# Does Jubilee's "Middle Ground" experiment work?

## A pilot analysis of YouTube comments

Presentation by :

**KAWTHER ALBADER**


# What is Middle Ground?

- Video series exploring “social and controversial issues through both sides”
- People arguing for and against a series of claims about a topic
- Goal is to promote mutual understanding of both sides

# Method

- **Aim:** to treat study as a pseudo-experiment
- Explored the Middle Ground channel playlist to control for topic
  - *Chosen topic:* Religion
- “Subjects” were “randomly assigned” to each condition
  - 0 = control
  - 1 = Christian video
  - 2 = Muslim video
  - 3 = Mormon video
- Note: One video had an in-person moderator

CONTROL

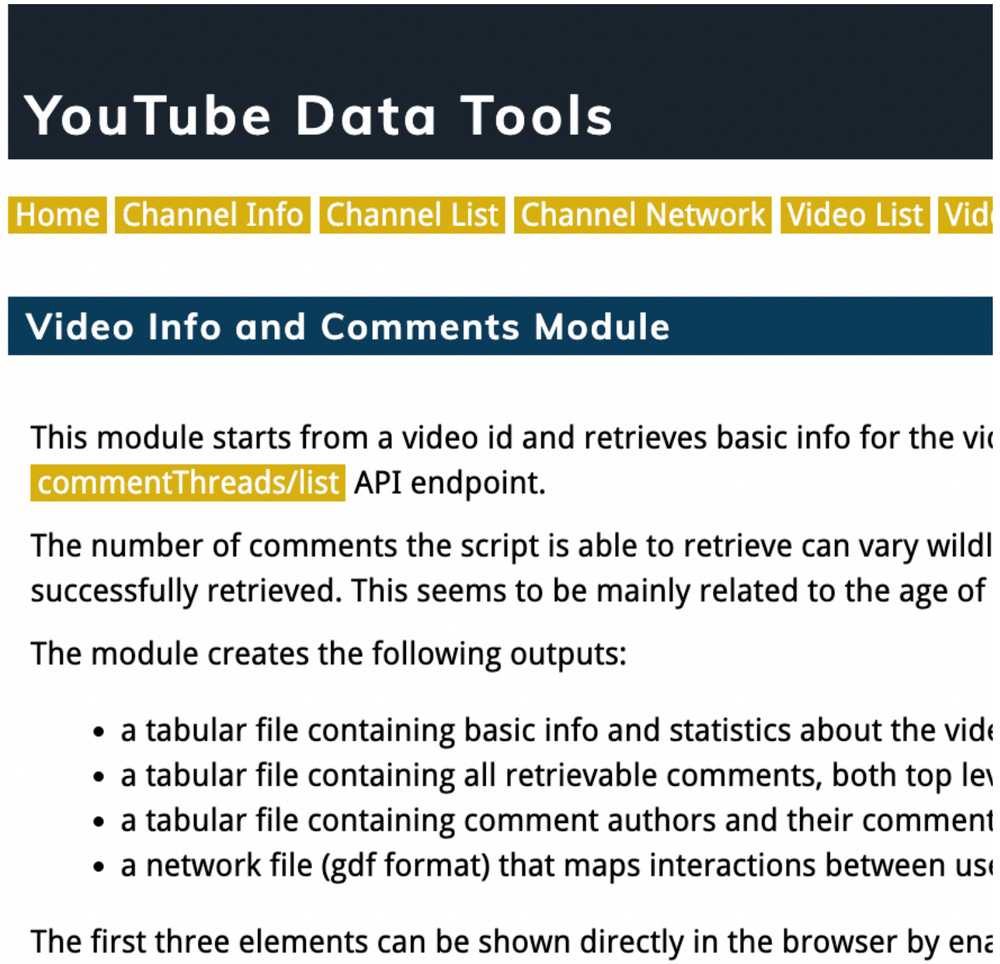

MUSLIM


CHRISTIAN


MORMON

# Data Collection

- **Sourcing:** Manually selected three religion-themed videos from the Middle Ground playlist on the Jubilee channel
  - All topics had similar titles to maintain comparability across conditions
  - Titles were: “Can [RELIGIOUS GROUP] and ex-[RELIGIOUS GROUP] see eye to eye?”
  - All videos were uploaded within the last year
- **Collection:** Extracted comments using **YouTube Data Tools**; sampled 2,000 top-level comments (per video) ranked by relevance



