years vanished, resulting in a decrease in the space for these communities to exist [4]. As a result, users explored other social media platforms in search of a new place to safely express sexual and nonsexual aspects of their identity online [32].

'NSFW', an initialism of the phrase 'not safe for work', is nowadays generally used to describe content that is sexually explicit. Although its first known use was in the year 2000 to describe the shock website Stile Project [22], which featured both pornographic content and gore, the term has since moved away from its original meaning to primarily denote pornographic content [23]. At the same time, there has been a significant increase in the prevalence of unpaid amateur online pornography [28]. For example on Tumblr, taking and posting NSFW pictures of oneself served as a way to create an empowered identity and control the narrative surrounding what is 'sexy' [31] and as a way for young people to develop a sense of belonging while experiencing the thrill of edgework [11].

Twitter, now known as X, is a microblogging social media platform where users can share information through text, photos, or videos. Upon the creation of an account, users are prompted to create a bio describing themselves in 160 characters or less. This voluntary self-identification has been previously used as a way to study how people self-identify and perceive themselves and the content they post over time [18]. The exact words in a Twitter bio can serve as an indicator of a user's belonging to a group or a warning to other users about the content they post. For example, including a pride flag emoji or a 'they/them' in one's bio signals the user identifies as queer or nonbinary. Sampling tweets allows estimating the rates at which tokens are used in the active population through time. For example, the increase in the use of sexualities and pronoun lists in Twitter bios can be taken as evidence for the increased visibility of queerness online [17, 35].

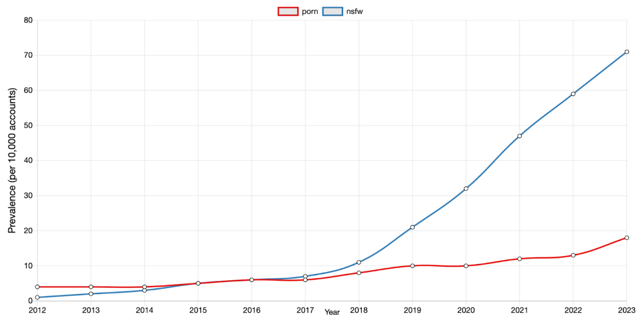


Figure 1: Prevalence of 'nsfw' and 'porn' in Twitter bios from 2012-2023

1.2 NSFW and Porn in User Bios

In this paper, we examine the context and temporal changes for the signifiers `NSFW` and `porn` within Twitter users' self-authored self-descriptions. Including `NSFW` as opposed to `porn` in one's Twitter bio may signify that the content that one posts is sexual in nature, but not necessarily pornographic, suggesting a distinction between the meanings of NSFW and pornographic. To observe the frequency with which these words are used , we measure their frequency in Twitter bios for a given year using the formula for prevalence shown below.

$$\text{Prevalence}(\text{keyword}, \text{year}) = \frac{\#\text{ bios with keyword in year}}{\text{total number of bios in year}} \cdot 10000$$

Prevalence of the `NSFW` and `porn` signifiers diverged over time. Figure 1 shows prevalence history of 'nsfw' and 'porn' in US Twitter bios between 2012 and 2023. In 2012, `NSFW` had a prevalence of 1 bio per every 10,000, with `porn` being also relatively infrequent with a prevalence of 4. In 2015 `NSFW` caught up to `porn`, with both keywords sharing a prevalence of 5. From 2018 forward, the growth of `NSFW` prevalence strongly outpaced the growth of `porn`. In 2023, `porn` reached a prevalence of 18 - having grown to nearly four times its 2015 prevalence - while `NSFW` multiplied its prevalence more than fourteen times to 71 bios per 10,000. Annualized growth estimates in the prevalence of `NSFW` and `porn` were +10.7 and +2 respectively between the years of 2017 and 2023, increasing four and one standard deviations respectively every year [18], highlighting the extreme increase in the usage of these words relative to other tokens in Twitter bios.

1.3 Internet Migration

The push-pull theory of migration distinguishes between factors that influence someone to *leave* a given space and factors that incentivize people to *join* a new one [20]. Internet migrations, such as the migration from Facebook to Instagram [14] and a recent migration from Twitter to Mastodon [15], have been studied using this framework. Different social media platforms attract distinct user groups because of the affordances they provide and the communities those affordances sustain. Pre-NSFW ban Tumblr, with its loose regulations on NSFW content and large fandom communities, provided communities interested in NSFW with unique pull factor: a safe space to upload and share their content while interacting with that of similar users. Once communities have been established on such a platform, banning adult content may instead serve to push users to other platforms [36].

Looking at the history of fandom migrations allows us to place Tumblr's NSFW ban and subsequent migrations in a larger context. The first online platform commonly used to host fanfiction and fandom content was Usenet [9, 13]. Even at its peak, Usenet didn't garner the number of users that LiveJournal or Tumblr boasted according to a 2018 survey [9], with the pull to LiveJournal being driven by the new features it had to offer rather than anything pushing users away from Usenet [9, 34].

LiveJournal hosted plenty of fandom communities that existed outside of the norm of what was considered palatable by mainstream audiences, including fandom communities centered around adult content [36]. In May 2007, LiveJournal launched the initiative 'Strikethrough' to remove journals run by child predators, banning any instances of pornography featuring minors. However, the ban was defined loosely enough that cases of NSFW artwork with characters of ambiguous ages were also censored [36]. Many users felt this was an indication that Livejournal did not respect adult content creation on principle or care about the desires of its users [9, 36]. Tumblr was not meant to imitate LiveJournal prior to Strikethrough due to differences in the features each site had to offer [36], but loose regulations surrounding NSFW content likely resulted in Tumblr being a place of landing for many LiveJoural artists. The migration of users from LiveJournal to Tumblr casued by Strikethrough is also unique in being the only online mass-migration prior to 2018 being driven by a push- rather than a pull-factor [9].

Strikethrough closely mirrors the Tumblr NSFW ban in being a major policy change on the biggest platform hosting alternative fandom content at the time. Prior to 2018, about 11 to 20 percent of content posted on Tumblr was adult in nature [16], with one estimate placing over 20% of Tumblr users using the app for its NSFW content [6]. In the two months following the ban, Tumblr lost 29% of its traffic [30], with the censorship primarily causing an exodus of the queer and feminist users that disproportionally used Tumblr compared to other platforms [29]. This exodus was heightened by the AI used to detect adult content commonly and disproportionately misflagging gay non-adult content, fan art, and sex education material [16]. One of the biggest deterrents to online mass migrations is the lack of a critical mass of users to start similar communities on a new platform [9], but like with Strikethrough, the Tumblr NSFW ban resulted in the communities most affected by the new censorship to need to find a new platform all of a sudden, largely resolving the 'critical mass' problem that occurs with slower mass migrations.

Twitter, which was used by 24% of American adults in 2018 [2], proved to be a promising migration site for the displaced Tumblr users. In the years before 2018, Twitter began to be a platform where sexual communities could form around minority racial identity [5, 37], minority gender identity [24], sex work [19], and other marginalized or underrepresented identities. At the same time, Tumblr users not belonging to fandom communities but still consuming the Tumblr's pornographic content may have been also drawn to Twitter due to its due to its relaxed regulations regarding sexual content [7] and rise in self-directed pornographic content [24]. Together, this set the stage for an online mass-migration of users from Tumblr to Twitter.

1.4 Our Goal

In this paper, we explore the emergence of NSFW and pornographic communities on Twitter in light of the Tumblr NSFW ban using computational methods that give new insight on how these communities evolve over time. We do so by investigating the ways that two Signifiers of Interest (SoI)—`NSFW` and `porn`—evolved in meaning and associations from 2017 to 2022[1]. First, we examined these two SoIs as a signal resulting from user migration from Tumblr to Twitter. Then, we repeated the methods for 1628 other words and emojis – all tokens appearing in at least 7 bios per 10,000 from 2017 to 2022. This broader analysis opens the possibility for many subsequent studies similar to this one.

