Write-up: Facebook Field study with Local Politicians

# Introduction

Online political hostility, such as hate speech on the internet, is a widespread phenomenon that poses a threat to democracy by undermining political engagement. This unfortunate tendency possibly entails users to withdraw from the political debate. While political science research has extensively studied aggressors and the development of politically hostile behaviour, there has been relatively little focus on bystanders and their potential pro-social reactions, and whether it is possible to encourage pro-social online behaviour. Pro-social bystander behaviour refers to behaviour where users respond to hateful content by speaking up and steering the communication in a more positive and constructive direction, in contrast to simply ignoring the content or responding with more hate.

# Study objective

This project is part of the Standing by: Pro-social Bystander Reactions to Online Political Hostility (STANDBY) project, aiming to explain pro-social bystander reactions and their consequences. The project is funded by the European Research Council (ERC).

The main objective of this study was to investigate if Facebook users respond more pro-socially to hateful content on the Facebook posts of Danish local politicians from Århus and Copenhagen, when an intervention comment encouraging users to: ‘Sig Fra, Anmeld, støt op’ has been implemented. Furthermore, to investigate any more general tendencies in the commenting on politicians Facebook posts for the two municipalities.

Several research questions were formulated:

1. Is there a difference in the Facebook comments when the intervention is applied compared to when it is not?
2. How do people respond to negative comments? do negative comments also entail negative subcomments (Comment to a comment) and vice versa?
3. Is there a difference across the two municipalities? Is the commenting different in Århus compared to Copenhagen?
4. Are there general tendencies in the commenting on the politicians’ posts? Do people generally comment negatively, aggressively, encouraging etc.?

# Study Design

Local politicians from Århus and Copenhagen was invited to participate in the study. 11 politicians, of which 5 were from Copenhagen, took part in the intervention part of the study. Politicians were instructed to comment a planned, pre-written text to their post, alongside with a short, animated video emphasising the message. In short terms, the text encourages users to speak up against hateful content by creating awareness of the hatefulness of certain comments, report the comment or user, and showing your support to the receiver of the hateful comment.

Et billede, der indeholder tekst, skærmbillede, pattedyr

Automatisk genereret beskrivelse

Figure 1: Intervention text and video

The full text is available in the appendix. The politicians were instructed not to change anything else about the post or the content of the post itself. The treatment weeks were randomly assigned to the politicians, hence the intervention was not implemented on all posts of a participating politician. The intervention schedule ran for 8 weeks in total.

Furthermore, Facebook thread data was collected for 66 additional local politicians from Århus and Copenhagen to expand the data pool, making comparison between treatment and control posts possible.

# Data

Data from 1601 Facebook threads were collected, of which 51 were intervention posts. After data cleaning, a total of 13.102 comments remains, of which 489 are comments from intervention posts. Posts are collected from both the private and public accounts of politicians. Date of posting runs from 01/05/23 to 02/07/23 totalling 9 weeks. The data was collected and formatted to Json files from 02/07/23 to 25/08/23.

To extract emotions and attitudes as a way of quantifying the comments, 5 types of sentiment analysis were conducted for every comment:

1. VADER: giving a negative, positive, neutral and compound score
2. ROBERTA: giving a negative, positive, neutral and compound score
3. Danish BERT: giving 1 of 9 different labels like tillid/accept, foragt/modvilje, glæde/sindsro etc.
4. Attack: giving 1 of 2 different labels: linguistic attack/no linguistic attack
5. Rec-nition: giving 1 of 2 different labels: appreciative language/no appreciative language

# Results

1. **General tendencies for commenting**

The average ROBERTA compound score (which ranges from -1 to 1, with -1 being very negative and 1 being very positive) for all comments is 0,13 meaning neutral leaning towards positive.

Et billede, der indeholder tekst, skærmbillede, linje/række, Kurve

Automatisk genereret beskrivelse

Plot 1: Roberta Compound for all comments across weeks

Looking at the distribution of labels, 35,1% of all comments contain appreciative language, and very few comments contain some form of linguistic attack. The BERT distributions reveal that while 22,3% and 17,4% of the comments are labelled as positive emotions like ‘tillid/accept’ and ‘forventning/interesse’, 19,4% of the comments are labelled as ‘foragt/modvilje’.

Et billede, der indeholder skærmbillede, cirkel, diagram, Font/skrifttype

Automatisk genereret beskrivelseEt billede, der indeholder tekst, skærmbillede, cirkel, diagram

Automatisk genereret beskrivelse

Et billede, der indeholder tekst, skærmbillede, diagram, cirkel

Automatisk genereret beskrivelse

Plot 2: distribution of labels for all comments

1. **Sentiment of subcomments**

The intervention text encourages people to never respond to hateful comments with more hate, but on the contrary to create awareness of the hatefulness content. Study results reveal that for negative labelled comments, subcomments are mostly neutral or negative. Positive labelled comments similarly entail neutral or positive subcomments. For the intervention data, a different pattern occurs, where subcomments to positive comments are mostly negative or neutral.

