Star Wars Twitter analysis

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 September 2018. I analyzed data from Star Wars fan Twitter and argued that (1) offensive language and hate speech have a modest but clear presence, (2) abusive posts are not due to automated accounts (bots), and (3) posters use more profanity and slurs to talk *about* women and minorities and to talk *to* female fans.

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

[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.

contrast, a search on Twitter's website provides results according to relevance.

I conducted four searches for tweets in reverse chronological order from before June 2, 2018 at 1700 GMT, about when Kelly Marie Tran quit Instagram. The four batches of tweets were: (1) results of a keyword search for Star Wars (SW) or *The Last Jedi* (TLJ); (2) results of a keyword search for Kelly Marie Tran or Rose Tico (the name of Tran's TLJ character); (3) mentions and replies sent to male fan podcasters; (4) mentions and replies sent to female fan podcasters.

1.1 How much is missing?

I estimate about 5% of Star Wars Twitter is not available in the Twitter archives because the tweets have been deleted. Deleted tweets may have violated community standards, such as rules against abusive language or spam. If someone deletes their Twitter account, their tweets become unavailable via the historical search API.

My 5% estimate is based on asking the API for count data. According to Twitter, "The counts . . . reflect the number of Tweets that occurred and do not reflect any later compliance events (deletions, scrub geos). Some Tweets counted may not be available" to download via an API search.[2] I compared the number of Tweets available to download versus Twitter's total counts of tweets.

1.2 How I found the fan accounts

Part of my article examines tweets sent to male and female fans. These tweets are "replies" to or "mentions" of fan podcast Twitter accounts.

I began with the 23 podcast links on the StarWars.com community webpage.[3] I noted show and host Twitter accounts listed on each podcast webpage. I divided host accounts into male-run and female-run. A show account is included in my female-run category if any of the hosts were female.[4] After this process, I had 37 male-run accounts[5] and 17 female-run accounts.[6] I added 9 more female-run

[2] See Twitter's developers' guide.

[3] Until July 2018, the Rebel Force Radio podcast was linked to StarWars.com. The hosts have since deleted their Twitter account, making their tweets unavailable via API.

[4] There was only one co-ed hosting team although several of the male-run shows have regular female guest hosts.

[5] REDACTED

[6] REDACTED

accounts[7] from podcasts that had been recommended by Lucasfilm employees on Twitter. None of the podcasters are celebrities or Lucasfilm employees.

Why podcasts from StarWars.com and Lucasfilm employees' recommendations? A recurring debate on Star Wars Twitter is about the relationship between Lucasfilm and Star Wars media. All of the podcasters on my list are vulnerable to the charge of being indebted to Lucasfilm. If the female podcasters are treated differently on Twitter it is not because they are more or less aligned with Lucasfilm.

1.3 How did I know if the fans were male or female?

I guessed whether individuals were male or female based on their pictures and names. I excluded one account that had no picture and a sex-neutral name. Some of my guesses may be wrong. I have no information about podcasters' gender identity, sexuality, race, heritage, religion, and so forth. I also do not know their opinions about the latest Star Wars films.

My reliance on guesswork means I am in the same position as a stranger on Twitter. My implicit question is whether Twitter accounts that seem to be run by women get different posts than accounts that seem to be run by men.

2 Measuring offensive language and hate speech

The analysis of offensive language and hate speech are based on two different algorithms developed at the Social Dynamics Lab (SDL) at Cornell University.

To measure offensive language, I used an SDL tool called HateSonar. For hate speech detection, I used another SDL program developed by Thomas Davidson and coauthors.[8]

Offensive language is profanity. Hate speech includes slurs and threats of violence. The SDL models are more than a dictionary of offensive and hateful terms. Text can be flagged as hate speech or offensive speech even if it does not include explicit language and vice versa.

[7] REDACTED

[8] Davidson, T., D. Warmsley, M. Macy, and I. Weber. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. International AAAI Conference on Web and Social Media, North America, May 2017.

2.1 False negatives and false positives

I was most concerned about overestimating the presence of hate speech. I read all of the tweets coded as hate speech and deleted obvious false positives. I did not hand code for false negatives. The amount of hate speech I report is more likely to be an underestimate than an overestimate.