By investigating which tokens predict the presence of the two SoIs over time, we identify the social and semantic niches that `NSFW` and `porn` occupy on Twitter and how these niches evolve. Tracking changes in the tokens most strongly associated with each SoI allows us to observe shifts in meaning and patterns of co-occurrence from 2017 to 2022. This approach makes it possible to not only see whether the meaning of these words broaden or narrow in use, but how the identities associated to them change over time. In doing so, we capture trends in identity signaling and community (re-)formation on Twitter, and reveal how Twitter's affordances made it a suitable landing place for Tumblr users following Tumblr's NSFW ban. More broadly, we use this predictive framework to shed light on the large-scale movement of NSFW communities from Tumblr to Twitter.

2 Methodology

Our analytical approach proceeds in the following steps: (i) collecting Twitter bios and preprocessing the data, (ii) performing logistic regression to predict the presence of `NSFW` and `porn`[2] and interpreting the estimated predictive strength of predictor tokens, (iii) evaluating model accuracy on a held-out testing set and calculating ROC curves, (iv) plotting and visualizing trends in token weights, testing accuracy, and ROC curves. A schemetized overview of the methodology is provided in Figure 2.

The usefulness of Twitter bios to analyze identity trends online is well supported by previous work [10, 26]. The methodology presented here is ideal for the analysis of identity at scale and is easily adapted for the analysis of other keywords provided in the Biographical Keyword Prediction Dataset (BioKPD) that is made available with this paper[3]. A snapshot of this dataset demonstrating its functionality is provided in Appendix A.

[1]2017 was selected as a start year, because in our sample of 500,000, this was the first year where both 'nsfw' and 'porn' had a prevalence of 7 or greater. 2022 was selected as an end year, because changes to the Twitter API pricing made it impractically expensive to collect data after June, 2023.

[2]As discussed later, the authors note that `NSFW` and `porn` do not exhaustively cover the possibilities of adult content within users' self-descriptions.

[3]The full dataset can be found at https://zenodo.org/records/18852754

2.1 Source of Bios

All Twitter bios used in this study were collected according to the methodology described in [18]. Specifically, the dataset was derived from a random sample of tweets. The Twitter Streaming API was used to observe a random 1% sample of all tweets. These were filtered the tweets authored by likely US users by filtering on the profile location field: Locations likely not in the US (e.g. 'London, UK') were filtered out and locations indicating a US place name (e.g. state names and formal abbreviations) were kept. Some tweet authors were observed many times, however, only one bio per user was sampled for each year. Therefore, the stream of random tweets was reduced to a cross-sectional, annual sample of snapshots of American tweet author bios.

2.2 Preprocessing

These bios are then preprocessed and each transformed into a set of tokens. We define the tokens in a bio as every item present in a bio (e.g. words, emojis, punctuation) after splitting on the following regular expression: `\b|\s+`, which separates strings on word boundaries (such as punctuation marks) and whitespace. Duplicate tokens are removed, such that every bio is represented as the set of present tokens. In order to avoid model overfitting, we define our *Predictor Token Dictionary* as all tokens with prevalence greater than or equal to 3 for every year in 2017 through 2022; tokens with lower prevalence were discarded from further analysis. No out of vocabulary token was used, though a unique token was used for empty bios and bios with no tokens in the Predictor Token Dictionary.

Similarly, we constructed a *Signifier of Interest (SoI) Dictionary*, consisting of all tokens with a prevalence greater than 7 for each year from 2017 through 2022. The SoIs represent tokens whose predictability we aim to track over the years in our range using the larger set of Predictor Tokens as features. The *Predictor Token Dictionary* is comprised of 3535 unique tokens, while the *Signifier of Interest (SoI) Dictionary* is comprised of 1630 tokens. We make available weights for *all* predictor tokens per *each* SoI in the Biographical Keyword Prediction Dataset released alongside this paper (BioKPD), but within this manuscript we restrict analysis to the two SoIs `NSFW` and `porn`.

2.3 Model Training

For each SoI–year pair, we train a logistic regression model to predict whether the SoI appears in a given bio. We randomly sample 500,000 unique bios per year and split them 80–20 into training and test sets. Prior to training, we remove all occurrences of the SoI from the bios while recording which bios originally contained it. We then train a logistic regression model for four epochs on the 400,000 training bios, using the remaining tokens as binary features to predict whether the SoI was originally present. For example, if the SoI is basketball and a bio reads 'Professional basketball player and sommelier', we remove basketball (the SoI) and sommelier (which falls below the minimum prevalence threshold of 3 bios per 10,000), leaving 'Professional player and' while retaining the label that the bio originally contained basketball. The model must then infer the presence of basketball from the remaining tokens. After training, we extract the learned weights for each predictor token,

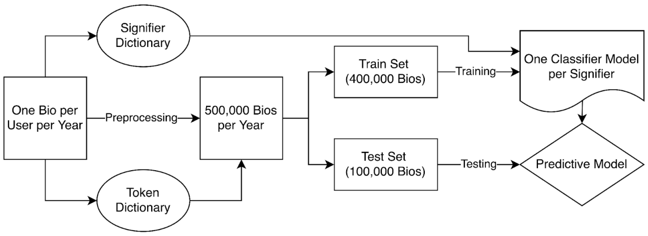


Figure 2: Flowchart of model training and evaluation.

which quantify their predictiveness of the SoI. Model performance is evaluated by predicting SoI presence in the 100,000 held-out test bios and computing test accuracy.

We repeat this methodology for each of the 1630 keywords in our Signifier of Interest Dictionary to create the Biographical Keyword Prediction Dataset (BioKPD). As such, one model was created for each of the 1630 SoIs over the 6 years in 2017 to 2022, yielding a total of 9780 logistic regression models. All models were trained across four GPUs for a total of approximately four GPU days of training.

3 Results

3.1 Logistic Regression Weight Analysis

We start by investigating the mean weights over the span of 2017-2022 for all tokens predicting NSFW. We do the same for porn. Figure 3 plots the 20 tokens with the highest mean weight for the two SoI we focus on in this paper. The red line represents the mean negative bias line. For a bio to be predicted as having the SoI, the weights of all other tokens must sum to a value higher than the negative bias. For porn, this means that given the weights averaged out over the years 2017-2022, a bio containing the token 'star' will automatically be predicted as also containing the SoI porn (unless other words with negative weights overwhelm this signal). Interestingly, the model learns to assign low weight values to other words that frequently appear with 'star' in order to account for bigrams that contain 'star', such as 'star wars', with 'wars' having the third lowest weight for porn at -4.06. Top mean predictors for porn include words describing those shown in the videos (e.g. 'star', 'stars', 'model', 'performer'), genres (e.g. 'amateur', 'gay'), and terms associated with restrictions or filtering of users (e.g. 'no', 'block', 'blocked').

From Figure 3 we can similarly see the top 20 tokens and weights for the SoI NSFW. Unlike for porn, there is no one token whose weight exceeds the negative bias value. Tokens that highly predict NSFW include words that highlight its occasional presence in the content of the poster (e.g. 'occasionally', 'sometimes', 'occasional', 'often', 'may'), warning for younger audiences (e.g. 'warning', '18', '+', with the last two often appearing together as '18+'), or the source (e.g. 'tumblr', 'rts' - short for retweets) or genre of the content (e.g. 'tumblr', 'soft', 'gay', 'rp' - short for roleplay). The differences in the tokens that predict NSFW and porn suggests a difference in the niche they inhabit on Twitter.

We can also track predictor weights for these two SoIs across the span of the 6 years to see how they have evolved over time. In Figure 4, we look at the 10 tokens with the highest mean predictive power for a given SoI and track their evolution across 2017 to 2022. The black line represents the negative value of the bias, which also changes over time. On the left, we see that certain tokens have stayed steady in their predictive power of porn, while others have changed drastically. For example, the top two predictors, 'star' and 'amateur' both remain high across the span of the six years, suggesting that their use in bios containing porn does not fluctuate much. However, 'addict' increased from a weight of 0.42 to 3.61, representing a rapid and large increase in people identifying as a 'porn addict'. Table 1 presents four randomly sampled bios containing the keywords porn and addict in 2022, exemplifying the ways in which porn and 'addict' may be presented in a bio and how it may co-occur with other aspects of identity. Though there are bios that contain both NSFW and 'addict', there were no bios in our database for any year between 2017 and 2022 that contain the bigram 'nsfw addict'. Redundant words, such as 'sex' and 'xxx', decreased in use as Twitter porn gained popularity and became more mainstream [7].

