Bayesian analysis of the NESTA study of interventions against verbal aggression online Technical Report

Rafal Urbaniak

Contents

Data and exploration

```
Hate <- readRDS(file = "datasets/RAWNESTA/Hate.rds")
Comments <- readRDS(file = "datasets/RAWNESTA/Comments.rds")
summaries <- read.csv(file = "datasets/Summaries.csv")

dates <- colnames(Hate)[-1]
dates <- as.Date(dates)
startDate <- dates[1]
interventionDate <- "2020-07-08"
observationDate <- "2020-09-09"
end <- dates[length(dates)]

periods <- numeric(length(dates))
periods <- ifelse(dates < interventionDate, "pre-treatment", periods)
periods <- ifelse(dates >= interventionDate & dates < observationDate,
    "treatment", periods)
periods <- ifelse(dates >= observationDate, "post-treatment", periods)

hateTS <- as.data.frame(colSums(Hate[, -1]))
hateTS$date <- as.Date(rownames(hateTS))
rownames(hateTS) <- NULL
colnames(hateTS) <- c("attacks", "date")
hateTS$periods <- periods

interventions <- readRDS(file = "datasets/interventions.rds")
interventionsTS <- as.data.frame(table(interventions$day))
interventionsTS$Var1 <- as.Date(interventionsTS$Var1)
```

```
colnames(interventionsTS) <- c("date", "interventions")

periodsDF <- merge(x = hateTS, y = interventionsTS, by = "date", all.x = TRUE)

idx <- c(1, diff(periodsDF$date))
i2 <- c(1, which(idx != 1), nrow(periodsDF) + 1)
periodsDF$grp <- rep(1:length(diff(i2)), diff(i2))

periodsDF$interventions[is.na(periodsDF$interventions) & periodsDF$periods ==
    "treatment"] <- 0

periodsPlot <- ggplot(periodsDF) + geom_line(aes(x = date, y = attacks,
    group = grp), alpha = 0.8, size = 0.6) + geom_line(aes(x = date, y = interventions,
    group = grp), alpha = 0.8, size = 0.6) + geom_vline(xintercept = startDate,
    lty = 2, size = 0.2, alpha = 0.5) + geom_vline(xintercept = as.Date(interventionDate),
    lty = 2, size = 0.2, alpha = 0.5) + geom_vline(xintercept = as.Date(observationDate),
    lty = 2, size = 0.2, alpha = 0.5) + geom_vline(xintercept = as.Date(end),
    lty = 2, size = 0.2, alpha = 0.5) + labs(title = "Attacks and interventions time series",
    subtitle = "no line at data gaps", caption = "days with data: 81 (pre-treatment), 62 (treatment), 72 (post-treatment)") +
    theme_tufte() + theme(axis.title.x = element_blank(), plot.caption = element_text(hjust = 0.5,
    face = "italic")) + scale_x_date(date_labels = "%b %d", breaks = c(startDate,
    as.Date(startDate), as.Date(interventionDate), as.Date(observationDate),
    end), limits = c(startDate - 30, end + 10)) + ylab("count") + annotate("rect",
    xmin = as.Date(interventionDate), xmax = as.Date(observationDate),
    ymin = -1, ymax = 360, alpha = 0.2, fill = "darkgreen") + ylim(c(-1,
    370)) + annotate("text", label = "pre-treatment", x = as.Date(startDate) +
    2, y = 370, hjust = 0) + annotate("text", label = "treatment", x = as.Date(interventionDate) +
    2, y = 370, hjust = 0) + annotate("text", label = "post-treatment",
    x = as.Date(observationDate) + 2, y = 370, hjust = 0) + annotate("text",
    label = "interventions:", x = as.Date(interventionDate) - 55, y = 15,
    hjust = 0) + annotate("text", label = "attacks:", x = as.Date(startDate) -
    30, y = 215, hjust = 0)

periodsDF$weekdays <- weekdays(as.Date(periodsDF$date))
periodsDF$weeks <- week(as.Date(periodsDF$date))

periodsDF$weekdays <- as.factor(periodsDF$weekdays)
levels(periodsDF$weekdays) <- c("Monday", "Tuesday", "Wednesday", "Thursday",
    "Friday", "Saturday", "Sunday")
weeksPlot <- ggplot(periodsDF) + geom_smooth(aes(x = weekdays, y = attacks,
    group = 1)) + geom_line(aes(x = weekdays, y = attacks, group = weeks),
    alpha = 0.1) + theme_tufte() + labs(title = "Personal attacks throught weekdays (six months)",
    subtitle = "No weekly patterns") + xlab("") + ylab("count")
```

