Do YouTubers Promote Bullshitting using ChatGPT? Exploring the Use of Large-Language Models in YouTube Videos and Their Risk Landscapes

Linh Bui[1], Stephen Gerson[4], Megan Lincicum[2], Leyat Samson[3], Zhenan Zhang[1], Julia Hsu[1], Myeong Lee[1]

[1]George Mason University [2]Lake Braddock Secondary School [3]South County High School [4]College of William & Mary

Aspiring Scientists Summer Internship Program

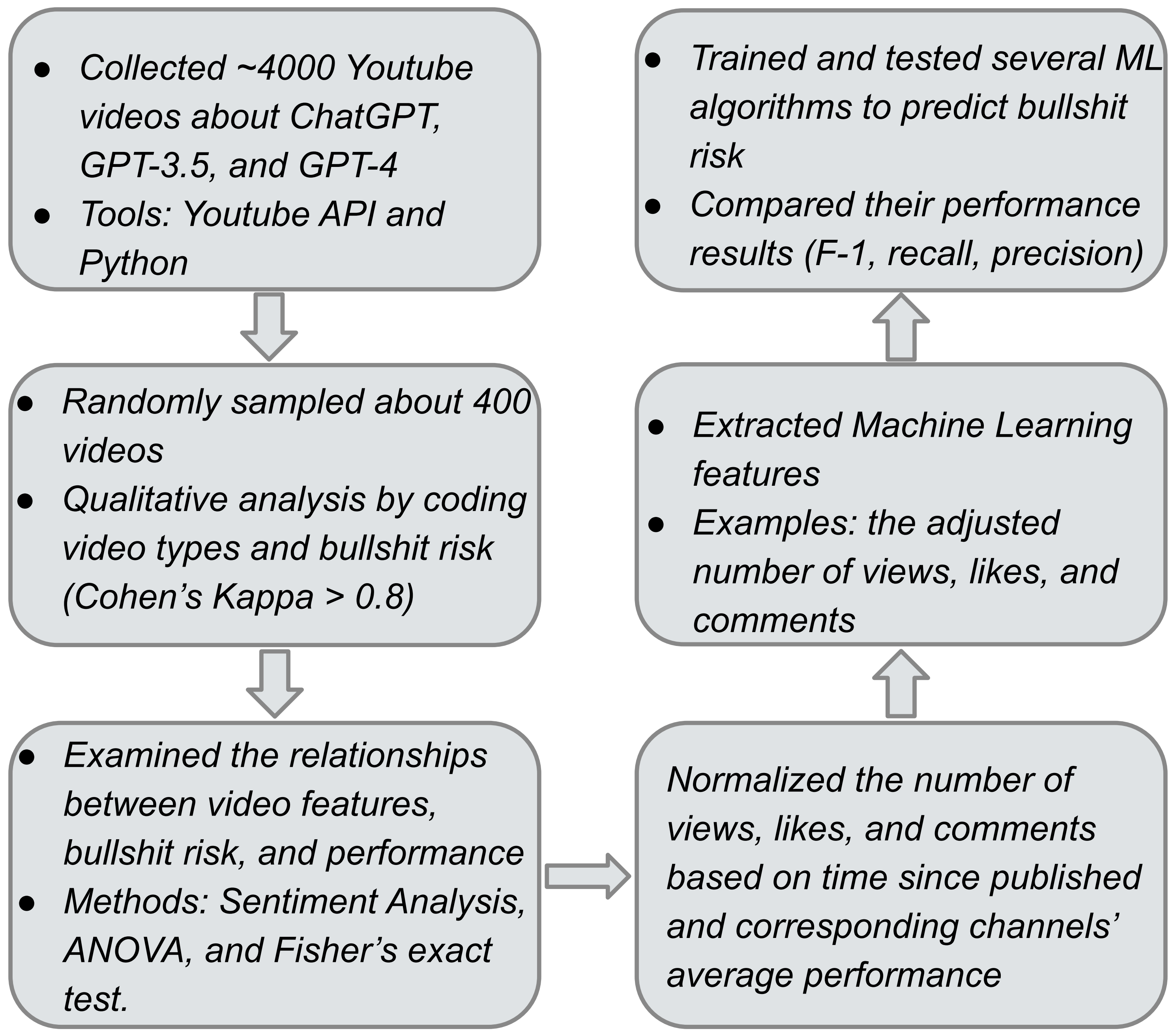
Abstract

Large language models (LLMs) like ChatGPT are, as per Frankfurt's definition, "bullshit" generators since they lack the ability to recognize truths. As ChatGPT gains popularity, there is a growing concern that it could facilitate the spread of "bullshits" (utterances made without recognition of truths) on the Internet, exacerbating the media environment. To understand this potential risk, we collected about 4,000 YouTube videos related to ChatGPT that have been published since December 2022, conducted qualitative analysis on approximately 400 randomly-sampled videos, and examined the relationships