The most common reason for a false positive was that part of a tweet was meant to be a quotation of someone else. A denunciation of hate speech could be recorded as hate speech. The algorithm also stumbled over fictional violence.

2.2 What exactly is "hate speech"?

The SLC uses the term "hate speech" to describe what their tools measure. I deferred to them. However, the term "hate speech" is a potential source of confusion.

Hate speech does not have a legal definition in the USA. The FBI's website has the following discussion of hate crimes:

> For the purposes of collecting statistics, the FBI has defined a hate crime as a "criminal offense[9] against a person or property motivated in whole or in part by an offender's bias against a race, religion, disability, sexual orientation, ethnicity, gender, or gender identity."

The FBI definition of hate crime does not include a case in which an offender is biased against an ideology or political orientation. If hate speech is defined along similar lines, saying incendiary things about a political or ideological group would not qualify.

Why does this matter? Twitter users call each other racist, sexist, homophobic, bigoted, and so forth. If they are not subtle about it, the SDL algorithm may code their language as hate speech. But labels like "racist" or "homophobic" describe attitudes (including unconscious beliefs) and do not fall into the ascriptive categories in the FBI definition above.

It is useful for me that the SLC algorithm captures people berating each other for their presumed beliefs. Those interactions are part of how popular culture becomes politicized. Calling this "hate speech" is debatable, however.

[9] I.e., an action that is a crime according other laws. Examples include physical assault and homicide.

3 Is the backlash due to bots?

Automated Twitter "bots" were used prior to the 2016 US national elections and the Brexit referendum in order to create a misleading sense of public opinion. *Star Wars* fans sometimes wonder whether similar "astroturfing" is responsible for online anger.

To check, I used the Botometer Python program from the Observatory on Social Media at Indiana University. Botometer estimates that about 4.4% of the results of a Twitter search for *Star Wars* were produced by bots.[10] For comparison, one study suggests 14% of Brexit Twitter traffic came from bots.[11]

The automated accounts on *Star Wars* Twitter seem benign. Bots use less offensive language than other posters and did not generate any of the hate speech in my sample (Table 1).

Table 1: Offensive speech and hate speech by automated accounts on Star Wars Twitter

	Offensive language	Hate speech	N
Tweets by unlikely bots			
Without retweets	6.2%	1.1%	1169
Including retweets	5.7%	1.3%	2388
Tweets by likely bots			
Without retweets	0%	0%	67
Including retweets	2.7%	0%	112

[10] Botometer scores a Twitter account from zero to one. 4.4% of the queried accounts had a Botometer score of 0.5 or more.

[11] Howard, P.N. and B. Kollanyi. 2016. "Bots, #StrongerIn, and #Brexit: Computational Propaganda during the UK-EU Referendum." SSRN Abstract 2798311.

4 Monkey Cage results

Table 2 is the data used in my Monkey Cage blog post. The columns display the share of tweets classified as having offensive language or hate speech. The rows represent four groups of tweets: (1) results of a keyword search for Star Wars or *The Last Jedi*; (2) results of a keyword search for Rose Tico or Kelly Marie Tran (RT/KMT); (3) mentions and replies sent to male fan podcasters; (4) mentions and replies sent to female fan podcasters.

Tweets are also broken down within each category according to whether they expressed a negative or positive sentiment. I used a sentiment analysis tool called VADER[12] to measure whether a tweet says something positive or something negative about its subject. The scale runs from negative one to positive one.

Negativity is a much broader characterization than offensive language or hate speech. For example, the statement "the flu is bad" is a negative sentiment and has a VADER score of -0.76. This sentence is not offensive or hateful and is not rated as such by HateSonar.

Interestingly, tweets about RT/KMT were less likely to have negative sentiment than tweets about Star Wars in general (40% versus 70%). People were negative or critical more frequently in the general Star Wars discussion than in the RT/KMT discussion, even though there was more offensive and hate speech in the latter.

Similarly, even though female fans receive more hate speech in their mentions, female podcasters received fewer tweets with negative sentiment: 20% versus 36%.