For the duration of the project we selected 486 Reddit users and tracked their activity (with some breaks resulting from API restrictions and technical issues, which were mostly sorted out in the observation period), starting on `rstartDate‘, beginning the intervention period on 2020-07-08, leading to a further observation period starting on 2020-09-09 and ending on 2020-11-20. The time series of attacks observed and of interventions conducted can be inspected in Figure 1.

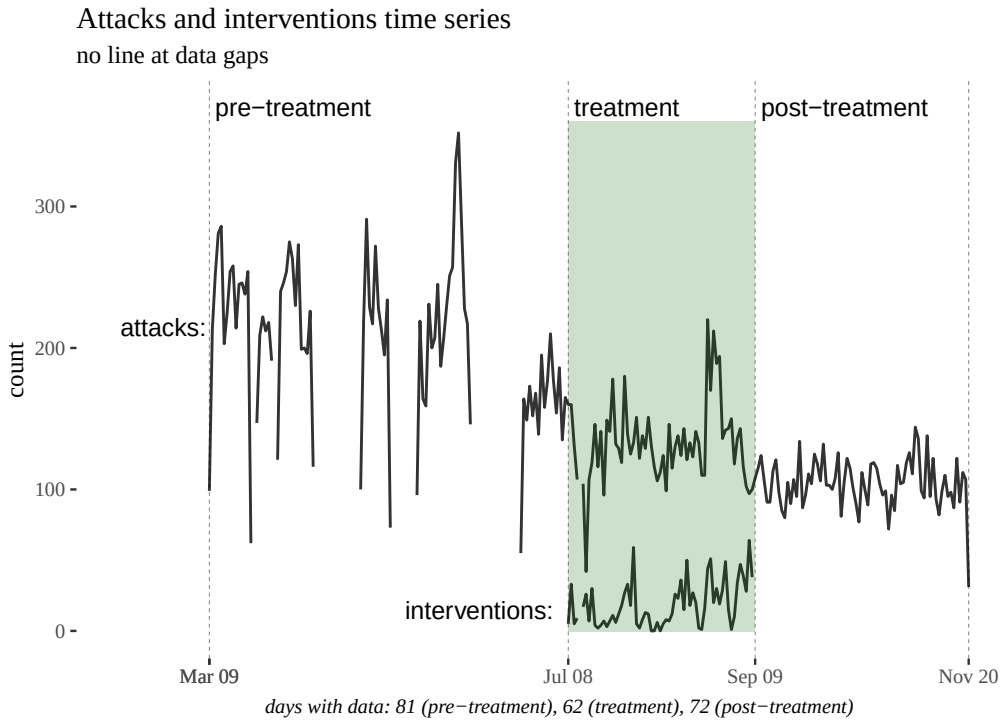

Figure 1: Daily sums of attacks and interventions throughout the three experimental periods, with GAM smoothing (blue).

Interestingly, no weekly patterns of overall aggressive behavior seem apparent, as can be seen from plotting multiple weeks alongside, as in Figure 2.