For NSFW, we see the predictiveness of 'tumblr' and 'likes' increase dramatically from 2017 to 2018, then plateauing, suggesting an initial large migration of users posting NSFW content from Tumblr to Twitter after Tumblr's December 2018 ban on adult content. We additionally see a sharp increase in the predictiveness of 'likes', often used by users who signal that their liked posts are NSFW in nature or simply signaling the interests of the poster. Randomly sampled bios from 2018 containing NSFW co-occuring with likes and tumblr can be seen in Table 2, with emojis represented between colons. Meanwhile, 'furry' faces a steady decrease in predictive power from year to year, representing that 'furry', though often paired with NSFW in 2017, progressively widened its reach over the course of the following years, supported by the fact that its prevalence increased every year between 2017 and 2022. Finally, the steady decrease in the value of the negative bias in Figure 4 represents the space which NSFW inhabits on Twitter broadening throughout the years, as fewer tokens are able to reliably predict the presence of the word.

yes im a husky guy... yes i love all kinds of sex ... i an addicted to porn and edging almost no limits very hard kinks
sex addict - mostly to gay porn
24 yo NSFW 18+ minors not allowed! pronouns (sperm/dump) i am a porn addict / gooner / perv / pussyfree / degenerate / my sex life is my bed & iPhone
recovering addict studying dark psychology. sarcastic with a hint of horny but always shitposting. definitely on a revenge porn website somewhere

Table 1: Sample bios with the keywords porn and addict in 2022.

Additionally, given a SoI and a year, we can look at the distribution of the weights of the tokens that predict it. Figure 5 plots the number of predictor tokens at varying levels of mean logistic regression weight. For a bio to be classified as containing a given SoI, the sum of the weights of the tokens needs to exceed the negative bias. From these plots, we can see the high percentage of tokens that

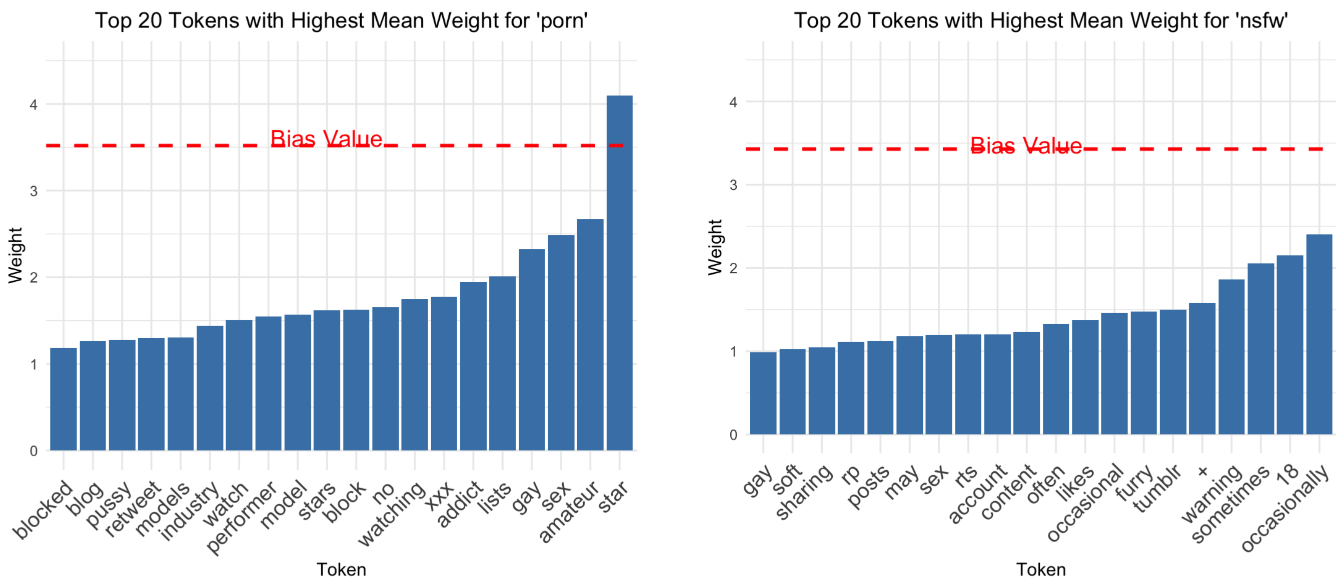


Figure 3: Top 20 tokens predicting 'porn' and 'nsfw' with highest mean weight over 2017–2022.

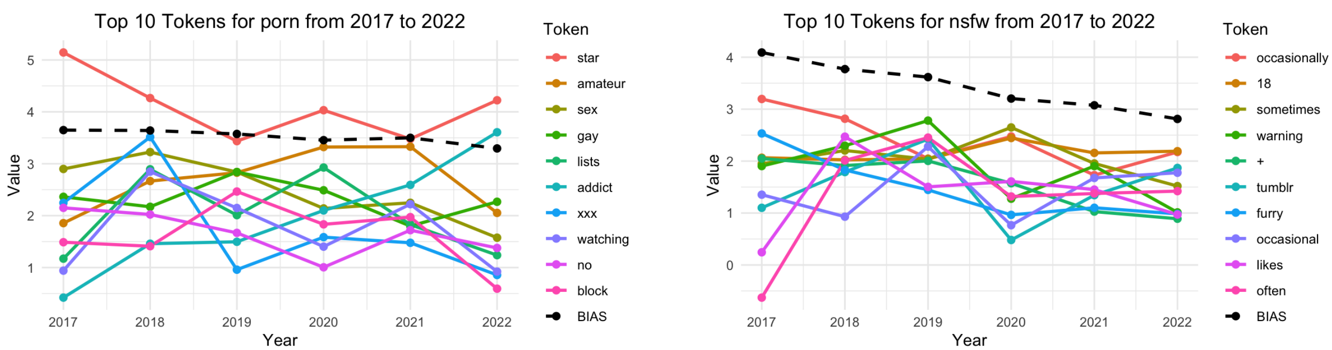


Figure 4: Top 10 tokens with highest mean weight.

:MexicanFlag:25:ColombianFlag: US Army vet, Hentai connoisseur :Squid: Tech/animal geek :Rat:Gaming dude :Dragon:Aspiring actor :PerformingArts: Pan:RainbowFlag: Here there be NSFW likes :NoUnder18: :BlackHeart: [@ ANONYMIZED] :BlackHeart:	Hey ya'll, Tumblr died so here I am :3 also there will be NSFW
Furry/Kemono | 22 | Art Enthusiast | some #nsfw retweets/likes	#Skyrim Monster #NSFW tumblr refugee. Gfycat: [LINK ANONYMIZED] Piczeltv: [LINK ANONYMIZED]
Just a trashy gal who draws trashy art. Mostly retweets. (DISCLAIMER: My likes are full of NSFW/Fetish art, sorry in advance) My NSFW/Fetish art: [@ ANONYMIZED]	Refugee from Tumblr as Liminal-Wanderings (NSFW, 18+ only) late 30s/He-Him/Poly/Married, Hypnosis/Mind Control Kink, Femsexual. Subject/Bottom primary.
Genderfluid (she/they mostly) | 25 | Show me gryphons & werewolves pls. | Icon [@ ANONYMIZED]| Header [@ ANONYMIZED]| Sometimes #NSFW in likes, rare in retweets.	22 year old artist. I draw NSFW (18+) stuff and all things soft. SFW:[@ ANONYMIZED], Rant:[@ ANONYMIZED]. Find my commission info on Tumblr ([USERNAME ANONYMIZED])

Table 2: Sample bios with the keywords 'nsfw' and 'likes' (left) and 'nsfw' and 'tumblr' (right) in 2018.

have negative weights (93.86% for porn and 92.11% for NSFW). For both of our SoIs, the majority of tokens predict their absence, with only a relative few tokens contributing to our models predicting the presence of either SoI.

