Unveiling negativity, offensive language and hate speech

An analysis of the effects of content creators' socio structure and platform characteristics using a representative sample of YouTube channels from German-speaking countries

Claudia Buder, Nina-Sophie Fritsch, Chiara Osorio Krauter, **Aaron Philipp**, Roland Verwiebe, **Sarah Weißmann**

30th International Conference of Europeanists
Radical Europe: Violence, Emancipation & Reaction
July 3 - 5, 2024 | Lyon

State of the art

- **Relevance**: Negativity and hate speech in comments influence the **well-being** and mental **health of CCs** (Krämer et al., 2021; Lutz & Schneider, 2021), whose **working conditions** are already quite **stress- and harmful** (we tend to accept this but wouldn't for many other occupations).
- There is a <u>smaller number of studies</u> analysing **risk factors** for the occurrence of **negativity** and **hate speech** in **comments** (especially missing in a quantitative, comprehensive view).

State of the art: At-risk factors

Mainly **qualitative research** shows, affected are:

- **Female** CC (especially in science, politics or gaming) (Kim, 2023)
- **BIPoC** CC (Sue et al., 2007, Tynes et al. 2018)
- **Disabled** CC (Heung et al. 2024)
- **Young** CC (Obermaier & Schmuck 2022)
- A visible **religious affiliation** CC (Keipi et a. 2017)
- **Large audience** (Thelwall et al., 2012) , fluid communities or **viral** videos (Mathew et al. 2019; Uyheng, J., & Carley, K. M.2021)
- CC producing **polarizing content** or **draw hate bubbles** (Goel et al., 2023; Xin, 2023)
- CC producing content in **topics** <u>gaming</u> or <u>science</u> (Döring & Mohseni, 2019)

Research question

To what extent are **negativity** and **hate speech** in YT comments influenced by the content creators' **socio structure** (e.g. age, gender, religion, race) and specific **platform characteristics of YouTube** (e.g. channel topic, number of subscribers, channel age, community strength)?

Hypotheses

- **H1**: The incidence of negativity and hate speech in comments is determined by content creatos' **socio structure** that "provoke" people to criticize or hate certain content.
  - The current state of the art indicates that women, BIPoCs, religious CCs and young CCs are at risk (effects of education are open for debate).
- **H2**: The incidence of negativity and hate speech in comments is determined by **platform characteristics** that create a more toxic environment for CCs.
  - The current state of the art indicates that larger/more popular channels and certain topics are at risk.

Data, design of the study

N	
115.976	• full coverage of channels in Germany, Austria & Switzerland (Dec. 2022; min. 10 videos)
5000	• Random Sample and classification of the channels
4352	• Exclusion of media and promotional channels
3746	• Filter channels which produce content for the German-Speaking Community • 36 million YT-Comments

Data, design of the study

- Jul-Dec 2023: Web scraping (with the YouTube Data API v3)
  - all comments under each video of our classified channels
  - platform variables: number of subscribers, views, videos, channel foundation

- We used a hand-coded classification survey to categorize
  - Content Creators' socio-structure: sex, age, race, education, migration background, religious visibility
  - platform related variables of the channel: channel topic, visibility of the CC etc.

Unique combination of socio-structural and platform variables → Research gap

Sample composition

Variables	N
Sex	
Female	565
Male	2597
Mixed	76
Missing	508
Race	
BIPoC	351
White	1834
Mixed	13
Missing	1548
Migration Background	
Yes	398
No	1235
Mixed	7
Missing	2106
Education	
Low	198
High	345
Mixed	0
Missing	3203

Variables	N
Age	
$\leq$ 20 years	557
21-30 years	817
31-40 years	584
40+ years	661
Mixed	33
Missing	1094
Religion	
Yes	26
No	529
Mixed	0
Missing	3191
Subscriber	
Mean	24,342
Median	498
Channel Age [in years]	
Mean	9.13
Median	8.67

