Pognalysis: An Analysis of Gender and Language on Twitch

Julia Cardillo, Pace University

Introduction

A Rising Industry

- Booming video game and eSports industry (Entertainment Software Association, 2021)
- Known hostility towards women (Buyukozturk et al., 2018)
- Twitch, eSports, and video game industry/community go hand in hand

Main Questions:

- Is the atmosphere created by female Twitch streamers distinct from male Twitch streamers?
- Is there more non-gaming related or gendered language/harassment in female Twitch streams than male Twitch streams?

We want to foster a safe and inclusive environment, as well as assess current moderation practices.

Literature Review

(Nakandala et al., 2016)

- Streamer gender is associated with the type of message they receive
- Pulled statistically overrepresented words

(Ruvalcaba et al., 2018)

- More sexual harassment comments received by female than male streamers (pertaining to the most competitive eSports)

(Farrell et al., 2019)

- Use of various lexicons to detect presence of misogynistic rhetoric on Reddit
- Violence and hostility are increasing toward women online

Pre-Estimation Discussion

- If a streamer’s gender is female, there may be a larger presence of gendered harassment/language, non-activity related words, or profanity.
- Shooter games, compared to other activities may contain more profanity or harassment.
- Twitch language can be difficult to analyze due to its nature.

Data Generating Process

Variables

Male

1 - Indicates Male
0 - Non-Male (just female for the scope of this study)

Genre_e

Game or Activity occurring during the stream

Followers

Number of followers (in millions)

dt_e

Time of day (Day, Night, Overnight)

Profanity

1 - Profanity string is present in observation
0 - Profanity string is not present in observation

Harassment

1 - Harassment string is present in observation
0 - Harassment string is not present in observation

Harpro

1 or 0, Profanity | Harassment, either a Harassment or Profanity string is present

Channel_e

Channel that the chat is being pulled from

*Words used for Harassment, Profanity, and Harpro were selected from Hatebase, and Offensive/Profane Lexicon

*Genre_e and dt_e are indicator variables, which shows the marginal effects of each genre and the time of day

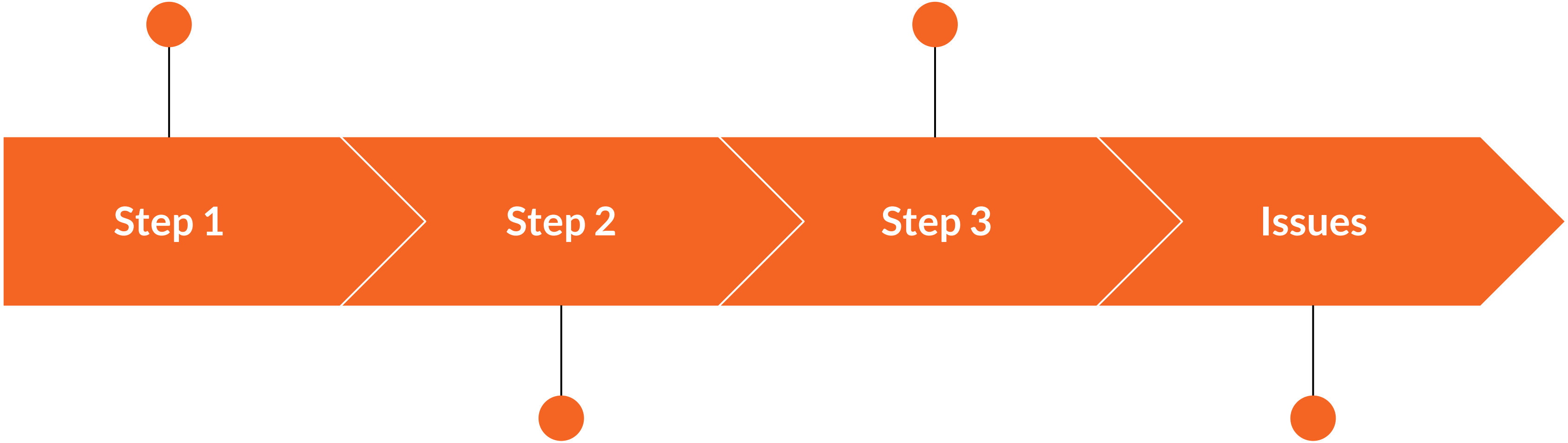
Summary Statistics Table

Summary Statistics

	mean	sd	min	max
harpro	.0122883	.1101709	0	1
profanity	.0120426	.1090772	0	1
harassment	.0006881	.0262238	0	1
genre_e	4.293716	1.664827	1	7
dt_e	1.750768	.7183063	1	3
channel_e	8.246012	3.727981	1	15
followers	3.600683	3.045029	.141	8.4
N	40689			

Web scraped 30 minute intervals of **live** Twitch chats with Python socket scripts (with Author's permission). Includes 7 different games/activities, 1 M 1 F streamer per game/activity.