Repeating the same methodology for the weights within individual years paints a different picture. Figure 6 shows the same histograms for porn and NSFW, but only for the tokens for the models trained on 2017 data and 2022 data in order to visualize the change in the distribution of tokens' predictive power over time. Taking Figures 6a and 6b, which represents the models trained on 2017 data as an example, we see a clear bimodal distribution, with 83.43% and 84.34% of tokens having a negative weight for porn and NSFW respectively. The large negative peak comes from tokens that very rarely, or never, co-occur with the SoI, resulting in these tokens being assigned a large negative weight. Meanwhile, the smaller peak near zero comes from tokens that do co-occur with the SoI, but which do not strongly predict its presence or absence. The right tail of positively weighted tokens are the ones that contribute the most to a model predicting the presence of a token. Figures 6c and 6d show the same histograms for our SoIs but for the year 2022. In them, we see a similar bimodal distribution, but as the prevalence of both SoIs increased over time, logically more tokens came to predict their presence. This can be seen by the right peak increasing between Figures 6a and 6c and Figures 6b and 6d.

We can also plot the weights of individual tokens for both NSFW and porn. In Figure 7, we visualize the individual tokens that predict the presence of our two SoIs while including the labels of some of the tokens. Overall, we have the same general patterns of clustering we observed in Figure 6, with the largest number of tokens significantly predicting the absence of either SoI, and a smaller number of tokens predicting neither the presence nor absence of the SoI. Interestingly, we find two additional smaller clusters which strongly negatively predict the presence of one SoI but are neutral or slightly positively predict the other.

We can examine these plots for tokens that do not fall neatly on the y=x line to learn about keywords that predict one SoI but not the other. In 2017, we see tokens predicting NSFW but not porn signaling an aspect of the identity of the user (e.g. 'bisexual', 'gender', 'witch') or the type of NSFW content they post (e.g. 'art', 'round', 'wolf', 'daddy'), though the line between these two is not always clear. Meanwhile, for porn, tokens signaling the category of porn posted by a given account tend to fall near the y=x line, since they often refer to identities that an individual posting NSFW content might also have. Away from the y=x line though, we find a number of tokens that seem to exist in an account of an individual that either produces porn (e.g. 'star', 'model', 'goddess') or wants to signal values that tend to co-occur with ideologies against porn (e.g. 'america', 'leader', 'married').

In 2022 for NSFW, we similarly see tokens that signal aspects of the identity of the user or the type of NSFW content they post (e.g. 'athletic', 'bear', 'gentleman', 'adventure', 'shy'). For porn, like in 2017, we see tokens that signal either creating porn (e.g. 'shoot', 'actress', 'wannabe') or having values that tend to co-occur with being against porn (e.g. 'israel', 'maga', 'church'). In 2022 but not in 2017, we see the new tokens that signal porn addiction (e.g. 'addiction', 'addicted', 'addict').

3.2 Logistic Regression Accuracy Analysis

Now, we turn our attention to the predictive capabilities of the models we trained. For each keyword for each year, we used the trained models to predict whether a SoI was present in a testing set of 100,000 held-out bios and recorded the testing accuracy, calculated as the percent of correct classifications. This accuracy can be thought of as corresponding to the 'predictability' of a SoI, where a higher testing accuracy represents a higher 'predictability' of a SoI given tokens in the same bio. Here we look at the predictability of NSFW and porn given all the tokens calculated for the years 2017 and 2022. The testing accuracy for the vast majority of keywords is above 97% due to the models correctly predicting that most keywords are absent from most bios. To render these results interpretable, we compare the testing accuracy, which we use as a 'predictability' of a keyword, against the testing accuracy for every other SoI in our SoI dictionary of 1630 total tokens. Hence, we plot the testing accuracy of porn and NSFW against all other SoIs in our dataset for the year 2017, 2020, and 2022. Figure 8 allows us to visualize the change in predictability for our two target SoIs relative to all other tokens in the SoI dictionary across six years. In 2017, we see that both NSFW and porn are extremely predictable. With each year, both SoIs greatly decrease in predictability, with NSFW going from the 92nd percentile to 10th percentile over the course of the six years. Meanwhile, porn falls from the 97th percentile to the 60th percentile, a stark but smaller decrease in predictability. These results show that both NSFW but especially porn occupied semantic niches on Twitter, with their presence in a bio being predictable based on the other tokens in the bio. With time, both tokens widened in their niches, with NSFW showing a drastic decrease in predictability as the tokens that it co-occurred with increased in variability.

In the same way that we looked at the predictability for each SoI for each year, we can also look at change in test accuracies for NSFW and porn compared to all other tokens from 2017-2022. Figure 9 visualizes the mean change in accuracy across the six years in the dataset. Here, of all the tokens in the SoI dictionary, NSFW is the 11th most negatively sloped token, suggesting that over the span of 2017 to 2022, NSFW showed one of the largest decreases in predictability in our corpus. Similarly, porn being in the 9th percentile of slope shows a similar but lesser decrease in predictability relative to all other SoIs.

3.3 Model Validation

We validate the logistic regression models using test ROC curves and their respective AUC. Figure 10 shows one such test ROC curve for the model for porn for 2022. The large AUC value (~0.80) indicates high model performance, even with a strongly imbalanced sample containing primarily bios lacking any given keyword. The ROC line being closer to the y-axis than the x-axis means that the model rarely misclassifies negatives as positives, instead losing accuracy by misclassifying positives as negatives. While there was variation in the AUC for the models trained based on the SoI and year, it was never less than 0.76 for our keywords between 2017 and 2022, usually hovering around 0.80, indicating a strong capability of our models to distinguish between bios containing and lacking a given SoI.

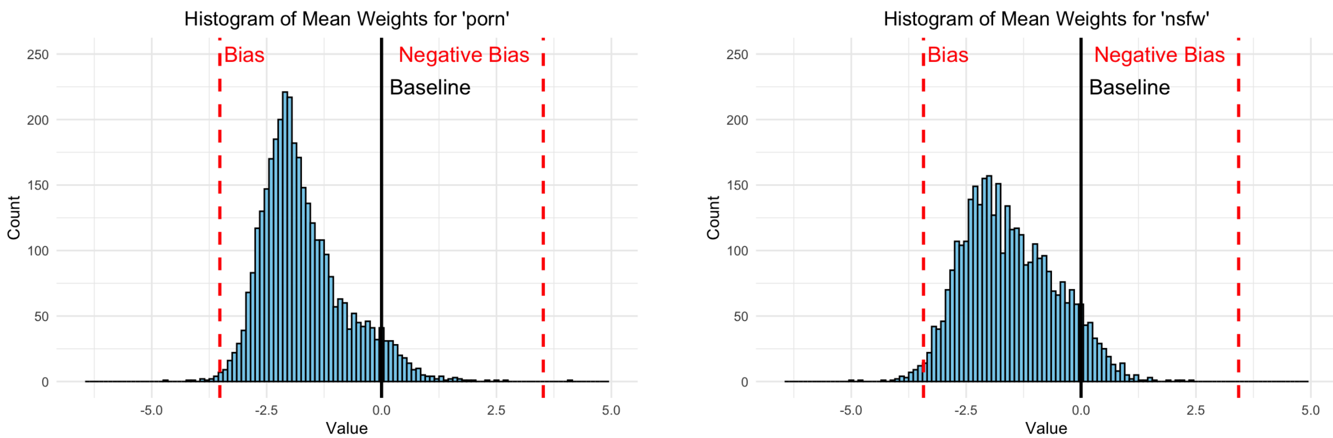
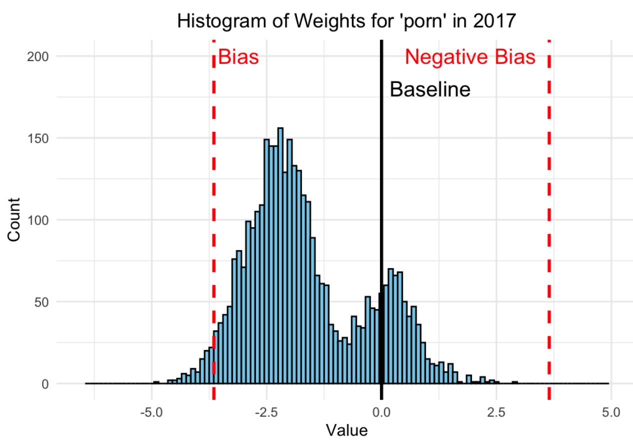
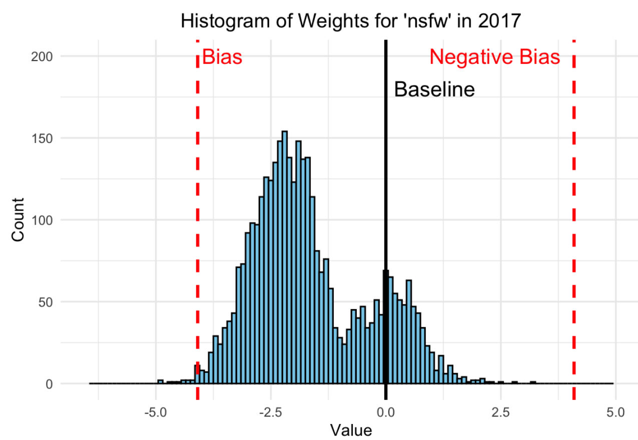


