Predictability of 'NSFW' and 'Porn' in Twitter Bios

University of Washington, Seattle

Alessio Tosolini, Jason Jones

Stony Brook University

Abstract

Twitter has increasingly become a popular platform for NSFW (not safe for work) and pornographic content. Users may include these keywords in their bios to express to other users that their profiles contain explicit content. To see which tokens predicted a bio containing the keyword 'nsfw' and/or 'porn', **we created a linear regression model which predicts whether a bio contains a given keyword given the remaining unique tokens in the bio**. The weights assigned to each token represented the ability of that token to predict the presence of a given keyword, a measure we called 'predictive power'. We analyzed which tokens changed in predictive power over time and which keywords were most predictable by the linear regression model.

- 'nsfw' is a highly predictable keyword, with tokens relating to the poster's identity increasing in predictive power and tokens relating to pornographic content surprisingly decreasing in predictive power.
- 'porn' is an averagely predictable keyword, with tokens relating to the type of porn and porn addiction increasing in predictive power while semantically broad pornographic words decreasing in predictive power.

These results suggest that users posting content with 'nsfw' in their bios do not post it with the same intent as users posting content with 'porn' in their bios. 'nsfw' bios increasingly post nudity in the context of their identity as opposed to porn, while 'porn' bios increasingly post nudity to appeal to a niche on Twitter.

Predictability of Tokens - Methodology

A linear regression model is used to predict whether a certain keyword appears in a given bio.

Data Collection:

- For each year analyzed, bios of **1% of tweeters** were collected randomly into a dataset representative of that year's bios, consistent with past studies on a larger scale (Rogers 2021).
- Of those tweets, 200,000 were selected randomly, 2% of which **contained the keyword** (either 'nsfw' or 'porn') the model would be trained to predict.

Tokenization:

- Each string of letters separated by a whitespace or punctuation was considered a separate **token**.
- All the unique **tokens** in a bio were collected into a **bag of words (BoW)** representing that bio.
- The keywords were filtered from the BoW.
- **Tokens** with a **prevalence** of less than 3 were discarded, since they led to **overfitting**.

Linear Regression Model:

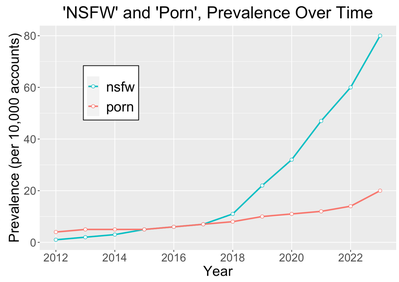
- A linear regression model was trained on the **BoW** representations for each unique token in each bio.
- A model would face a **binary classification task**:
- Given a **BoW** representing a bio, does that bio also contain the keyword of interest?

Regularization was used with lambda=10e-5 and the model performance was calculated with the test *critical success index*, used to emphasize true positives over true negatives.

'NSFW' and 'Porn' in Bios over Time

Prevalence Over Time:

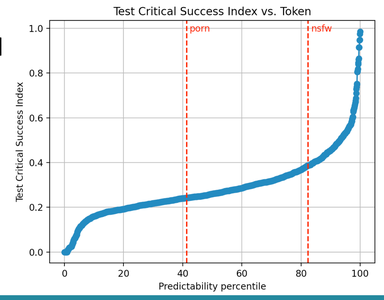
- Since 2012, both 'nsfw' and 'porn' have seen a **steady increase** in prevalence in Twitter bios.
- Since 2018, 'nsfw' has increased **significantly faster** than 'porn' in prevalence, possibly due to the controversial 2018 ban of pornographic content on Tumblr (Pilipets 2022).



Predictability Index

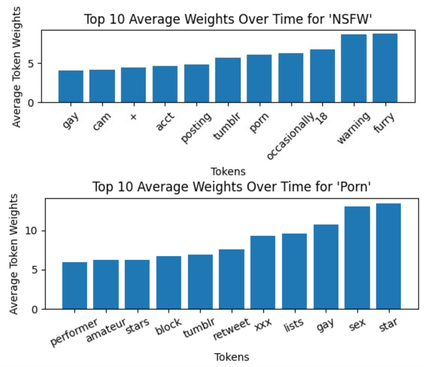
Predictability Index:

- For every keyword, we **train a linear regression model** to predict whether bios contain it.
- The test critical success index is used as a measure of a token's **predictability**.
- Given bios from the year 2022, 'NSFW' is **more predictable** than 'porn'.



Best Predictors

For every keyword, we trained models over 2017-2022 to **find the tokens which best predict the presence of a given keyword** and plot the average best predictors.