# Data Cleanup

**STEP ONE: R Studio (R)**

- Created an R script to transform all the raw YouTube Data Tools .csv files into dataframes
- Cleaned the dataframes to only include variables of interest, e.g., like count, comments, isreply
- Added a new variable detailing which condition each video is (i.e., 0 = control, 1 = Christian, 2 = Muslim, 3 = Mormon)
- Randomly sampled 750 comments from each condition (since some conditions had more subjects than others)
    - Final sample, *N* = 3,000 (750 comments per condition)
- Combined all the dataframes into one CSV file to upload to Google Co-Lab

# Data Processing

**STEP TWO: Google Co-Lab (Python)**

- Loaded and ran the Pysentimiento package using Python to analyze comments for dependent variables:
  - **Sentiment:** negative, neutral, positive
  - **Emotion:** anger, disgust, fear, etc.
  - **Hate speech:** hatefulness, aggression, targetedness
- Appended the sentiment, emotion, and hate speech comment scores as variables to the existing dataset
  - Resulting in three new variables: sentiment scores, emotion scores, and hate speech scores (each with embedded scores)

*Example of what sentiment, emotion, and hate speech score data looked like pre-parsing:*

| sentiment_scores |
| --- |
| {'NEG': 0.005250483751296997, 'NEU': 0.14093117415905, 'POS': 0.8538183569908142} |
| {'NEG': 0.0016299014678224921, 'NEU': 0.012857824563980103, 'POS': 0.9855123162269592} |
| {'NEG': 0.0026768124662339687, 'NEU': 0.020113950595259666, 'POS': 0.977209210395813} |
| {'NEG': 0.39602282643318176, 'NEU': 0.5810597538948059, 'POS': 0.022917412221431732} |
| {'NEG': 0.0036871088668704033, 'NEU': 0.04416394606232643, 'POS': 0.9521489143371582} |
| {'NEG': 0.9707633852958679, 'NEU': 0.0260330718010664, 'POS': 0.00320355873554945} |
| {'NEG': 0.003488599555566907, 'NEU': 0.01940823905169964, 'POS': 0.9771031141281128} |
| {'NEG': 0.002030836883932352, 'NEU': 0.01894419640302658, 'POS': 0.979024946689605 7} |
| {'NEG': 0.0015142203774303198, 'NEU': 0.012582826428115368, 'POS': 0.9859030246734619} |

# Data Parsing

**STEP THREE: Back to R Studio (R)**

- Created an R script to parse through each of the scores and create separate dependent variables:
  - e.g., sentiment  was parsed into three variables: positive, neutral, negative
  - e.g., hate speech was parsed into three variables: aggression, hatefulness, and targetedness
  - e.g., emotion was parsed into variables like anger, disgust, fear, etc.


**Comments are now ready for analysis!**

# RQs & Analysis Plan (SPSS)

**Does exposure to debate-style videos of different religions result in lower negative sentiment, less negative emotion, and less hate speech for some religions more than others?**

**RQ1: How do each of the video conditions differ in terms of sentiment (as measured by negativity), emotion (as measured by anger, fear, disgust) hate speech (as measured by hatefulness, aggression, and targetedness)?**
--> ANOVA

**RQ2: Does the inclusion of an in-person moderator in a debate-style video result in less hate speech compared to unmoderated videos?**
--> ANOVA

**RQ3: Are positive comments more liked compared to negative comments?**
--> Correlation

# RQ1 Results

**RQ1: How do each of the video conditions differ in terms of sentiment (as measured by negativity), emotion (as measured by anger, fear, disgust) hate speech (as measured by hatefulness, aggression, and targetedness)?**

--> Ran an ANOVA using SPSS

*Sentiment, emotion, and hate speech scores of YouTube comments across condition*

| | Control | Christian | Muslim | Mormon |
|---|---|---|---|---|
| Negativity<br>*M* (*SD*) | $2.44(3.65)_a$ | $2.78(3.72)_a$ | $4.30(3.94)_b$ | $3.85(4.02)_b$ |
| Anger<br>*M* (*SD*) | $1.37(8.50)_a$ | $2.08(1.11)_a$ | $2.64(1.26)_a$ | $1.44(8.18)_a$ |
| Disgust<br>*M* (*SD*) | $1.12(2.65)_a$ | $2.17(3.55)_b$ | $3.57(4.17)_c$ | $3.25(4.17)_c$ |
| Fear<br>*M* (*SD*) | $4.96(1.38)_a$ | $6.91(5.51)_a$ | $3.72(6.10)_a$ | $6.15(4.81)_a$ |
| Hateful<br>*M* (*SD*) | $0.08(0.15)_a$ | $0.04(0.08)_b$ | $0.08(0.16)_{ac}$ | $0.04(0.10)_b$ |
| Targeted<br>*M* (*SD*) | $0.06(0.13)_a$ | $0.02(0.06)_b$ | $0.02(0.06)_b$ | $0.02(0.04)_b$ |
| Aggression<br>*M* (*SD*) | $0.01(0.01)_a$ | $0.01(0.01)_a$ | $0.02(0.03)_b$ | $0.01(0.01)_a$ |

*Note.* Means that do not share subscripts are significantly different ($p < .05$, Tukey's comparison).