between video features, bullshit risk, and performance via sentiment analysis, ANOVA, and Fisher's exact test. We then leveraged channel statistics to train a machine learning model that can predict the entire videos' risk of bullshit. The study emphasizes the high potential for AI-generated content to spread bullshit and the necessity for social media designers to develop more effective strategies that address the potential risk.

Research Questions

- *To what extent do ChatGPT-related Youtube videos spread "bullshit" content on the Internet?*
- *How do different video characteristics relate to the potential "bullshit" risk?*

Approach



- *Collected ~4000 Youtube videos about ChatGPT, GPT-3.5, and GPT-4*
- *Tools: Youtube API and Python*

- *Randomly sampled about 400 videos*
- *Qualitative analysis by coding video types and bullshit risk (Cohen's Kappa > 0.8)*

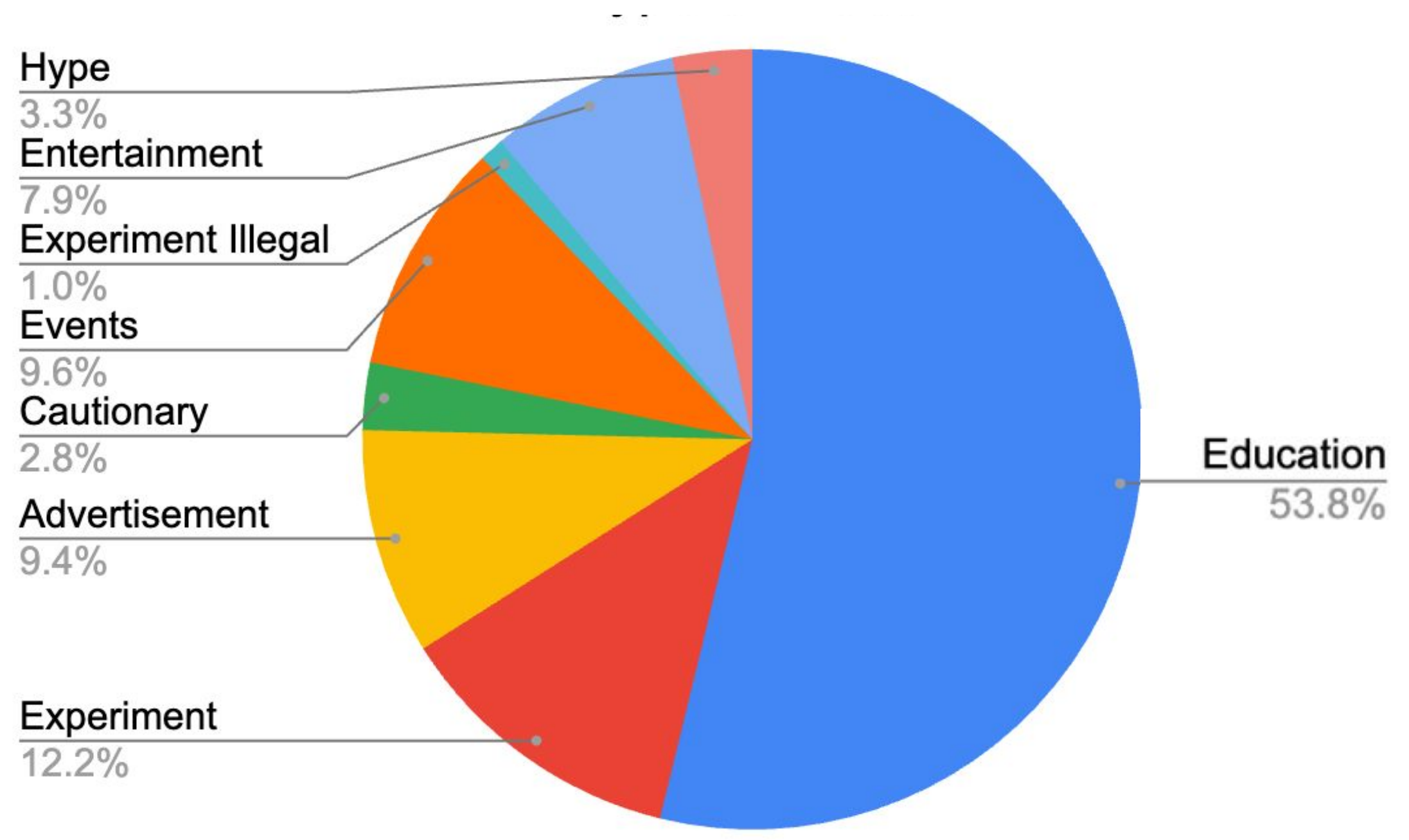
- *Examined the relationships between video features, bullshit risk, and performance*
- *Methods: Sentiment Analysis, ANOVA, and Fisher's exact test.*