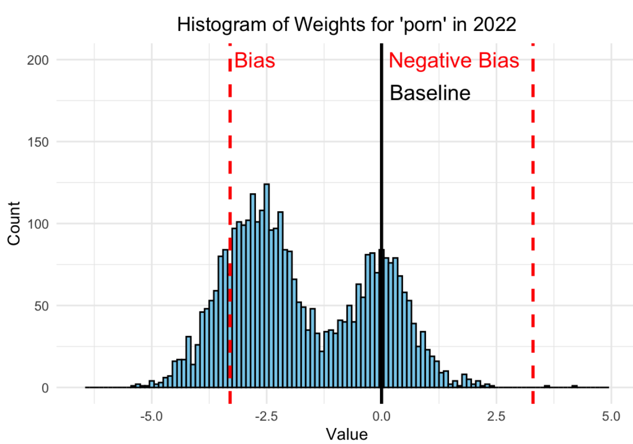
Figure 5: Histogram plotting number of tokens within a given mean value range across 2017–2022. Vertical lines indicate the bias, baseline, and negative bias values.



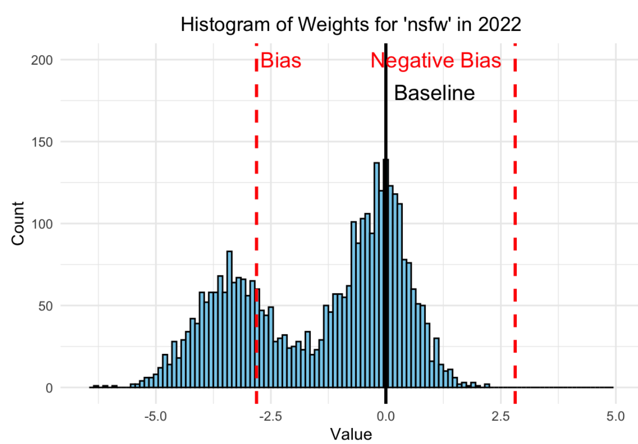
(a) Porn (2017)



(b) NSFW (2017)



(c) Porn (2022)



(d) NSFW (2022)

Figure 6: Histogram plotting number of tokens within a given value range for the years 2017 and 2022. Vertical lines indicate the bias, baseline, and negative bias values.

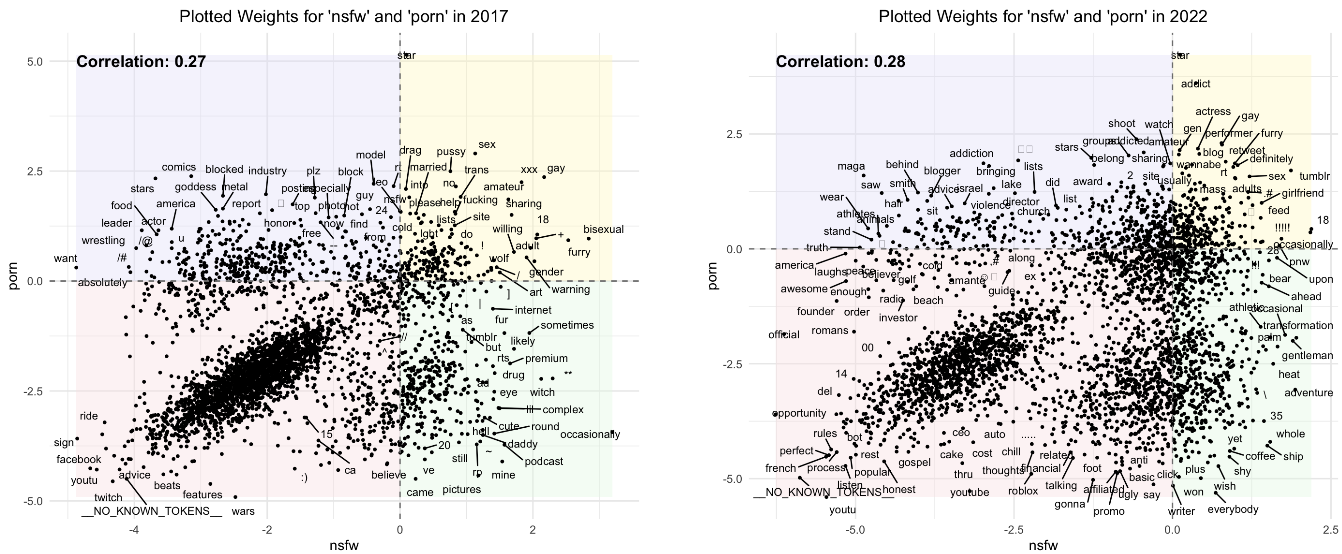
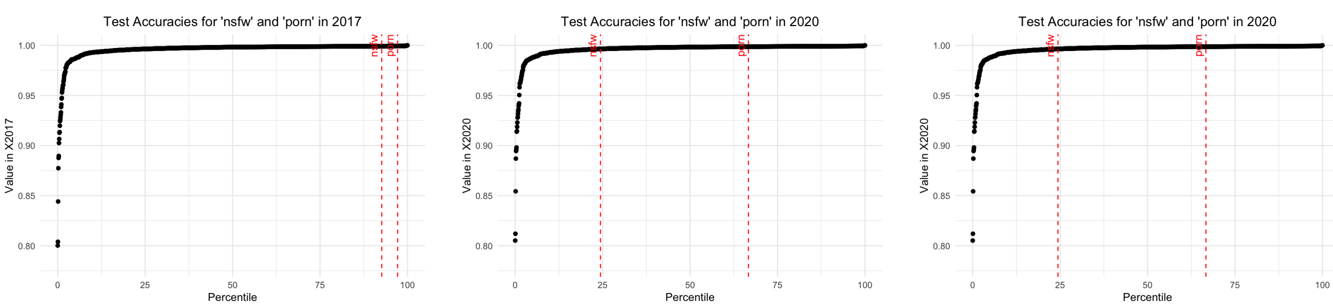


Figure 7: Scatterplots showing token weights for the signifiers 'nsfw' and 'porn' for the years 2017 and 2022, with Pearson correlation between the axes.



(a) 2017 **(b) 2020** **(c) 2022**

Figure 8: Testing accuracy for `NSFW` and `porn` in 2017, 2020, and 2022 compared to all other tokens with prevalence > 7.