Variables	N
Community Strength	
Mean	0.519
Median	0.545
Channel Topic	
Arts & Culture	486
Beauty & Lifestyle	122
Business & Finances	39
Conspiracy Theory & Spirituality	94
DIY	302
Education & Knowledge	97
Entertainment	822
Food & Culinary	71
Gaming	1300
Health	78
Politics & Society	24
Sport	120
Travel	176
Other	15
Observations	3,746

Methods: Sentiment analysis and hate speech detection

- Sentiment Analysis: Determination of the emotional message in either positive, neutral or negative → **Mean sentiment on channel level**

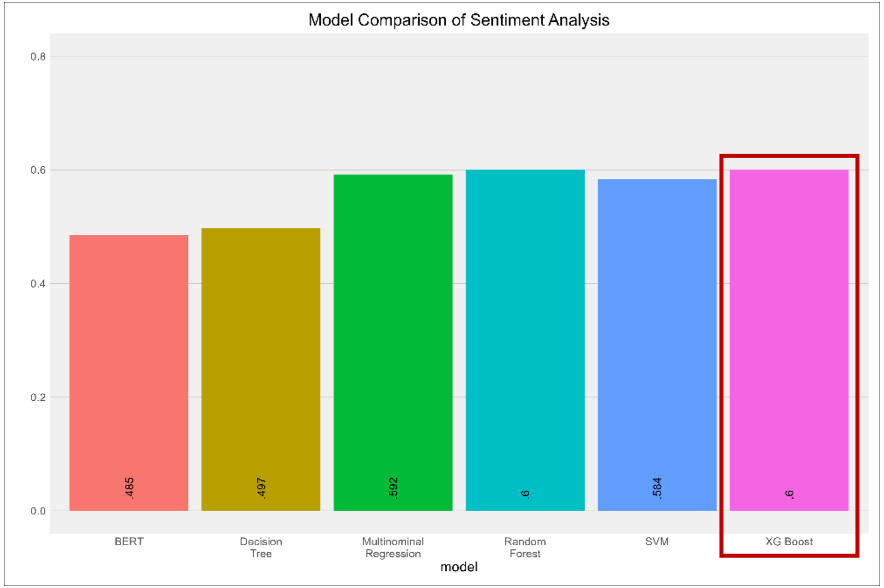
- Hate speech Detection: Classify acts of violence or expressions of hatred directed towards a specific/protected group of people or an individual belonging to such a group. (e.g., race, religion, sexual orientation, gender, disability, age)
  → **proportion of hate comments on channel level**

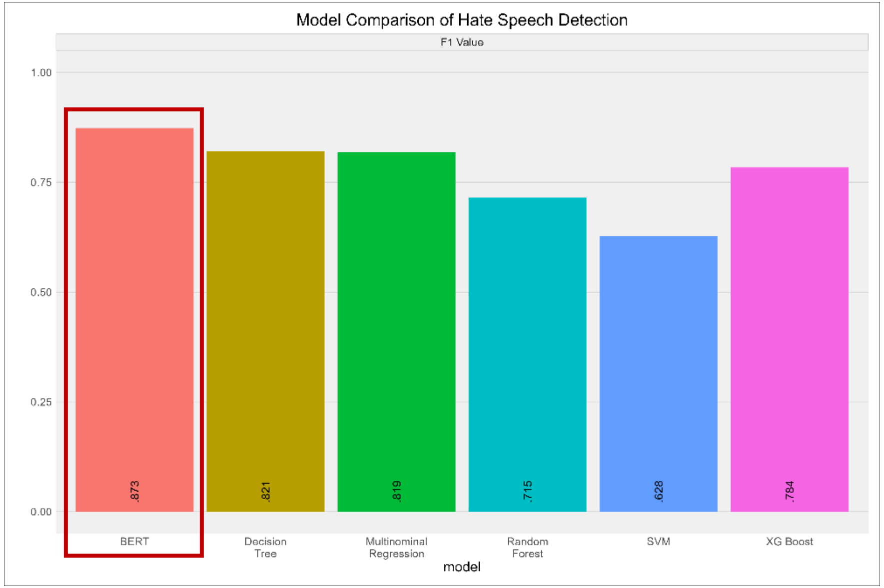
- Offensive Language: Comments that don't constitute hate but are characterized by hostility, insults, or toxicity → **proportion of offensive comments** (*including* hate speech) **on channel level**