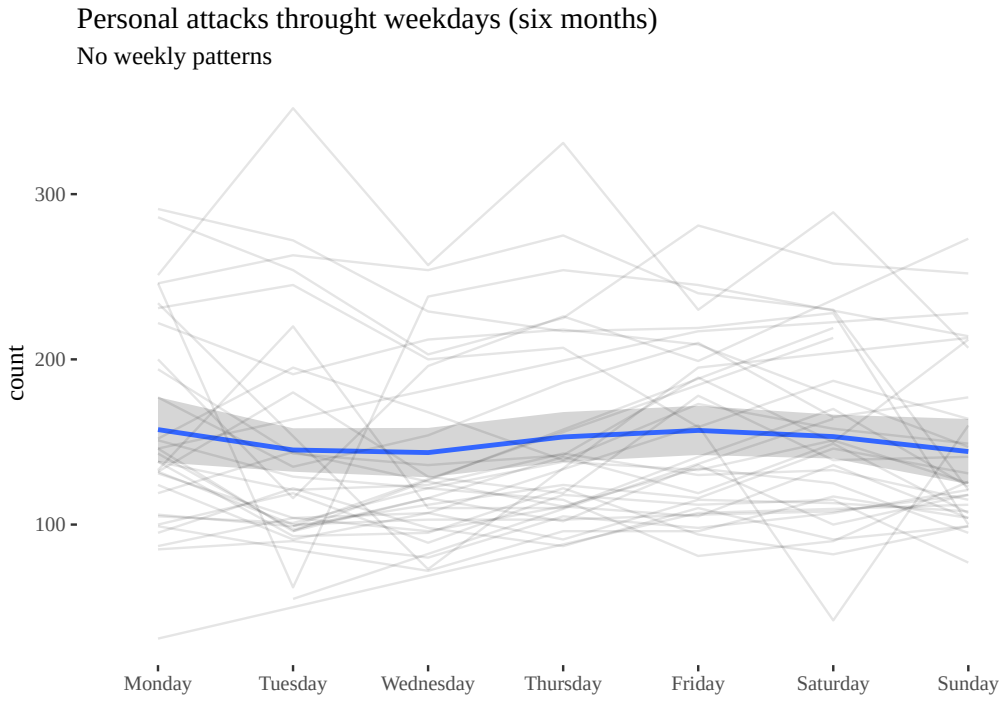

Figure 2: Attack sums from all weeks in the experimental period plotted against week days. No pattern seems to arise.

We analyzed the data from two perspective: we ran a before-and after analysis, comparing the summarized aggression levels before and after the intervention period (with various additional predictors), and a time-series perspective, which took a more fine-grained perspective. For now, we will focus on the Bayesian before-and-after analysis, for which the data were cleaned and converted into a summarized form, involving the variables listed in Table 1.

Further variables were defined in terms of those, in particular, we will be predicting AdiffS which is the standardized difference AA-AB, and AdiffS, which is the standardized difference CA-CB. The predictors were also standardized (and named ⟨variable_name⟩S), and a numerical index for the group (groupID) has been introduced.

```
summaries$ABS <- standardize(summaries$AB)
summaries$CBS <- standardize(summaries$CB)
summaries$AAS <- standardize(summaries$AA)
summaries$CAS <- standardize(summaries$CA)
summaries$CDS <- standardize(summaries$CD)
summaries$ADS <- standardize(summaries$AD)
```

variable	explanation
AB	attacks before (pre-treatment)
AD	attacks during (the treatment period)
AA	attacks after (post-treatment)
CB	comments before
CD	comments during
CA	comments after
group	treatment group
IC	intervention count

Table 1: Variables involved in the before-and-after analysis.

```
summaries$group <- as.factor(summaries$group)
summaries$groupID <- as.integer(as.factor(summaries$group))
```

The distribution of IC in the treatment groups is visualized in Figure 3. Note that the distributions are somewhat different, even though the total intervention counts are similar (110 for empathy and 119 for normative). The issue is discussed in Section XXXXX.

```
interventionsDistro <- ggplot(summaries[summaries$group != "control", ],
    aes(x = IC, fill = group)) + geom_bar() + theme_tufte() + xlab("interventions received") +
    labs(title = "Intervention counts in treatment groups") + scale_x_continuous(breaks = seq(0,
    40, 5))
```