4 Discussion

4.1 `NSFW` and `porn` are not synonyms

On Twitter, within user bios, the SoIs `NSFW` and `porn` do not have the same meaning. Though they may refer to similar content, i.e. content that involves nudity, the keyword prediction results in the previous section demonstrated that they are not used in the same contexts. Looking at the keywords that predict `NSFW` and `porn` across the entire span of six years reveals different trends in the identities of the users and the characteristics of the accounts that post pornographic and NSFW content. Tokens that highly predict `porn` tend to be impersonal, focusing around the type of adult content being posted (e.g. 'amateur', 'gay') and the specific performers within the content, which may include the poster themselves (e.g. 'star', 'xxx', 'model', 'performer'). Besides accounts that use their bios as advertising the type of pornographic content they post, we also see `porn` in bios of people with conservative values speaking out against porn (e.g. 'america', 'maga', 'israel'). Meanwhile, tokens that highly predict `NSFW` center around the habits of the poster (e.g. 'occasionally', 'sometimes', 'often') and warnings of the type of content (e.g. '18', '+', 'warning'), showing that posters of NFSW content want their viewers to be aware that NSFW content is not the entirety of their account and that the content they post is sensitive to begin with, an act that is best understood as these NSFW accounts existing and interacting within a larger community of NSFW and non-NSFW accounts. For NSFW posters, adult content is simply one dimension of their own personal Twitter account. This discrepancy is consistent with how porn accounts aim for a wide reach [7], while users posting NSFW content warn others to ensure they are reaching only those who want to see it.

The contrast between pornographic accounts advertising the type of content they want the viewers to see and NSFW accounts warning the viewers of occasional nudity highlights the distinct niches these accounts inhabit on Twitter. These two words are not synonymous, and though they may describe accounts posting

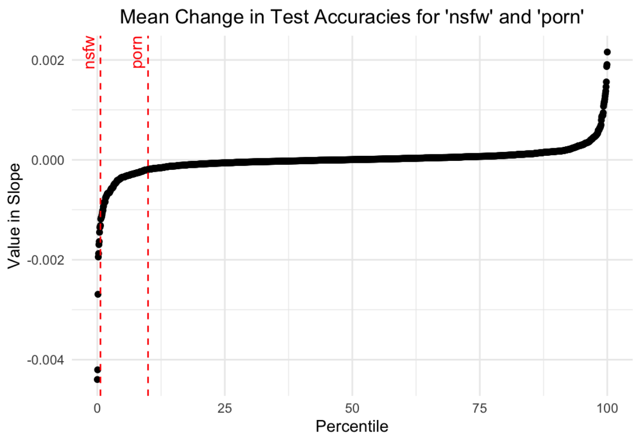


Figure 9: Mean change in testing accuracy between 2017 and 2022 for NSFW and porn compared to all other tokens with prevalence > 7.

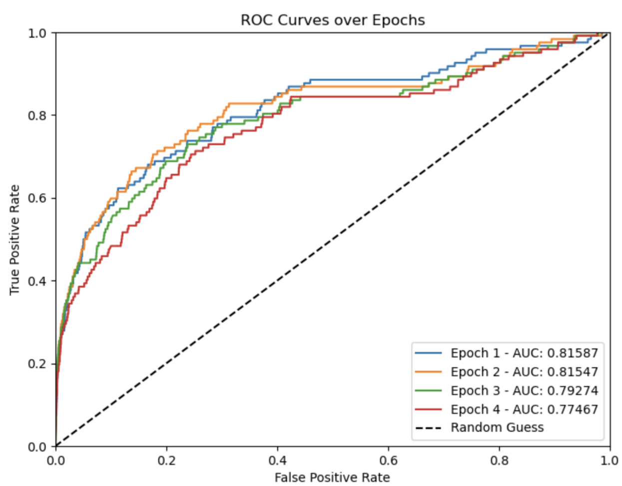


Figure 10: ROC curve for the model trained to predict porn in 2022.

similar content, including one word over another signals that they differ in how they expect their accounts to be interacted with in the Twitter ecosystem.

It must be acknowledged that NSFW and porn do not appear in exclusively sexual contexts, which additionally differentiates their meanings. For example, Figure 7 shows that the keyword 'food' highly predicts porn, but not NSFW, due to accounts posting 'food porn'. Bios with non-sexual uses of porn are a small minority, with 'food porn' never exceeding 0.5 bios per 10,000 between 2017 and 2022.

4.2 NSFW as an indicator of Tumblr NSFW migration

Looking at the tokens that most predict NSFW, we find no tokens that steadily increase in predictive power throughout the full timespan, suggesting a lack of any large, mainstream subcategories emerging on Twitter surrounding NSFW content. Instead, we find expansion of the semantic range of NSFW, as evidenced by the gradual decrease in the value of the negative bias in Figure 4, which represents the value that the sum of token weights must reach before the model predicts the presence of a SoI. A decrease in the bias score suggests that there are fewer words that highly predict NSFW and more words that slightly predict it, as seen by the token 'furry' which steadily lost predictive power as Twitter NSFW content expanded to include more types of content containing nudity. However, we do see a sharp increase in the predictive power of 'likes', which would often appear in the context of users letting others know that their likes are NSFW. On Tumblr, reposts from other accounts are salient and likes can also optionally be set to public, both of which are used for self-presentation and are an integral part of one's Tumblr account [36]. Before June 2024, Twitter likes were similarly public. Mentioning the content another user might find in a poster's likes was a way to extend one's reach and build community beyond the content that one posts. 'Tumblr' also experienced a steady increase in predictive power in 2018 and 2019, correlating with the ban of adult content on Tumblr. The rapid increase in predictive power of 'tumblr' and 'likes' in the year following the Tumblr NSFW ban are consistent with a large migration of users posting NSFW content from Tumblr to Twitter, where users moving to Twitter aimed to preserve similar platform dynamics as Tumblr and its reposting feature by mentioning how their likes were an extension of their self-presentation on Twitter. Finally, the high predictive power of marginalized sexuality and gender identities (e.g. 'gender', 'bisexual') for NSFW – but not porn – is consistent with movement of Tumblr's large queer userbase from Tumblr to Twitter [16].

4.3 porn as nudity for sexual consumption

The changes in the predictability and most predictive tokens for a given SoI reveal how these words may change over time. By looking at how the predictive power evolved for the 10 tokens with the highest average weights for porn and NSFW, we can observe evolution in the niches that porn and NSFW inhabited on Twitter. Beginning with porn, the most striking trend is a sharp increase in the weight for 'addict'. 'Addict' having the second highest predictive power for porn in 2022, preceded only by 'star', shows that Twitter increasingly served as a platform for a community of people who identify as porn addicts. This stands in contrast with the SoI NSFW, where the token 'addict' is at the 88th percentile for predictive power. Words generally synonymous with porn (e.g. 'sex', 'xxx') show a slight negative trend, as pornographic content on Twitter became more common and normalized there was less of a need to reinforce the sexually explicit nature of the account. Taken together, these findings point towards accounts posting porn and having porn in their bios wanting to signal that the nudity they post (or watch, in the case of porn addicts) is for sexual consumption.

4.4 NSFW and porn have both widened in meaning

The decreasing predictability of the signifiers NSFW and porn implies an increase in their scope of meaning. Consider the distribution plotted in Figure 8a. **Predictability varies across identity signifiers.** Signifiers with higher test accuracy (rightward on the x-axis) reliably co-occur (or do not) with predictor tokens. Their context must be narrower than less predictable signifiers.

Predictability varies across time. Signifiers change in test accuracy over years both in value and in relative rank. Sweeping across the panels in Figure 8, one sees that both `NSFW` and `porn` became less predictable (i.e., moved leftward on the x-axis). In every year, our models were consistently able to predict the presence or absence of `porn` better than that of `NSFW`. Even in 2017, prior to the Tumblr NSFW ban, `porn` accounts occupied a smaller niche on Twitter than `NSFW` accounts. This adds to evidence that NSFW accounts are usually personal accounts of an individual who wishes to share multiple facets of themselves online and can be expected to show greater variety in other identifying terms compared to those present within a pornographic account.