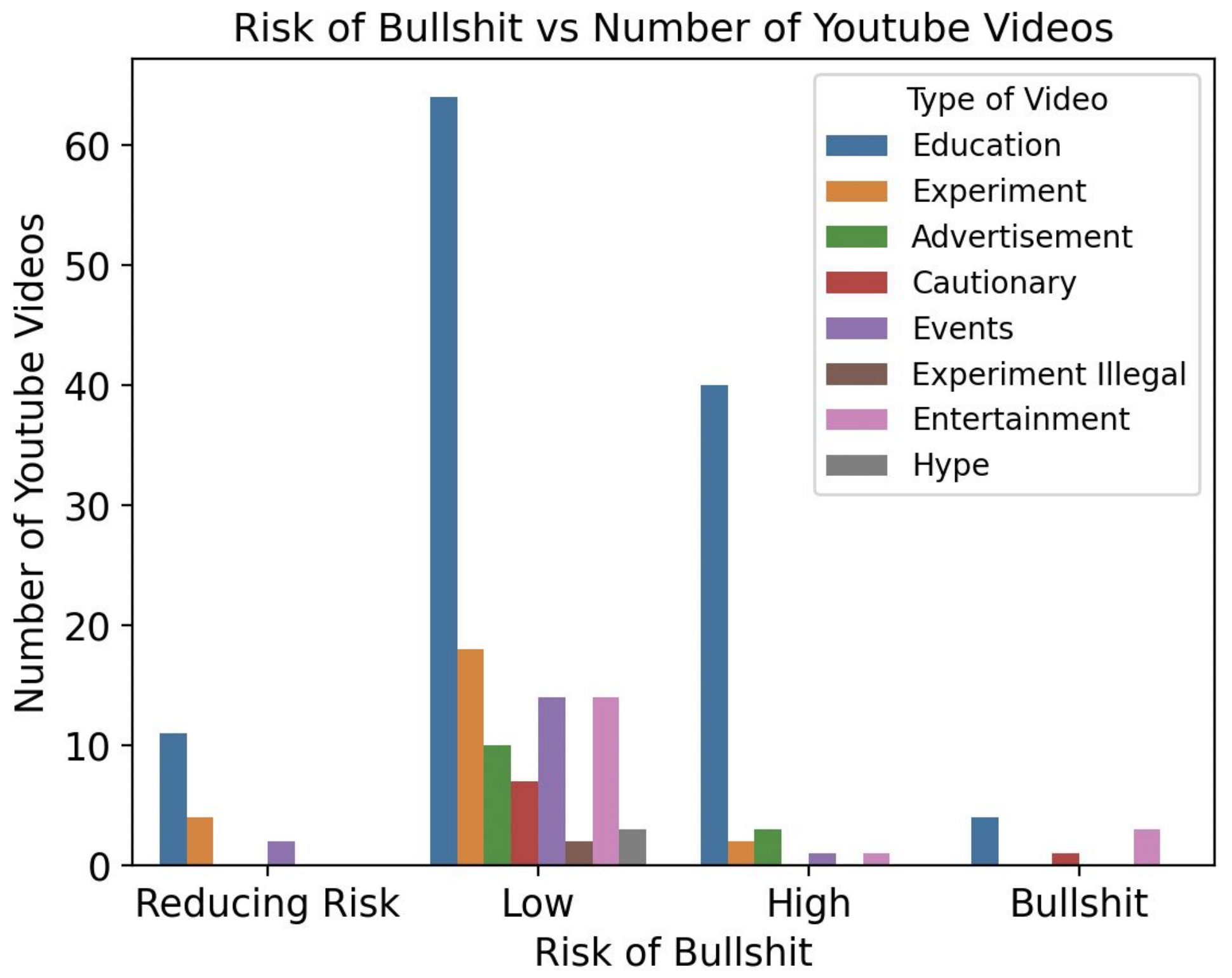
Normalized the number of views, likes, and comments based on time since published and corresponding channels' average performance