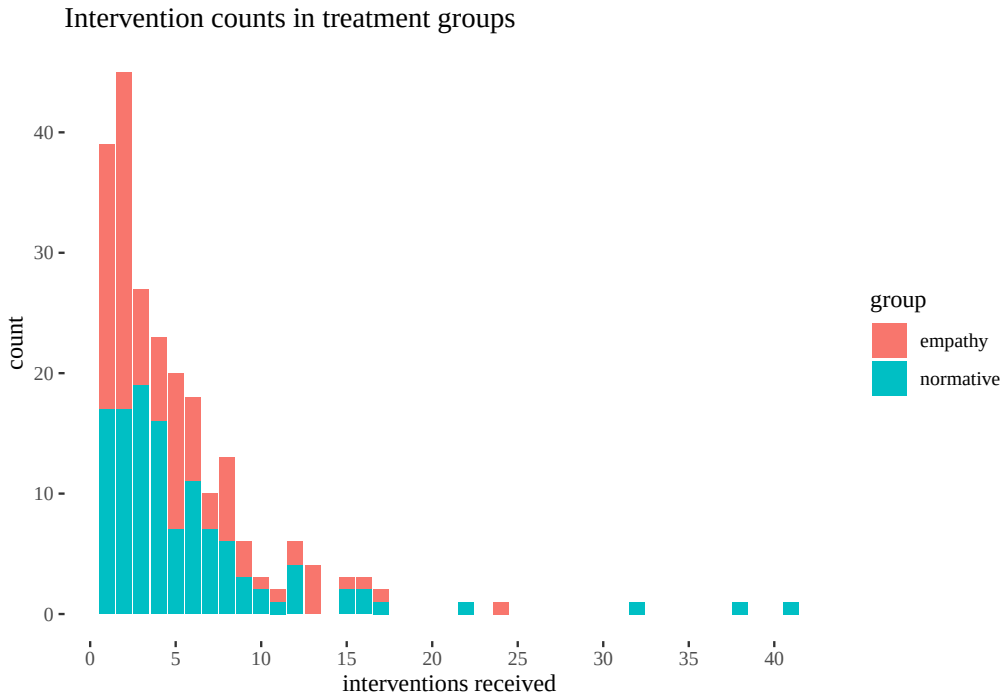

Figure 3: Distribution of daily interventions, by treatment group.

Second, when we look at the distribution of standardized difference in attacks, when restricted to (-1,1), the peaks of distributions are shifted a bit between the groups, with lowest median for the normative group, but the differences seem minor (Figure 4). This might suggest no impact of the interventions, but this conclusion would be too hasty, as the impact of other predictor variables and interactions involved can mask actual associations. We will take a closer look at this issue in our analysis.

```
violAdiffS <- ggplot(summaries, aes(x = group, y = AdiffS)) + geom_violin() +
    theme_tufte() + theme(plot.title.position = "plot")
violJoint <- ggarrange(violAdiffS + ggtitle("whole range"), violAdiffS +
    ylim(c(-1, 1)) + geom_boxplot(width = 0.2) + ggtitle("restricted to (-1,1)")) +
    theme(plot.title.position = "plot")
violJointTitled <- annotate_figure(violJoint, top = text_grob("Empirical distribution of change in attacks (standa
    size = 12))
```

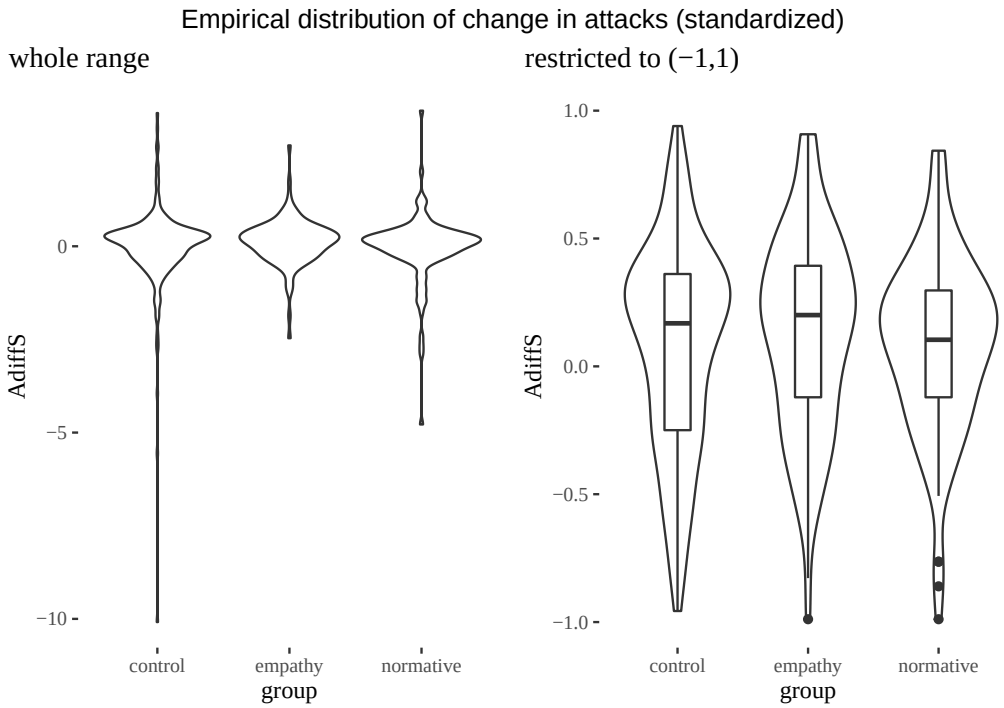

Figure 4: Empirical distribution of change in attacks, by treatment group.