The rate at which predictability changes varies across identity signifiers and time. Figure 9 presents the distribution of estimated annual *change* in predictability for all predicted signifiers. Most signifiers retain a stable predictability (i.e., slope values near zero). The negative slope for `porn`—meaning it became less predictable—was somewhat unusual. `NSFW` displays an extreme negative slope. Compared to other signifiers, `NSFW` took on increasingly diverse contexts; keeping signifiers, predictors and methods constant, the model struggled most in later years to correctly place `NSFW` within bios.

We interpret the above results to mean that `NSFW` and `porn` both widened in their meaning over time.

4.5 Additional Considerations and Future Directions

While the Tumblr NSFW ban explains the increase of `NSFW` and `porn` in the first half of our time sample, there are other factors that led to the increase in `NSFW` and `porn` around 2020 that are not addressed here and are left as a direction for future research. Firstly, there was a sharp rise in accounts self-directing pornographic content through OnlyFans or similar adult subscription websites during the COVID-19 pandemic [27]. In fact, in our dataset, we find that the token 'onlyfans' increased in prevalence from 3 bios per 10,000 in 2019 to a peak prevalence of 28 bios per 10,000 in the year 2021, before gradually decreasing to 16 bios per 10,000 in 2023. Additionally, the rise of bots, which are notoriously difficult to discriminate from real accounts [12], possibly artificially affected the prevalence of `NSFW` and `porn`.

4.6 Conclusion

The changes in the predictability of `NSFW` between 2017 to 2022 comport with what we would expect to see from a large-scale migration of users from Tumblr to Twitter. In the years following the Tumblr Adult Content ban, `porn` and `NSFW` accounts expanded in number and proportion of US Twitter users. `porn` and especially `NSFW` decreased significantly in predictability as their semantic spaces expanded. Communities centered around self-expression of marginalized identities found a new home on Twitter after Tumblr was no longer able to foster them. Similarly, new communities that developed more recently centering on taboo topics such as porn addiction grew on Twitter due to its relaxed regulations.

This paper additionally makes available the BioKPD, a large dataset of weights and test accuracies for 1630 signifiers of interest for a token dictionary of size 3535. Future work should investigate the evolution in predictability of other SoIs that may have garnered new meanings during 2017 to 2022. For example, the same methodology may be applied to words relating to how Twitter users describe their race, identity, and community in relation to the Black Lives Matter movement and 2020 protests [8], or to words signaling belonging to a fangroup or online stan community as their followed celebrity's popularity ebbs and flows [21]. Additionally, future work may attempt to use larger datasets as to allow for the inclusion of SoIs that did not make the threshold of a prevalence of 7 bios out of 10,000 across 2017 to 2022. This would allow for the inclusion of more keywords, such as 'onlyfans', a platform that revolutionized self-directed pornographic content on Twitter by both normalizing sexuality and allowing sex workers a way to control their earnings [1] while also opening up new avenues for online sex trafficking [3]. Finally, we urge researchers to further explore the dynamics of identity signifier predictability in new domains using our data and methods.

References

[1] Nazil Mumtaz Al-Mujtahid, Zainun Zainun, and Erwan Efendi. 2024. Phenomenological-Communicative Review of OnlyFans Normalization on Twitter Media. *JHSS (JOURNAL OF HUMANITIES AND SOCIAL STUDIES)* 4, 2 (May 2024), 286–294. doi:10.33751/jhss.v8i2.9755

[2] Aaron Smith and Monica Anderson. 2018. Social Media Use in 2018. https://www.pewresearch.org/internet/2018/03/01/social-media-use-in-2018/

[3] Maricarmen Arenas, Pratheeksha Nair, Reihaneh Rabbany, and Golnoosh Farnadi. 2023. Social Media as a Vector for Escort Ads:A Study on OnlyFans advertisements on Twitter. In *Proceedings of the 15th ACM Web Science Conference 2023 (WebSci '23)*. Association for Computing Machinery, New York, NY, USA, 260–270. doi:10.1145/3578503.3583609

[4] Carolyn Bronstein. 2020. Pornography, Trans Visibility, and the Demise of Tumblr. *TSQ: Transgender Studies Quarterly* 7 (May 2020), 240–254. doi:10.1215/23289252-8143407

[5] Ruepert Jiel Dionisio Cao. 2021. Amateur porn in Filipino Twitter alter community: affordances, commodification, ghettoization, and gay masculinity. *Media International Australia* 179, 1 (May 2021), 52–65. doi:10.1177/1329878X211002845 Publisher: SAGE Publications Ltd.

[6] Mauro Coletto, Luca Maria Aiello, Claudio Lucchese, and Fabrizio Silvestri. 2017. Pornography consumption in Social Media. doi:10.48550/arXiv.1612.08157 arXiv:1612.08157 [cs].

[7] Selat Dewi Seribayu Nordin, Aziz Ahlam Abdul, and Adan Wardatul Hayat. 2023. Aftermath of NSFW in Twitter towards Adolescents. *Forum Komunikasi* 18, 1 (2023), 68–87.

[8] Zackary Okun Dunivin, Harry Yaojun Yan, Jelani Ince, and Fabio Rojas. 2022. Black Lives Matter protests shift public discourse. *Proceedings of the National Academy of Sciences* 119, 10 (March 2022), e2117320119. doi:10.1073/pnas.2117320119 Publisher: Proceedings of the National Academy of Sciences.

[9] Casey Fiesler and Brianna Dym. 2020. Moving Across Lands: Online Platform Migration in Fandom Communities. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1 (May 2020), 42:1–42:25. doi:10.1145/3392847

[10] Margot Hare and Jason Jones. 2023. Slava Ukraini: Exploring Identity Activism in Support of Ukraine via the Ukraine Flag Emoji on Twitter. *Journal of Quantitative Description: Digital Media* 3 (March 2023). doi:10.51685/jqd.2023.005

[11] Matt Hart. 2017. Being naked on the internet: young people's selfies as intimate edgework. *Journal of Youth Studies* 20, 3 (March 2017), 301–315. doi:10.1080/13676261.2016.1212164

[12] Chris Hays, Zachary Schutzman, Manish Raghavan, Erin Walk, and Philipp Zimmer. 2023. Simplistic collection and labeling practices limit the utility of benchmark datasets for Twitter bot detection. In *Proceedings of the ACM web conference 2023*. 3660–3669.

[13] Karen Hellekson and Kristina Busse. 2014. *Fan Fiction and Fan Communities in the Age of the Internet: New Essays.* McFarland. Google-Books-ID: UgZsi_DOKoQC.

[14] Avus Hou and Wen-Lung Shiau. 2019. Understanding Facebook to Instagram migration: a push-pull migration model perspective. *Information Technology &*

People ahead-of-print (July 2019). doi:10.1108/ITP-06-2017-0198

[15] Ujun Jeong, Paras Sheth, Anique Tahir, Faisal Alatawi, H. Russell Bernard, and Huan Liu. 2024. Exploring Platform Migration Patterns between Twitter and Mastodon: A User Behavior Study. doi:10.48550/arXiv.2305.09196 arXiv:2305.09196 [cs].

[16] Jeremy D. Johnson. 2025. *Algorithmic worldmaking the rhetorical craft of networked order.* The University of Alabama Press, Tuscaloosa.

[17] Jason Jones and Irissa Cisternino. 2024. LGBTQ Visibility Measured Consistently and Persistently in US Twitter Bios from 2012 through 2023. *Journal of Quantitative Description: Digital Media* 4 (Oct. 2024). doi:10.51685/jqd.2024.017

[18] Jason Jeffrey Jones. 2021. A dataset for the study of identity at scale: Annual Prevalence of American Twitter Users with specified Token in their Profile Bio 2015–2020. *PLOS ONE* 16, 11 (Nov. 2021), e0260185. doi:10.1371/journal.pone.0260185 Publisher: Public Library of Science.

[19] Christiany Juditha. 2021. The Communication Network of Online Prostitution in Twitter. *Jurnal ASPIKOM* 6, 1 (Jan. 2021), 13–28. doi:10.24329/aspikom.v6i1.826 Number: 1.