**Sentiment - Negativity:**

- Muslim and Mormon conditions differed from control, meaning comments were more negative for the Muslim and Mormon videos than both the Christian and control groups.

**Emotion - Disgust:**

- All groups differed from control, but the Muslim and Mormon groups didn't differ from each other (note: they had the highest means).

**Hate Speech - Hatefulness:**

- Christian and Mormon groups had lower hate than the control. Muslim group had higher hatefulness than the Christian and Mormon conditions.

**Hate Speech - Targeted:**

- All groups differed significantly from control but not from each other.

**Hate Speech - Aggression:**

- Muslim video comments were rated as more aggressive compared to all other conditions.

*Sentiment, emotion, and hate speech scores of YouTube comments across condition*

| | Control | Christian | Muslim | Mormon |
|---|---|---|---|---|
| Negativity *M* (*SD*) | $2.44(3.65)_a$ | $2.78(3.72)_a$ | $4.30(3.94)_b$ | $3.85(4.02)_b$ |
| Anger *M* (*SD*) | $1.37(8.50)_a$ | $2.08(1.11)_a$ | $2.64(1.26)_a$ | $1.44(8.18)_a$ |
| Disgust *M* (*SD*) | $1.12(2.65)_a$ | $2.17(3.55)_b$ | $3.57(4.17)_c$ | $3.25(4.17)_c$ |
| Fear *M* (*SD*) | $4.96(1.38)_a$ | $6.91(5.51)_a$ | $3.72(6.10)_a$ | $6.15(4.81)_a$ |
| Hateful *M* (*SD*) | $0.08(0.15)_a$ | $0.04(0.08)_b$ | $0.08(0.16)_{ac}$ | $0.04(0.10)_b$ |
| Targeted *M* (*SD*) | $0.06(0.13)_a$ | $0.02(0.06)_b$ | $0.02(0.06)_b$ | $0.02(0.04)_b$ |
| Aggression *M* (*SD*) | $0.01(0.01)_a$ | $0.01(0.01)_a$ | $0.02(0.03)_b$ | $0.01(0.01)_a$ |

*Note.* Means that do not share subscripts are significantly different ($p < .05$, Tukey's comparison).

*No significant differences between groups for anger or fear.

# RQ2 Results

**RQ2: Does the addition of an in-person moderator in a debate-style video result in less hate speech compared to unmoderated videos?**

--> Ran an ANOVA using SPSS with video condition as the independent variable (1 = Christian, 2 = Muslim, 3 = Mormon-moderated) and hate speech (a scale created with the mean of the three hate speech variables) as the dependent variable.

**Results:**

- *Assuming the only difference is the addition of the moderator* (which it's not, because the "religions" in each video differed), the moderator video condition (i.e., the Mormon video) had significantly less hate speech compared to the Muslim condition.
- However, both the Christian and Mormon video conditions did not differ in terms of hate speech, indicating the presence of a moderator likely did not mitigate hate speech.

# RQ3 Results

**RQ3: Are positive comments more liked compared to negative comments?**

--> Ran a correlation via SPSS

**Results:**
Like count was significantly and negatively related to neutrality, r = -.05, *p* = .007, CI [-.09, -.01].
The more neutral a comment, the lower the likelihood that it would be liked.

However, like count was not significantly associated with either positive nor negative sentiment.

# Limitations (Validity)

- Pseudo-experiment:
  - Sentiment may be a function of the personalities/characters chosen in the videos, not necessarily the topic itself
  - No control over the stimuli
  - “Subjects” who watch Middle Ground series may be fundamentally different from those who watch more “fun” videos like the control
  - Assumptions (e.g., one comment per subject, users only exposed to one video each)
- Limitations with Pysentimiento:
  - Sarcasm and humor
- Sampled most relevant comments but unclear as to how “relevance” is measured according to YouTube and YouTube Data Tools
  - Download all comments next time?

# Challenges & Future Directions

**Challenges:**

- Finding the right setting to apply pseudo-experiment to: Difficult to conceptualize a YouTube experiment because conditions need to be kept somewhat comparable
- Require more control over stimuli

**Future Directions:**

- Apply to different contexts
  - TikTok comments of politician videos (TikTok works better than YouTube b/c politicians have higher following/more comments on TikTok)
- Employ natural language processing approaches to explore comment content further and extract own themes (versus being limited to Pysentimiento)
- Inclusion of human coders as a validity check for Pysentimiento

# Thank You!