- *Extracted Machine Learning features*
- *Examples: the adjusted number of views, likes, and comments*

- *Trained and tested several ML algorithms to predict bullshit risk*
- *Compared their performance results (F-1, recall, precision)*

Results





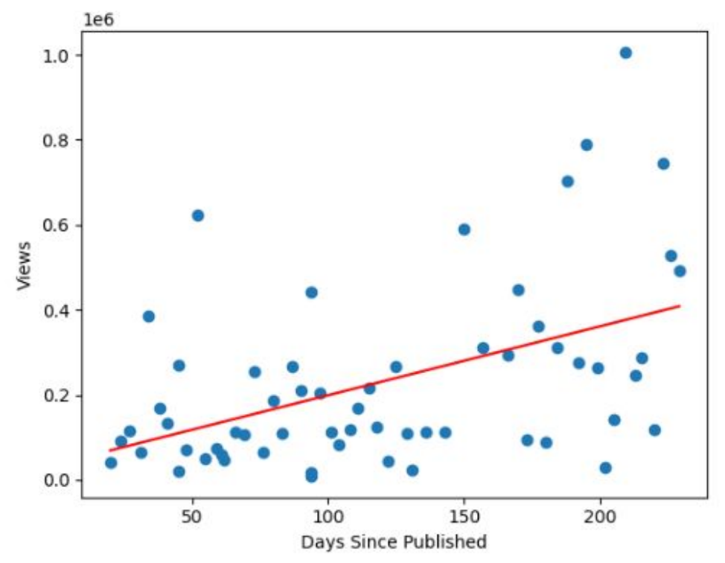
Machine Learning

ML models such as Random Forest, XGBoost, and Logistic Lasso Regression were trained on video and channel level features to predict bullshit risk. Most achieved accuracy scores around 70%, but F-1, precision, and recall scores were much lower. Further feature engineering is needed to improve the performance.

Data Pre-Processing

Number of likes, views, and comments on videos were adjusted using Lasso Regression to account for the difference in time since published.

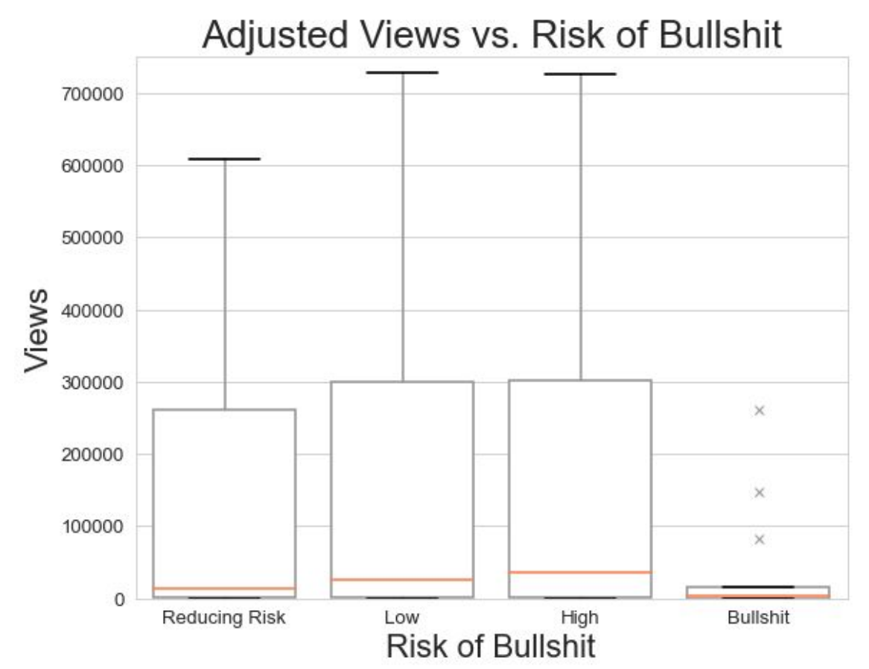
$$P_{200} * \left(1 + \left(\frac{A_{Observed} - P_{current}}{P_{current}}\right)\right)$$