5 Is Rose Tico just unpopular?

In my article, I argue that people use different language people use to talk about Star Wars and to talk about Rose Tico/Kelly Marie Tran. About 6% of SW/TLJ posts use offensive language. The rate of offensive language in posts about RT/KMT is double that: 12%. Hate speech was about 60% more common in tweets about RT/KMT compared to other Star Wars topics: 1.8% versus 1.1%.

This difference is not due to Rose Tico being an unpopular character. If anything, Twitter about KMT/RT tends to be more positive than other Star Wars Twit-

[12] Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

Table 2: Offensive language and hate speech in Star Wars Twitter

	Offensive language	Hate speech	N
Keyword searches			
Star Wars/*The Last Jedi*			
All tweets	5.8%	1.1%	1236
Only negative tweets	6.9%	1.2%	903
Negative tweets as share of all tweets: 73%			
Rose Tico/Kelly Marie Tran			
All tweets	12.1%	1.8%	1516
Only negative tweets	18%	3.8%	577
Negative tweets as share of all tweets: 38%			
Tweets sent to fan podcasters (replies and mentions)			
Men			
All tweets	7.7%	0.22%	1337
Only negative tweets	12%	0.63%	480
Negative tweets as share of all tweets: 36%			
Women			
All tweets	8.5%	0.35%	1429
Only negative tweets	17%	1.4%	292
Negative tweets as share of all tweets: 20%			

ter, as noted in the previous discussion.

Table 2 shows that the differences in abusive language are even larger in a comparison of only negative posts. About 7% of negative posts about SW/TLJ use profanity. Offensive language is 2.5 times more common (18%) in negative posts about RT/KMT. Hate speech is four times more common when fans complain about Star Wars' first non-white female lead (4%) compared to when they complain about other parts of the franchise (1.2%).

Negative posts sent to female fans are also more likely to have offensive language and hate speech than negative mentions for male fans. 17% of negative tweets sent to women contain offensive language, twice the rate (8.5%) in tweets to male fans. Hate speech is also about twice as common in a negative tweet sent to a female fan compared to a negative tweet sent to a male fan (1.4% versus 0.63%).

6 Extended results on offensive language

Since conducting the research reported in The Monkey Cage, I have analyzed a larger sample of Star Wars tweets focusing on offensive language. I drew these tweets from randomly selected days in the period December 15, 2017 to May 27, 2018. I used the HateSonar algorithm to classify offensive language.

Table 3 shows a comparison between tweets drawn from an extended keyword search for Star Wars/*The Last Jedi* ($n = 8000$) and an extended Kelly Marie Tran/Rose Tico search ($n = 5427$). The columns show what percentage of tweets scored above 0.5 on the 0 to 1 scale for offensive language. The extended sample has lower rates of offensive speech than the late May/early June sample. The latter period seems to have been especially bad on Star Wars Twitter.

The results are very similar to the earlier analysis. There was more offensive language in the RT/KMT discussion compared to tweets on other Star Wars topics. This difference was most pronounced in a comparison of negative tweets about RT/KMT to negative tweets about SW/TLJ.

Table 3 also summarizes a sample of 7,767 replies to male podcasters' accounts and 7,786 replies to female podcasters' over the period December 15, 2017 to May 27, 2018.[13] The difference in the use of offensive language is small: 4% of replies to women and 3% of replies to men use offensive language. There is a larger difference in the amount of offensive language in negative tweets. 6% of negative replies sent to men use offensive language and 10% of negative replies sent to women use offensive language.

7 Why measuring harassment is difficult

For a variety of reasons, it is more difficult to capture the amount of harassment directed at a particular Twitter user than it is to characterize how people talk about a topic.

It is useful to look at an example. In July 2018, a Lucasfilm employee, Andi Gutierrez, was targeted on Twitter. The incident prompted StarWars.com to cut ties with one of the oldest fan-run media sources, Rebel Force Radio. The incident is summarized here:

> Andi Gutierrez is a familiar face to anyone who watches the Star Wars YouTube channel, co-hosting The Star Wars Show and Rebels Recon.