Compiled into Stata, run probit models to test for change in probability of gendered language/harassment and profanity.



Processed data using Python Pandas and print to Excel, created and ran script that pulls 200 most common words present in dataframe, as well as presence of various words/strings.

DGP was time consuming, and lack of buffer space.
Many technical issues I needed to solve.

Models

$harpro_i = β1male_i + β2genre\_e_i + β3dt\_e_i + β4followers_i + β5channel\_e_i + u_i$

$profanity_i = β1male_i + β2genre\_e_i + β3dt\_e_i + β4followers_i + β5channel\_e_i + u_i$

$harassment_i = β1male_i + β2genre\_e_i + β3dt\_e_i + β4followers_i + β5channel\_e_i + u_i$

Probit Results

	(1) Harpro	(2) Harassment	(3) Profanity
male	0.362***	-0.281*	0.416***
	(0.0644)	(0.169)	(0.0678)
1.Apex Legends	0	0	0
	(.)	(.)	(.)
2.Among Us	0	0	0
	(.)	(.)	(.)
3.Chess	-1.246***	0	-1.203***
	(0.227)	(.)	(0.228)
4.GTA V	-0.475***	-0.283	-0.515***
	(0.114)	(0.305)	(0.121)
5.Just Chatting	-0.214***	3.060	-0.192***
	(0.0625)	(132.3)	(0.0644)
6.League of Legends	-0.263***	3.381	-0.270***
	(0.0880)	(132.3)	(0.0887)
7.VALORANT	-0.445***	0	-0.467***
	(0.119)	(.)	(0.124)

Pseudo R^2:

Harpro: 0.0392
Harassment: 0.0452
Profanity: 0.0429

Cont.

1.Day	0	0	0
	(.)	(.)	(.)
2.Night	-0.133**	0.0102	-0.130**
	(0.0583)	(0.206)	(0.0586)
3.Overnight	-0.0622	3.343	-0.0636
	(0.0997)	(132.3)	(0.103)
channel_e	0.00416	-0.0330	0.0105
	(0.0114)	(0.0213)	(0.0123)
followers	-0.0488***	-0.0618*	-0.0488***
	(0.0168)	(0.0369)	(0.0173)
_cons	-1.999***	-5.615	-2.103***
	(0.122)	(132.3)	(0.131)
N	35372	30250	35372

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Marginal Effects

	(1) harpro	(2) harassment	(3) profanity
male	0.0126***	-0.000855	0.0142***
	(0.00229)	(0.000534)	(0.00237)
1.Apex Legends	0	0	0
	(.)	(.)	(.)
2.Among Us	0	0	0
	(.)	(.)	(.)
3.Chess	-0.0267***	0	-0.0259***
	(0.00370)	(.)	(0.00372)
4.GTA V	-0.0189***	0	-0.0193***
	(0.00458)	(.)	(0.00455)
5.Just Chatting	-0.0108***	0	-0.00967***
	(0.00363)	(.)	(0.00364)
6.League of Legends	-0.0127***	0	-0.0126***
	(0.00425)	(.)	(0.00417)
7.VALORANT	-0.0182***	0	-0.0182***
	(0.00474)	(.)	(0.00472)

Cont.

1.Day	0	0	0
	(.)	(.)	(.)
2.Night	-0.00465**	0.0000203	-0.00446**
	(0.00210)	(0.000410)	(0.00207)
3.Overnight	-0.00234	0.397	-0.00234
	(0.00366)	(36.67)	(0.00367)
channel_e	0.000145	-0.000100	0.000359
	(0.000398)	(0.0000668)	(0.000419)
followers	-0.00170***	-0.000188	-0.00167***
	(0.000588)	(0.000117)	(0.000596)
N	35372	30250	35372

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Common Words

Female - Words	Occurrences
invite	853
pog	410
omegalul	328
kekw	316
lul	279
hair	272
<3	236
pepelaugh	228
lulw	202
lol	188

Male - Words	Occurrences
kekw	1808
lul	935
omegalul	522
letsgo	486
lol	470
lmao	316
ayaya	279
pepela	279
sadge	279
pogu	273

Post Estimation Discussion

- Male streamer indicates that the likelihood of presence of harassment and profanity increases by over 1%
  - Likelihood of presence of just profanity increases by over 1% as well
  - Likelihood of just harassment decreases by .08%

Main issues:

- There are a lot of moderation practices in place, and streamers are able to pre-filter out words.
- Twitch language is complex to analyze.
- Buffer size of my computer was not enough to collect larger samples of data.

Conclusion

Based on my results, there is somewhat a presence of gendered language/harassment and profanity on Twitch.tv.

There is a slight but statistically significant difference in probability of profanity, and environment between male and female Twitch streamers. Harassment however, showed to be statistically insignificant.

Future areas of research could focus more on time of day, region, etc.

In addition, I would like to collect much larger sample sizes.