References

Rogers, N., & Jones, J.J. (2021). Using Twitter Bios to Measure Changes in Self-Identity: Are Americans Defining Themselves More Politically Over Time? J. Soc. Comput., 2, 1-13.

Pilipets, E., & Paasonen, S. (2022). Nipples, memes, and algorithmic failure: NSFW critique of Tumblr censorship. *New Media & Society*, *24*(6), 1459–1480.

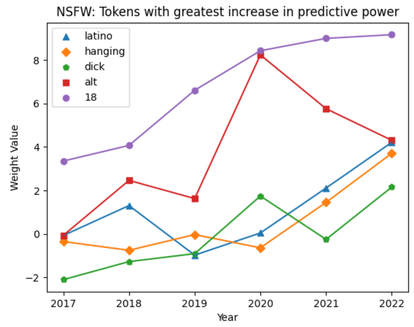
Change in Predictive Power

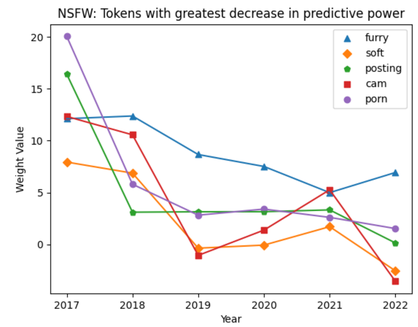
Predictive Power:

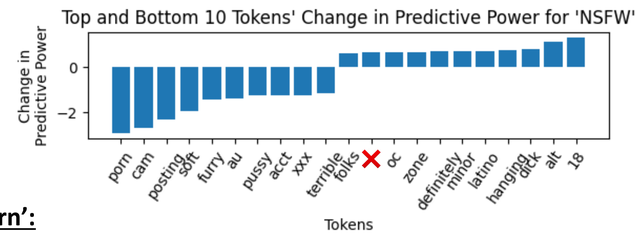
- When creating a model that predicts whether a bio contains a keyword, each **token is assigned a weight** representing how much it **indicates the keyword's presence** in the bio given the year.
- A unique model was run on bios from 2017-2022, the weights for each token normalized, and the tokens with **the greatest increase and decrease in weights** were plotted.

Results for 'nsfw':

- Tokens relating to the **user's identity and content** ('latino', 'dick'), **other accounts** ('alt') and **exclusion of minors** ('18') increased in predictive power.
- Tokens relating to **pornographic content** ('soft', 'cam', 'porn') the **user's sexual preferences** ('furry') and decreased in predictive power.

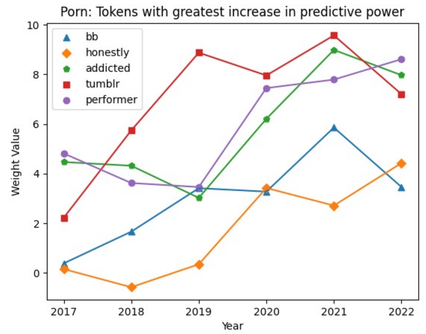


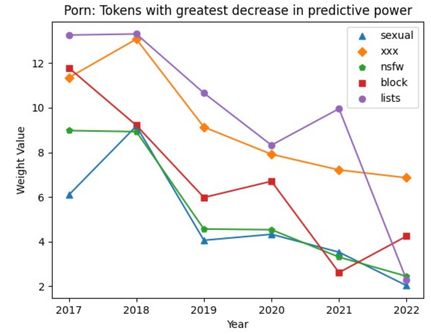


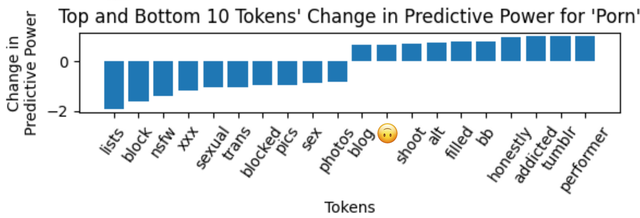


Results for 'porn':

- Tokens relating to **the type or origin of porn** ('bb' short for bareback, 'tumblr') and the **porn addiction** ('honestly', 'addicted') increased in predictive power.
- Tokens **vaguely semantically correlated to pornographic content** ('sexual', 'xxx', 'nsfw') decreased in predictive power, likely due to the redundancy in meaning with 'porn'.







This research was supported by NSF Award #1950052 Data + Computing = Discovery! REU (Institute of Advanced Computational Science at Stony Brook University)

tosolini@uw.edu