Proportions of hate: no hate 97,3% | offensive 1,9 % | hate 0,8%

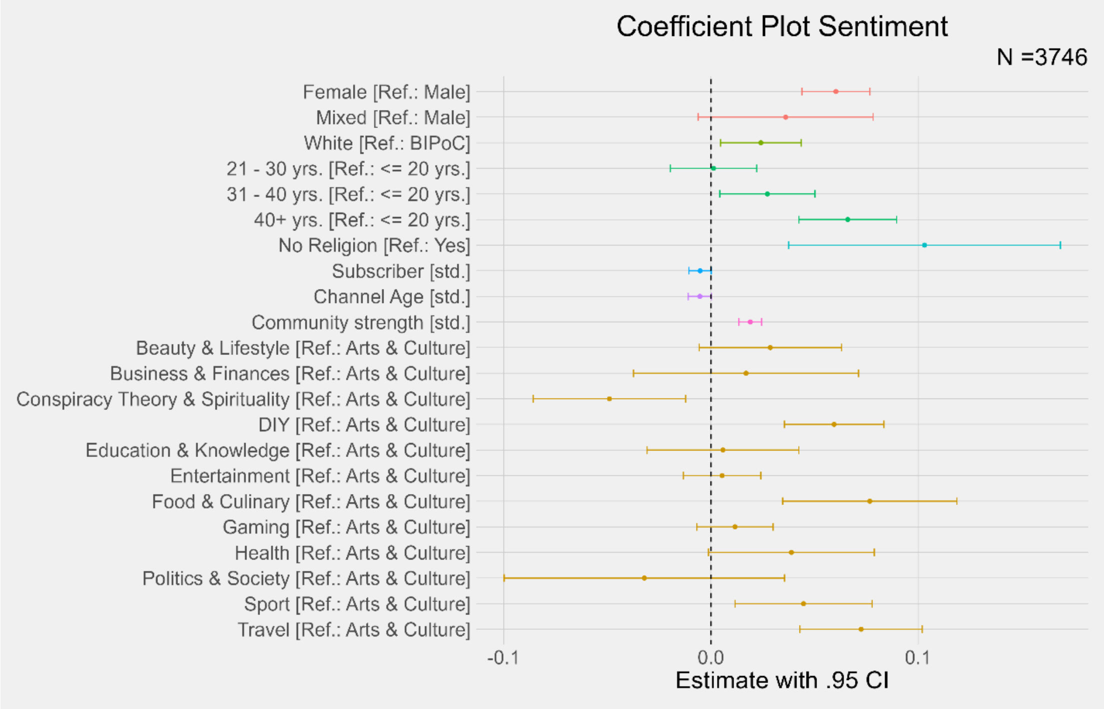
Performance of our models

- We predicted the sentiment and occurrence of hate speech in our 36 mio. YT comments using NLP techniques (ML & DL)
- Training data: 7500 german & english hand-coded YouTube comments





Channel sentiment (multiple predictors)



Note: Unknown category omitted from plot

H1 Socio-Structural Variables:

- Women
- White CCs
- Higher Age
- No Religious Visibility

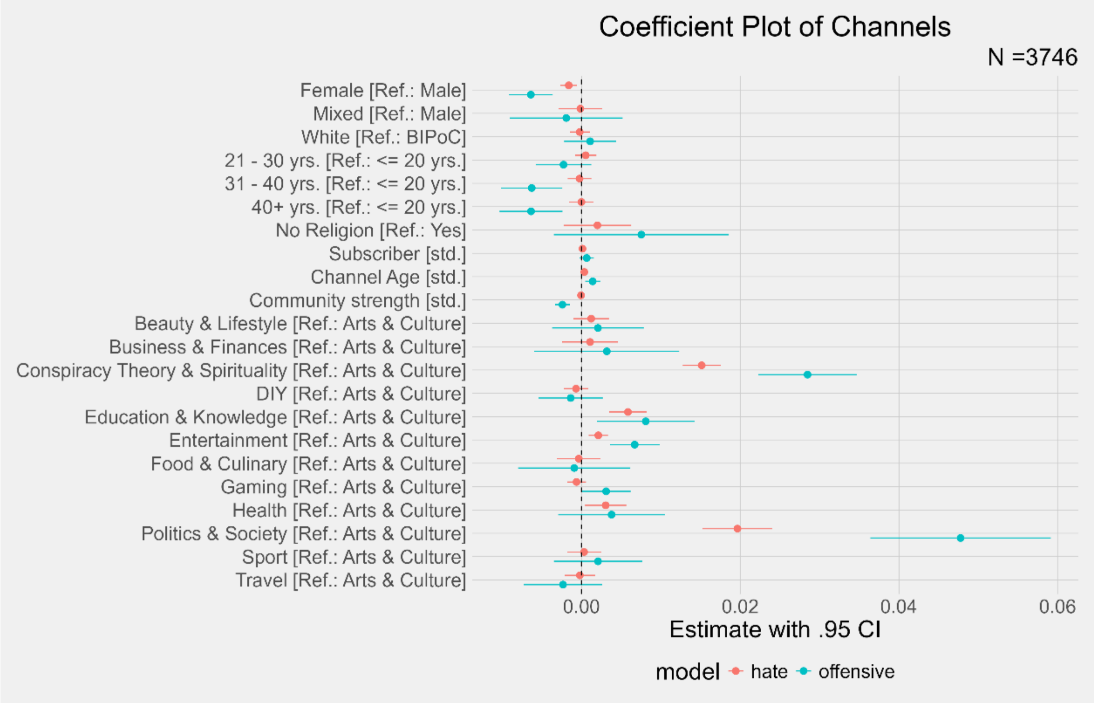
H2 Platform Variables:

- Community Strength
- DIY
- Food
- Travel
- Conspiracy

Positive Sentiment

Negative Sentiment

Hate speech and offensive comments (multiple predictors)



Note: Unknown category omitted from plot

H1 Socio-Structural Variables:

- Women
- Higher Age

H2 Platform Variables:

- Community Strength
- Channel Age
- Education
- Entertainment
- Health
- Politics
- Conspiracy
- Gaming

Less Hate/ Offensivness

More Hate/ Offensivness

Conclusion

- Sentiment, offensive comments and hate speech are structured by both socio-structural and platform variables in distinct ways
  - **Gender, race, age, and channel topic** are **at-risk factors** of receiving negative communication
  - However: **Counterintuitive gender effect**! (systematic differences in audiences?)
- Hate seems to be particularly structured around **channel topics** that generate controversial content (e.g., politics & conspiracy)
  - Content moderation as a considerable factor in hate speech detection

Our Contribution

1. **CCs** form a **new occupational group** on algorithm-based platforms. **Negativity** and **hate speech** is part of their **work environment**.
2. This has not been studied with a **comprehensive approach** before (focus on Germany): Combining digital trace data (which reflect the logic of **platform architecture**) with socio-structural data (which allows for **inequality analyses**).

Our Contribution

3. **Specific results** for negativity, hate speech and offensive language – indicating that they are **different phenomena**

   (but partially **similar socio-platform stratification logic**)

4. A '**low' prevalence** of **hate** on **YouTube** (0.8% of all comments)!? – the platform seems **less polarized** and **toxic** than X for example, reflecting different business philosophies, diverging cultures and specific audiences in the various arenas of the digital world.