To see how this masking can occur, let's inspect changes in attacks against intervention counts. It turns out that restricting attention to various activity levels results in fairly strong changes to the regression lines (Figure 5).

```
icplot1 <- ggplot(summaries, aes(x = IC, y = AdiffS, color = group, fill = group)) +
    geom_jitter(alpha = 0.6, size = 0.8) + theme_tufte() + theme(plot.title.position = "plot") +
    geom_smooth(alpha = 0.2, method = "lm") + xlim(c(0, 25)) + ylim(c(-2,
    2)) + ggtitle("sd restricted to (-2,2)") + theme(legend.position = c(0.65,
    0.2))

icplot2 <- ggplot(summaries, aes(x = IC, y = AdiffS, color = group, fill = group)) +
    geom_jitter(alpha = 0.6, size = 0.8) + theme_tufte() + theme(plot.title.position = "plot") +
    geom_smooth(alpha = 0.2, method = "lm") + xlim(c(0, 25)) + ylim(c(-1,
    1)) + ggtitle("sd restricted to (-1,1)") + theme(legend.position = c(0.65,
    0.2))

icplotJoint <- ggarrange(icplot1, icplot2)
icplotTitled <- annotate_figure(icplotJoint, top = text_grob("Change in attacks (standardized) vs interventions re
    size = 12))
```

Some interactions are also suggested by the differences in linear smoothing when

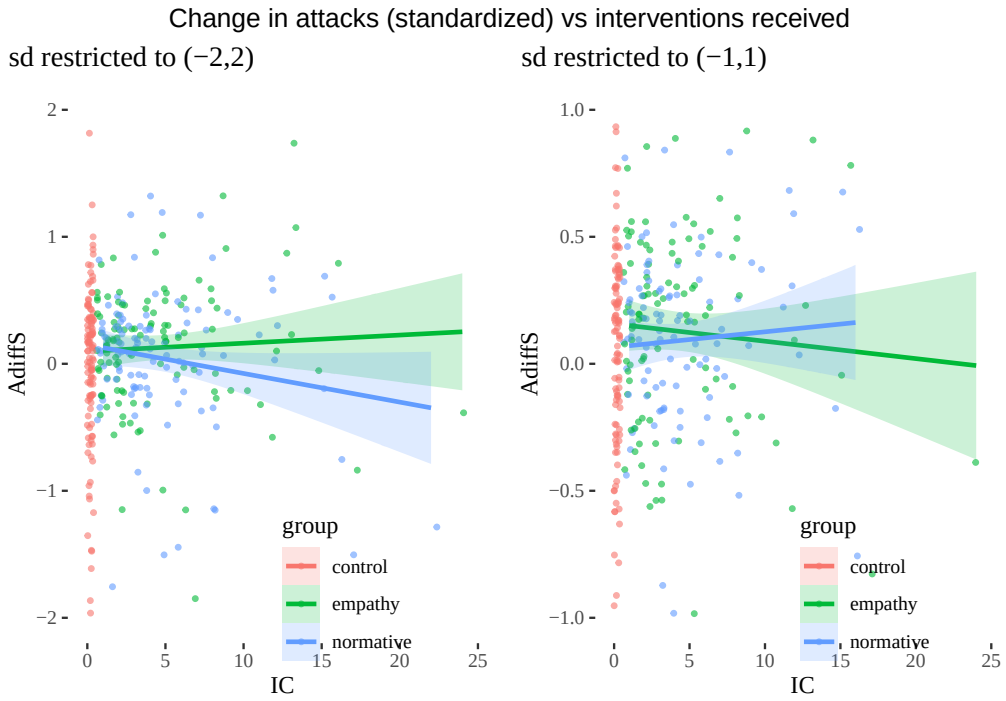

Figure 5: Change in attacks vs intervention counts by treatment group, jittered with linear smoothing.

attention is restricted when it comes to change in comments.