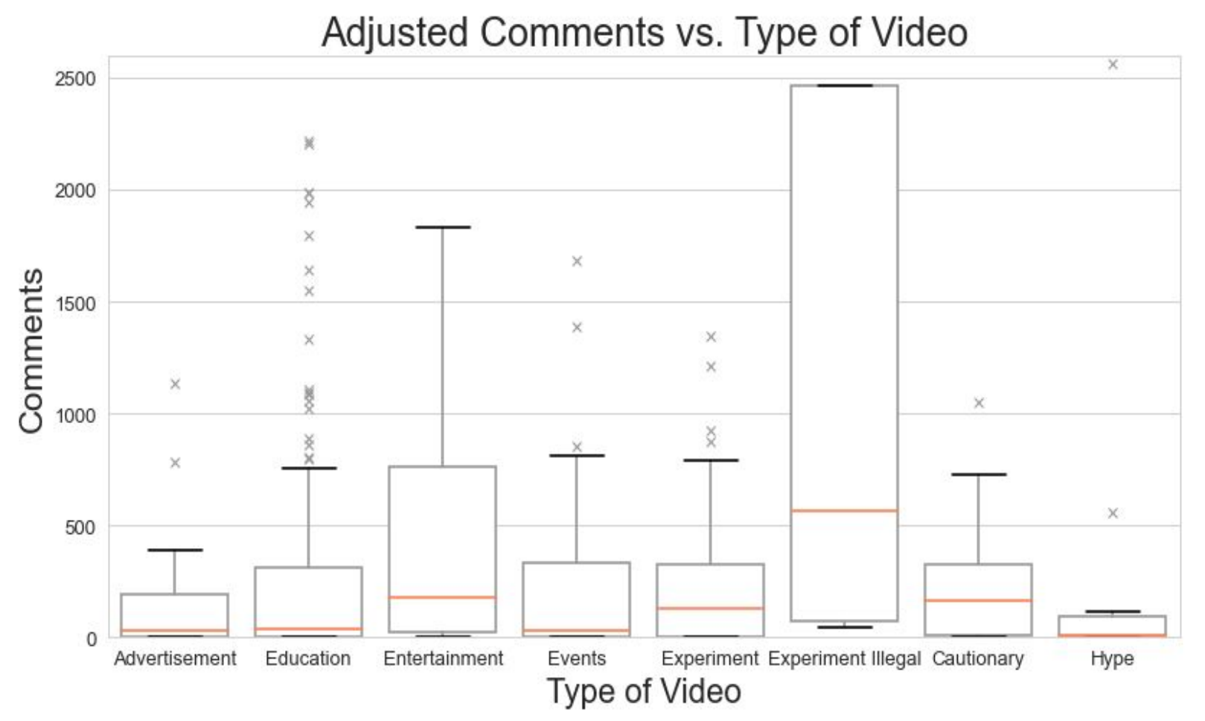


Statistical Analysis

H1: Number of views is associated with types of bullshit risk.

While our analysis does not reject the null hypothesis, bullshit videos appear to have less views on average than other types.