Et billede, der indeholder tekst, skærmbillede, diagram, Kurve

Automatisk genereret beskrivelse Et billede, der indeholder tekst, skærmbillede, diagram, Kurve

Automatisk genereret beskrivelse

Plot 3: sentiment of subcomments for all data and for treatment data only

1. **Differences in municipality**

Noticeable differences in sentiment across the two municipalities exists.

Et billede, der indeholder diagram, Kurve, skærmbillede, tekst

Automatisk genereret beskrivelse

Plot 4: density plots for sentiment across municipalities

Density plots reveal that the majority of Århus comments are positive, less are neutral, and the minority is negative. While the same overall picture is present for Copenhagen the spread of positive, neutral and negative comments is less pronounced.

The mean Roberta compound score for comments to Århus politicians Facebook posts are approximately 0,3 while it is 0,1 for Copenhagen, meaning that the comments from Århus generally are more positive than for Copenhagen.

Et billede, der indeholder tekst, skærmbillede, diagram, Rektangel

Automatisk genereret beskrivelse

Plot 5: Mean Roberta compound score across municipalities

1. **Effect of the intervention**

The main question of investigation was to determine whether the intervention had a positive effect on the sentiment of comments. To determine this, multiple models predicting the probability for the post being an intervention post was made. After accounting for the structure and nestedness in the data (sentiment could also be determined by what post it stem from, coming from a certain politician, which comes from a certain municipality), the models reveal that when the compound score increase with 1 (more positive sentiment) the log odds of the intervention being yes actually decreases: i.e. more positive sentiment entails a lower probability of the intervention being present. This indeed points to a ‘backfire effect’, in contrast to what could be expected.

Et billede, der indeholder tekst, skærmbillede, diagram, linje/række

Automatisk genereret beskrivelse

Plot 6: Mean Roberta compound score for intervention

# Discussion and implications

Several insights to the data might serve as explanation to the backfire effect of the intervention.

Posts evolving around certain controversial political topics can at times results in a more negative and harsh comment section. The intervention posts of this study do occasionally include such topics e.g. LGBT+, racism, Lynetteholmen, wokeness, municipal budgets etc. which could explain why comments from intervention posts hold a more negative sentiment. Furthermore, it is possible that Facebook users was negatively triggered by the intervention comment itself. Looking at the intervention comments, multiple examples of such situation exist.

Et billede, der indeholder tekst, Font/skrifttype, Website, Webside

Automatisk genereret beskrivelse

Figure 2: screenshots of comments from intervention posts

Comments like these all results in a negative sentiment score.

Finally, and most likely, the backfire effect is caused by randomness. when accounting for the data structure, no significant effects of the models exists, i.e. after accounting for factors like general tendencies within a certain post, from a certain politician, or within a certain municipality, not much of the sentiment can be explained purely by the intervention itself, at least not enough to make any of the models certain of an established intervention effect. More data from intervention posts is needed to draw any final conclusions on the effect of an intervention comment. Hence it is important to emphasize that the results regarding a definite treatment effect are inconclusive. Due to the random assignments of treatment, we can’t be certain that the backfire effect is exclusively caused by the intervention itself. The results however, being based on more than 13.000 comments, provides a good indication of general tendencies and patterns for comment sections on Danish political posts, and the differences that seems evident across the two most inhabited municipalities, Århus and Copenhagen.

# Appendix

NÅR DU OPLEVER FJENDTLIGE KOMMENTARER: SIG FRA, ANMELD, STØT OP!

Vi ønsker alle platforme, hvor vi kan diskutere åbent. Men fjendtlige kommentarer på sociale medier forekommer ofte, rammer alle og kommer i mange former. At opleve fjendtlige kommentarer giver folk lyst til at trække sig fra diskussioner. Når du oplever fjendtlighed på sociale medier, kan du vise din modstand på mange måder. Ved at vise din modstand bliver du en del af løsningen! Du kan gøre en stor forskel ved at følge tre enkle råd:

Sig fra: Gør høfligt opmærksom på indlæggets fjendtlige karakter, eller prøv at vække folks medfølelse. Reager aldrig på fjendtligt indhold med mere fjendtlighed.

Anmeld: Brug platformens anmeld-funktion til at gøre platformens moderatorer opmærksomme på den fjendtlige kommentar. Støt op: Vis din støtte til modtageren af det fjendtlige indhold i en kommentar eller i en privat besked.

Når du viser din modstand mod fjendtlige kommentarer, hjælper du offeret med at bearbejde det og få diskussionen tilbage på sporet. Målet er ikke at få afsenderen af den fjendtlige kommentar til at ændre mening, men at vise at du tager afstand fra det. Det, du synes er fjendtligt, er der sandsynligvis også andre, der synes er fjendtligt. Når du viser din modstand mod fjendtlige kommentarer er du ikke alene. Du er ikke alene, når du viser din modstand mod fjendtlige kommentarer.