[20] Everett S. Lee. 1966. A theory of migration. *Demography* 3, 1 (March 1966), 47–57. doi:10.2307/2060063

[21] Zunera Malik and Sham Haidar. 2023. Online community development through social interaction — K-Pop stan twitter as a community of practice. *Interactive Learning Environments* 31, 2 (Feb. 2023), 733–751. doi:10.1080/10494820.2020.1805773 Publisher: Routledge _eprint: https://doi.org/10.1080/10494820.2020.1805773.

[22] Know Your Meme. 2012. NSFW. https://knowyourmeme.com/memes/nsfw

[23] Susanna Paasonen, Kylie Jarrett, and Ben Light. 2024. *NSFW: Sex, Humor, and Risk in Social Media.* MIT Press. Google-Books-ID: R2bqEAAAQBAJ.

[24] Sophie Pezzutto. 2019. From Porn Performer to Porntropreneur: Online Entrepreneurship, Social Media Branding, and Selfhood in Contemporary Trans Pornography. *AG About Gender* (Jan. 2019). https://www.academia.edu/41168140/From_Porn_Performer_to_Porntropreneur_Online_Entrepreneurship_Social_Media_Branding_and_Selfhood_in_Contemporary_Trans_Pornography

[25] Elena Pilipets and Susanna Paasonen. 2022. Nipples, memes, and algorithmic failure: NSFW critique of Tumblr censorship. *New Media & Society* 24, 6 (June 2022), 1459–1480. doi:10.1177/1461444820979280 Publisher: SAGE Publications.

[26] Nick Rogers and Jason J. Jones. 2021. Using Twitter Bios to Measure Changes in Self-Identity: Are Americans Defining Themselves More Politically Over Time? *Journal of Social Computing* 2, 1 (March 2021), 1–13. doi:10.23919/JSC.2021.0002 Conference Name: Journal of Social Computing.

[27] Valeria Rubattu, Alicja Perdion, and Belinda Brooks-Gordon. 2023. 'Cam Girls and Adult Performers Are Enjoying a Boom in Business': The Reportage on the Pandemic Impact on Virtual Sex Work. *Social Sciences* 12, 2 (2023). doi:10.3390/socsci12020062

[28] Bonnie Ruberg. 2016. Doing it for free: digital labour and the fantasy of amateur online pornography. *Porn Studies* 3, 2 (April 2016), 147–159. doi:10.1080/23268743.2016.1184477 Publisher: Routledge _eprint: https://doi.org/10.1080/23268743.2016.1184477.

[29] Lauren Strapagiel. 2021. Tumblr Says It's Now The Queerest Social Media Platform, But Can It Hold On To That? https://www.buzzfeednews.com/article/laurenstrapagiel/tumblr-says-its-queerest-social-media-platform Section: US-News.

[30] Morgan Sung. 2019. Tumblr loses nearly 30 percent of its page views after banning porn. https://mashable.com/article/tumblr-lost-a-third-of-its-users-after-porn-ban Section: Life.

[31] Katrin Tiidenberg. 2014. Bringing sexy back: Reclaiming the body aesthetic via self-shooting. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 8, 1 (March 2014). doi:10.5817/CP2014-1-3

[32] Katrin Tiidenberg. 2019. Playground in memoriam: missing the pleasures of NSFW tumblr. *Porn Studies* 6, 3 (July 2019), 363–371. doi:10.1080/23268743.2019.1667048 Publisher: Routledge _eprint: https://doi.org/10.1080/23268743.2019.1667048.

[33] Katrin Tiidenberg and Andrew Whelan. 2019. "Not like that, not for that, not by them": social media affordances of critique. *Communication and Critical/Cultural Studies* 16, 2 (April 2019), 83–102. doi:10.1080/14791420.2019.1624797 Publisher: Routledge _eprint: https://doi.org/10.1080/14791420.2019.1624797.

[34] Mojtaba Torkjazi, Reza Rejaie, and Walter Willinger. 2009. *Hot today, gone tomorrow: on the migration of MySpace users.* doi:10.1145/1592665.1592676 Pages: 48.

[35] Liam Tucker and Jason Jones. 2023. Pronoun Lists in Profile Bios Display Increased Prevalence, Systematic Co-Presence with Other Keywords and Network Tie Clustering among US Twitter Users 2015-2022. *Journal of Quantitative Description: Digital Media* 3 (March 2023). doi:10.51685/jqd.2023.003

[36] Ana Valens. 2020. *Tumblr Porn.* Instar Books, New York.

[37] Shuaishuai Wang and Runze Ding. 2022. "Business Inquiries are Welcome": Sex Influencers and the Platformization of Non-normative Media on Twitter. *Television & New Media* 23, 8 (Dec. 2022), 822–839. doi:10.1177/15274764221077666 Publisher: SAGE Publications.

2020		2021	
Token	Weight	Token	Weight
swift	9.149768	swift	11.1060095
justice	2.791501	version	4.249052
tour	2.39677	met	3.2107613
west	1.9664543	wayne	2.7147028
films	1.9113102	stan	2.2780697

Table 3: Top 5 tokens and weights for 2020 and 2021 for the keyword 'taylor'.

mama		mom		mother	
Token	Weight	Token	Weight	Token	Weight
fur	4.23	wife	4.08	grandmother	4.74
wifey	4.10	wifey	3.38	wife	4.48
bear	4.02	grandma	3.07	twins	3.15
wife	3.47	fur	2.87	wifey	3.15
dog	2.25	cat	2.86	two	2.50

Table 4: Top 5 tokens and weights for *mama*, *mom*, and *mother* in 2020 (weights rounded to 2 decimals).

A BioKPD Snapshot

This appendix provides a couple examples of model weights and accuracies to reaffirm the functionality of the methodology presented in this paper. Of the 1630 keywords and 3535 predictive tokens in BioKPD, this appendix briefly looks at four keywords and the keywords with the top/bottom test accuracies.

To begin with, consider for example the token 'taylor', whose top 5 predictive tokens for 2020 and 2021 are shown in Table 3. In 2020, after the killing of Breonna Taylor, users began putting "Justice for Breonna Taylor" (and variants) in their bios to call for police accountability. Meanwhile, in 2021, after Taylor Swift released the rerecorded versions of her albums following the dispute over ownership of her masters, Taylor Swift fans would represent their support for the artist's decision by putting "Taylor's Version" (and variants) in their bio. Investigating the tokens that most predict the keyword 'taylor' during these two years allows us to investigate how Twitter users signaled their anger towards racial injustice and support of Taylor Swift in light of key events, highlighting how the same keyword can index distinct topics across time.

Comparing the weights for the keywords 'mom', 'mama', and 'mother' similarly show differences in the ways that individuals identifying with these three words present themselves online. Table 4 compares the weights for models predicting these words in 2020. Here, we notice users identifying themselves as 'mama' being most predictable by the mention of a pet (i.e. "fur mama"), something users identifying themselves as 'mother' do not do. Similarly, we see that users with 'mama' in their bio don't tend to identify as 'grandma' or 'grandmother' like users with 'mom' or 'mother'. Further conclusions can be drawn from the data presented here, such as the relative weights of 'wifey' and 'wife', but this snapshot is simply meant to show the potential utility of the BioKPD.

Turning our attention towards the testing accuracy, we can see which tokens are the most and least predictable by our model.

Top 5 mean accuracies		Bottom 5 mean accuracies	
Token	Mean accuracy	Token	Mean accuracy
HTTPS_T_CO	1.000000	.	0.801155
AT_GMAIL_DOT_COM	1.000000	,	0.806825
U+0336	0.999873	and	0.848742
U+064E	0.999838	the	0.882855
U+0650	0.999805	of	0.889918

Table 5: Top and bottom 5 signifiers by mean test accuracy.

Table 5 gives the 10 most and least predictable tokens and their corresponding weights for the testing accuracies averaged out over the years 2017 to 2022. The most predictable tokens are parts of URLs or email addresses, which always appear in a preset template, or certain combining diacritics, which appear with a closed number of other characters (e.g. U+064E and U+0650 are Arabic script diacritics). Meanwhile, the tokens with the lowest test accuracies are tokens whose appearance does not depend on the presence of any other specific tokens, such as punctuation and three of the most common words in English.