Star Wars Twitter
Biases in social media discussions of characters and relationships

Bethany Lacina

Associate Professor of Political Science,
University of Rochester
blacina@ur.rochester.edu

This document accompanies an article I wrote for the *Washington Post* Monkey Cage Blog in April 2019. I analyzed data from Star Wars fan Twitter and argued that (1) posters use more offensive language to talk about women and minority sequel trilogy characters than they use to talk about older, white original trilogy characters and (2) posters are particularly likely to use toxic language in posts that mention romantic storylines ('ships).

1 Where the tweets come from

The tweets I analyzed are from Twitter's Historical Search Premium API.[1] The API searches for tweets and presents them in reverse chronological order. By contrast, a search on Twitter's website provides results according to relevance.

[1] API stands for "application program interface." An API is a URL that transmits unformatted data instead of transmitting a website with a graphical interface.

2 Measuring toxic language

The analysis of toxic speech is based on an algorithm developed by Google called the Perspective API. The algorithm makers describe the concept of toxic speech as follows:

> This model was trained by asking people to rate internet comments on a scale from “Very toxic” to “Very healthy” contribution. Toxic is defined as... “a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion.”

Text flagged as toxic includes slurs of all kinds, insults, mentions of violence and self-harm, and obscenity. The homepage for the API has a section titled “writing experiment” where you can type some text and then see how the algorithm rates its toxicity.

3 Comparing use of toxic language according to the character(s) being discussed

The analysis in this section is based on 7,500 tweets from the period December 9, 2017 (premier of *The Last Jedi*) to May 27, 2018. 1,000 tweets were based on interval sampling, 6,500 on random sampling over the entire period. The search terms were “star wars”, “star war”, and “last jedi”; these searches are case insensitive.

I categorized whether tweets discussed sequel or original trilogy protagonists. The categories are non-exclusive. The rate of toxic language in each category is given in Table 1.

Table 1: Percentage of tweets rated as toxic in Star Wars Twitter disaggregated by character(s) discussed

Category of tweet†	Tweets rated as toxic*
All tweets	6.5%
Tweets mentioning original trilogy protagonists:	
Luke	10%
Han	7.9%
Leia	7.5%
One or more of Luke/Han/Leia	8.2%
Tweets mentioning sequel trilogy protagonists:	
Rose Tico	19%
Finn	15%
Rey	13%
Poe Dameron	12%
One or more of Rose/Finn/Rey/Poe	13%
Tweets mentioning original and sequel trilogy protagonists:	
One or more of Luke/Han/Leia AND one or more of Rose/Finn/Rey/Poe	22%

† Categories are not exclusive.
*Toxicity score of ≥ 0.5.

4 Shipping Twitter versus non-shipping Twitter

The sample described above was sorted according to whether one of the following relationships was mentioned in the tweet: Damerey (Poe Dameron and Rey); Finnrey (Finn and Rey); Finnrose (Finn and Rose); Gingerpilot (Armitage Hux and Poe Dameron); Kylux (Kylo Ren and Armitage Hux); Reylo (Rey and Kylo Ren); Stormpilot (Finn and Poe Dameron). I used that sorting to compare the level of toxic language in shipping and non-shipping tweets gathered in the same time frame.

I also hand checked a sample of shipping and the non-shipping tweets marked as toxic. I did this to adjust for several types of “false positives” that come up repeatedly in shipping Twitter: (1) insults (e.g., “trash”) used as terms of endearment; (2) sexually explicit language; (3) words related to homosexuality (e.g., “gay”) that the algorithm has misinterpreted as insults. I changed the toxic rating to non-toxic in places where a tweet (shipping-related or not) did not appear to be intended to provoke or abuse. If I was uncertain, I left the coding unchanged. I did not change any codings of non-toxic to toxic.

Based on the rate of false positives in this hand-coding, I adjusted my estimates of toxicity in shipping and non-shipping tweets in the full sample. The adjustment reduced the disparity between shipping and non-shipping Twitter.

As reported in the article, I estimated 4% of non-shipping Twitter contained toxic language while about 9% of shipping Twitter used toxic language.

5 Comparing use of toxic language according to the romantic relationship(s) being discussed

To estimate levels of toxic language in discussions of different romantic pairings, I drew a large sample of tweets that mentioned at least one ship. I drew 90,876 tweets from the period December 1, 2015 to February 14, 2019, selected in randomized batches of 500. The search was for the following ships: Damerey (Poe Dameron and Rey); Finnrey (Finn and Rey); Finnrose (Finn and Rose); Gingerpilot (Armitage Hux and Poe Dameron); Kylux (Kylo Ren and Armitage Hux); Reylo (Rey and Kylo Ren); and Stormpilot (Finn and Poe Dameron). For each pairing, I searched for a variety of combinations of the characters’ names. For example, the Damerey search was for “damerey”, “poe/rey”, “rey/poe”, etc. Only the “gingerpilot” pairing did not yield sufficient results for analysis. The Finnrose pairing also returned a small sample, in part because the sampling period included the years before the character of Rose Tico was introduced.

I categorized tweets according to which ship(s) they mentioned. The categories are non-exclusive.

As described above, I hand-coding a selection of the “toxic” tweets flagged for category. I estimated the rate of false positives based on that hand-coding and down-weighted the estimates of toxic speech accordingly. Results are in Table 2.

Table 2: Tweets rated as toxic in Star Wars shipping Twitter by relationship discussed

Category of tweet†	Tweets rated as toxic*
All tweets in the shipping sample	8.7%
Tweets by relationship mentioned:	
Finnrose‡	13%
Reylo	10%
Finnrey	8.5%
Kylux	8.0%
Damerey	7.5%
Stormpilot	6.3%

† Categories are not exclusive.

*Toxicity score of ≥ 0.5.

‡ Accounts for $<$1% of shipping tweets collected