[13] Unlike the data in Table 2, this sample does not include mentions other than replies.

Table 3: Extended analysis of offensive language in Star Wars Twitter

	Offensive language	N
Keyword searches		
Star Wars/*The Last Jedi*		
All tweets	3.1%	8,000
Only negative tweets	3.4%	5,008
Negative tweets as share of all tweets: 63%		
Rose Tico/Kelly Marie Tran		
All tweets	4.8%	5,427
Only negative tweets	8.4%	1,664
Negative tweets as share of all tweets: 31%		
Tweets in reply to fan podcasters		
Men		
All tweets	3.1%	7,767
Only negative tweets	6%	1,874
Negative tweets as share of all tweets: 24%		
Women		
All tweets	4.0%	7,786
Only negative tweets	10%	1,546
Negative tweets as share of all tweets: 20%		

> In other words, shes a public-facing Star Wars employee with no creative role in the movies. Fan podcast Rebel Force Radio singled her out on Monday [July 9, 2018] for posting a selfie with a "Fanboy Tears" mug, prompting a wave of criticism from their followers—and a deluge of support from other Star Wars fans and creators.[14]

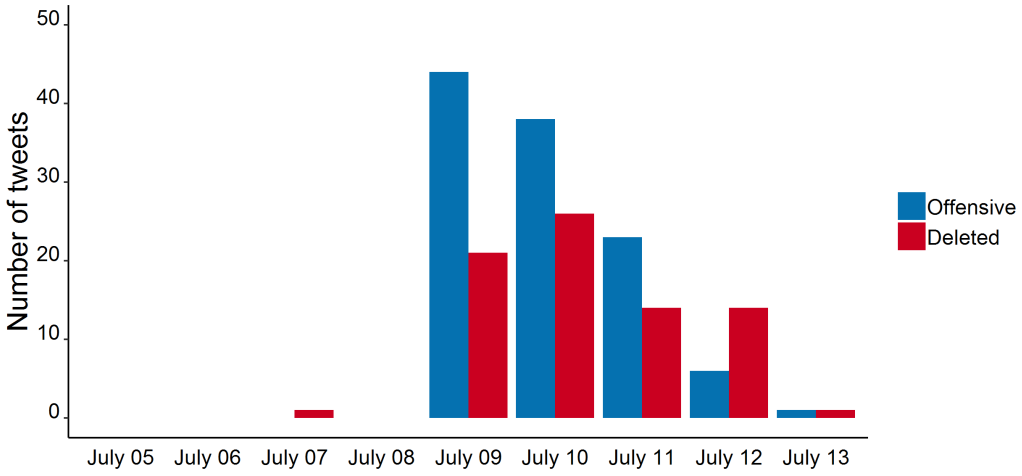
Gutierrez's picture was circulated with an incorrect date, making it appear more recent than it is.[15]

Figure 1 is based on replies and mentions to Gutierrez's Twitter account between July 5 and July 13. Her photo-with-mug was circulated in the middle of

[14] Baker-Whitelaw, G. July 11, 2018. "'Star Wars' backlash is the new Gamergate." *The Daily Dot*.

[15] Compare this thread with screen shots of tweets here.

Figure 1: Offensive and deleted tweets sent to Andi Gutierrez (@DeeGoots), July 5–13, 2018



Mislabeled picture of Gutierrez with ``Fanboy Tears'' mug tweeted on July 9, 2018.
Deleted tweets were removed by their author or violated Twitter community standards.

Figure: Lacina, Univ. of Rochester, goo.gl/YYWk4u. Data: Twitter.

that period, July 9.

The blue bars count the number of offensive tweets that Gutierrez received per day.[16] The red bars count the number of tweets that mentioned her account but were subsequently deleted, either by their author or because they violated Twitter community standards.

The graph highlights one difficulty with capturing harassment on Twitter. Harassment frequently occurs in short episodes. Gutierrez's account received no offensive tweets in the days before July 9, then twenty or more for the next three days. Activity in her account returned to normal on July 13. Since harassment is not spread evenly through time but bunched up at a few points, it is easy to miss an entire episode when drawing a random sample of tweets.