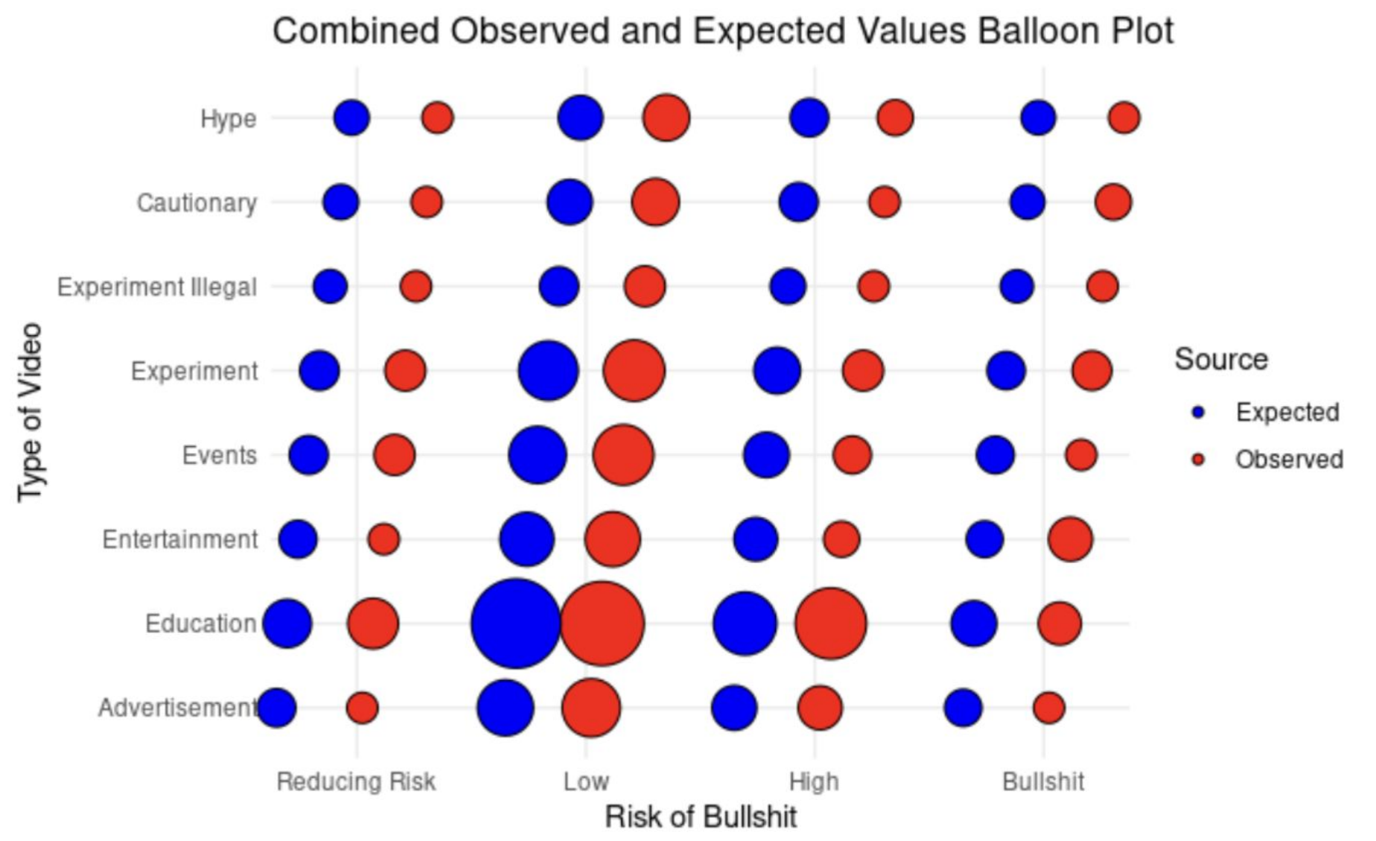




H2: Number of comments is associated with types of videos.

While the ANOVA result rejects the null hypothesis, significant differences are observed only between a small number of types.

Fisher's Exact Test: Post-Hoc Analysis



Types of Videos are related to the Risk of Bullshit (p=0.01) Bubble size differences indicates differences. "Educational" and "Entertainment" type videos tend to present more bullshit risk. Lower risks of bullshit amongst videoes labeled as "Events", "Experiment", or "Cautionary".

Discussion

- *Social media managers and researchers become possible to focus on people's motivation and behavior more rather than whether pieces of information are correct or not.*
- *Our project fosters a safer online experience for users, ensuring a more reliable and accurate information environment.*

Conclusions & Future Work

This study initiated the exploration of the potential risk of "bullshit" generation by large language models (LLMs) like ChatGPT on Youtube. To further advance our understanding, we intend to:

- *Expand the dataset by incorporating additional keywords to ensure a comprehensive coverage of AI-generated content.*
- *Develop a robust machine learning model to automatically tag and categorize data for in-depth statistical analysis.*
- *Utilize machine learning to predict bullshit risk across the entire dataset of videos, offering automated insights into content credibility.*

Citations

Frankfurt, H. G. (2005). On bullshit. Princeton, NJ, Princeton University Press.

Acknowledgements

This research was made possible through the support of George Mason University's College of Science, which supports the ASSIP Program.