**Background. The problem being addressed.**

Women are not well-represented in media, especially film. One study found that in over 120 movies across 11 countries, only 31% of named characters were female (Smith, Choueiti, & Pieper, 2014). Beyond simple numerical representation, the depth of female characters is often stereotypically shallow. For example, nearly half of movies in 2016 failed the Bechdel Test (in order to pass, a movie must have two female characters who speak to each other about something besides a man) (Calvario, 2016*).* Thus, the status quo is 1) many films have few female characters, and 2) when a movie does have female characters, they are often not well-developed. Thus, when women are cast in movies as strong characters, they may be met with criticism and and animosity from audiences whose expectations were violated. Strong female characters may also be psychologically threatening (e.g., by challenging traditional gender norms) and instigate people to act out to defend their world views.

One way people may express their discontent toward films featuring many and/ or strong female characters is by "trolling" online (e.g., by writing crude and offensive comments in online spaces like YouTube, Reddit, Twitter, etc.). Indeed, the Internet in general has been a hostile space for women. For example, recently women in the video game industry who advocated for women in gaming were harassed online with vitriolic language via Twitter and 4chan, known as the "Gamergate controversy." Similarly, the new Star Wars franchise has been met with online criticism for making its protagonist a woman.

However, the Star Wars example may be an anomaly because historically, its protagonist was a man and its fanbase was mostly men, so it is unclear whether in general, movies with more female characters are more likely to receive negative online backlash.

**User profile:**

The primary user for this system will be someone interested in social media responses to gender diverse films or someone interested in the social media response to a particular film included in this analysis. This requires little to no knowledge of programming. The user will only need to navigate to our github repository and open the files that will include a visual representation of the results. Those interested in scores for a specific film from our analysis will need basic basic programming knowledge.

Others may find the repository useful if they want to collect their own set of YouTube comments for other film trailers or other videos unrelated to the study or if they want to run sentiment analysis on their own dataset of YouTube comments or other text.

**Data sources. What data you will use and how it is structured.**

Dataset 1: Data on 5000 movies pulled from the International Movie Database (IMDB) by a Kaggle user. This file is currently saved as a .csv but it appears that is in a JSON format (each row is a movie, but for the column "cast" there is a lot of different information in each cell, for example, a list of all character names, the actor's name, gender). Currently, we do not know how data was collected on gender (e.g., whether this is available from IMDB or added by the Kaggle user). List of the cast for each movie appears to be in order of relevance (i.e., the main cast members are listed first).

Dataset 2: Online comments from each movie's trailer on YouTube. Each movie has its own .csv file with each row representing a comment (as strings). Comments are pulled for the first video that comes up when searching the movie's title + "trailer" on YouTube.

**Uses cases:**

**Case 1.**

- *Objective*: Determine whether gender diverse films receive more negative feedback online.
- *Expected interactions between system and user*: Reviewing a visual representation of the results of the study (observing correlation between a film's gender diversity score and sentiment score)

**Case 2.**

- *Objective*: Determine whether a specific film included in this study received majority positive or negative feedback online.
- *Expected interaction between system and user*: The user will input a specific film (included in our database) and receive the film's gender diversity score and sentiment analysis score.