A second difficulty is deleted tweets. Gutierrez received a lot of now-deleted

[16] I.e., tweets that score 0.5 or higher on HateSonar's offensive language scale. None of the available (not deleted) tweets from this period are classified as hate speech by that algorithm.

tweets at the same time the offensive tweets sent to her account surged. 4.5% of tweets to her account sent between July 9 to July 12 are now deleted. Offensive speech, hate speech, and threats could be missed entirely because Twitter does not make this kind of data available for research.

7.1 Offensive but complimentary

Another kind of problem arises with defining what sort of tweets are harassment. There are at least two kinds of exchanges on Twitter that use offensive language but don't fit an ordinary definition of harassment.

First, some replies and mentions sent to a fan account contain strong language but are intended to compliment or agree with the account owner. Below is a post by @swankmotron and one reply. (The examples are chosen for illustration. They may or may not be part of the quantitative analysis.)

> The comments beneath this @StarWars tweet is a good place to start developing your block list for a better twitter experience. - @swankmotron, June 2, 2018.

> Wow…you weren't kidding. That's just ridiculous. It will never cease to amaze me how people will say things on social media that they wouldn't be caught dead saying in person. Keyboard warriorism sucks. - @NHCPodcast, June 2, 2018.

From the point of view of a reader, these posts are potentially offensive or hurtful. They insult other accounts, which could be identified after some legwork. However, @NHCPodcast's reply agrees the original post and is not intended to antagonize @swankmotron. An account with many interactions like this is arguably an unpleasant forum. But the account owner is not really being harassed.

This distinction could be important in determining how different groups are treated on Twitter. Possibly, people use more offensive language when they are agreeing with or complimenting men (or women). That pattern would lead to more offensive language in one sex's mentions even though there was no difference in the amount of harassment.

7.2 Flame wars

A second kind of interaction that is not harassment, ordinarily defined, is the two-sided flame war. People receive strongly worded responses to strongly worded

tweets. All three tweets in this exchange between @DoctorRagnarok and @DocBullfrog666 are combative:

> Kelly Marie Tran gets attacked. Kathleen Kennedy has her legacy questioned. Colin Trevorrow's talent is ignored. Rian Johnson gets death threats. George Lucas, Jake Lloyd, Ahmed Best, and Ashley Eckstein get harassed for years. Solo underperforms. Star Wars fans suck. - @DoctorRagnarok, May 31, 2018.

> Star Wars is dead to me. Never will spend a single dollar on it again. Garbage. - @DocBullfrog666, June 2, 2018.

> Cool. Now go jerk off to a Geeks and Gamers video. Man baby. - @DoctorRagnarok, June 2, 2018.

From the point of view of a reader, the entire exchange might be offensive or hateful. But @DoctorRagnarok is not the target of unprovoked hostility. His original post contains an insult.

If male Twitter users both post and receive more inflammatory tweets than females, it does not follow they are being harassed more. (Or vice versa.)

7.3 What would it require to track harassment on Twitter?

When is a Twitter reply or mention harassment? The tweet should, at minimum, be adversarial toward the account targeted and use disproportionately hostile rhetoric compared to the target.

For example, @Ash1138's reply to @amy_geek is more confrontational than the original post and it is intended as a criticism.

> I wrote a Star Wars book about 75 of the incredible female characters in the universe with all new art by 18 female and non-binary artists. This is a dream come true. Details: https://www.starwars.com/news/star-wars-women-of-the-galaxy-announced - @amy_geek, May 31, 2018.

> Yay, for you being a sexist! - @Ash1138, May 31, 2018.

Another example is this exchange between @heathdwilliams and @spindry101:

> the last jedi is the best movie ever made [joke image here] - @heathdwilliams, March 1, 2018.

> You fucking idiot. Take your mouth off @rianjohnson cock and give your head a shake. - @spindry101, March 2, 2018.

The original tweet is not very pointed and does not use inflammatory language. The response is a negative comment on the original post and uses more hostile rhetoric.

A tool for tracking harassment on Twitter needs to pick up on both who is being criticized and whether a post has escalated the belligerence of a discussion.