Potential limitations

- Moderation of comments (deleting, reporting etc.) cannot be captured but might differ between different communities and CCs → topic of a survey we'll send out to CCs in 08/2024

- The algorithmic structure of the platform is unknown (difficult to measure effects of spam filters, recommender algorithms etc. on the commenting behavior).

Thank you for your attention!

Website: https://www.uni-potsdam.de/de/sozialstrukturanalyse/index/forschung/tubework

E-Mail: roland.verwiebe@uni-potsdam.de, claudia.buder@uni-potsdam.de, aaron.philipp@uni-potsdam.de, sarah.weissmann@uni-potsdam.de

GitHub web scraping: https://github.com/AaronPhilipp/youtube_data_api

References

- Amarasekara, I., & Grant, W. J. (2019). Exploring the YouTube science communication gender gap: A sentiment analysis. Public Understanding of Science, 28(1), 68–84. https://doi.org/10.1177/0963662518786654
- DeBerta: (He, P., Gao, J., & Chen, W. (2021). DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. arXiv preprint, arXiv:2111.09543.
- DeTox Dataset: Demus, C., Pitz, P., Schütz, M., Probol, N., Siegel, M., and Labudde, L. 2022. DeTox: A Comprehensive Dataset for German Offensive Language and Conversation Analysis. In Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), pages 143–153, Seattle, Washington (Hybrid). Association for Computational Linguistics.
- Keum, B. T., & Miller, M. J. (2017). Racism in digital era: Development and initial validation of the Perceived Online Racism Scale (PORS v1.0). *Journal of Counseling Psychology, 64*(3), 310–324.
- Md Saroar Jahan, Mourad Oussalah (2023), A systematic review of hate speech automatic detection using natural language processing, Neurocomputing, Volume 546,
- Nicola Döring & M. Rohangis Mohseni (2019) Fail videos and related video comments on YouTube: a case of sexualization of women and gendered hate speech?, Communication Research Reports, 36:3, 254-264, DOI: 10.1080/08824096.2019.1634533.

References

- Thomas, Kurt; Kelley, Patrick Gage; Consolvo, Sunny; Samermit, Patrawat; Bursztein, Elie (2022): “It’s common and a part of being a content creator”: Understanding How Creators Experience and Cope with Hate and Harassment Online. In: Simone Barbosa (Hg.): CHI Conference on Human Factors in Computing ACM Special Interest Group on Computer-Human Interaction; ACM SIGs. New York,NY, United States, S. 1-15.
- Sue, D. W., Capodilupo, C. M., Torino, G. C., Bucceri, J. M., Holder, A., Nadal, K. L., & Esquilin, M. (2007). Racial microaggressions in everyday life: implications for clinical practice. American Psychologist, 62(4), 271–286.
- Tynes, B. M., Lozada, F. T., Smith, N. A., & Stewart, A. (2018). From racial microaggressions to hate crimes: A model of online racism based on the lived experiences of adolescents of color. In C. Capodilupo, K. Nadal, D. Rivera, D. W. Sue, & G. Torino (Eds.), Microaggression theory: Influence and implications. Hoboken, NJ: Wiley.
- Veletsianos, G., Kimmons, R., Larsen, R., Dousay, T. A., & Lowenthal, P. R. (2018). Public comment sentiment on educational videos: Understanding the effects of presenter gender, video format, threading, and moderation on YouTube TED talk comments. PLOS ONE, 13(6), e0197331. https://doi.org/10.1371/journal.pone.0197331
- YouTube 2024: https://transparencyreport.google.com/youtube-policy/removals?hl=en&comments_removal_reason=period:2023Q3&lu=videos_by_country&videos_by_country=period:2023Q3