Application to unlabeled data

Cornelius Erfort

8/5/2021

Contents

1 Summary

We predict issue labels for all unlabeled German press releases and calculate the share of press releases dedicated to each issue area for each quarter.

2 Setting up

This script requires the files which are not included on GitHub.

At the end of this script, the file “issue_agendas.RData” is saved. It contains quarterly estimates for the share of press releases for each issue and party.

2.1 Loading packages

```r
start_time <- Sys.time()

packages <- c("quanteda", "quanteda.textmodels", "dplyr", "caret", "randomForest",
    "tm", "rmarkdown", "plyr", "readr", "ggplot2", "stringr", "formatR", "readstata13",
    "lubridate", "reticulate", "doMC", "glmnet", "kableExtra", "stargazer", "extrafont")

lapply(packages[!(packages %in% rownames(installed.packages()))], install.packages)

if (!("quanteda.classifiers" %in% rownames(installed.packages()))) {
    remotes::install_github("quanteda/quanteda.classifiers")
}

invisible(lapply(c(packages, "quanteda.classifiers"), require, character.only = T))
```

```
loadfonts()
loadfonts(device = "pdf")
theme_update(text = element_text(family = "LM Roman 10"))  # Set font family for ggplot

if (!dir.exists("supervised-files")) dir.create("supervised-files")

source("scripts/functions.R")
```

3 Classification of unlabeled data

3.1 Using the fine-tuned Transformers

We trained the models using a set of 2,612 labeled documents. In order to obtain aggregated measures of issue attention, we predict the issue categories of all ? labeled and unlabeled press releases in our sample.

```
# Load the predicted labels
alldocs <- read_csv("transfer-files/alldocs-pred.csv", col_names = F)
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   X1 = col_double(),
##   X2 = col_double()
## )
```

```
names(alldocs) <- c("label", "id")

# Translate labels back into CAP issues
labels <- read_csv("transfer-files/bert-pred.csv", col_names = F)[, 2:3] %>%
    unique
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   X1 = col_double(),
##   X2 = col_double(),
##   X3 = col_double(),
##   X4 = col_double()
## )
```

```
names(labels) <- c("issue_r1", "label")
labels$issue_r1[labels$issue_r1 == 191] <- 19.1
labels$issue_r1[labels$issue_r1 == 192] <- 19.2
labels$issue_r1 <- factor(labels$issue_r1, levels = c(1:7, 9:10, 12, 15:17, 19.1,
    19.2, 20, 99))
alldocs <- merge(alldocs, labels, by = "label") %>%
    select(-c(label))

# Load and merge unlabeled data
load("data/all/germany.RData")
nrow(germany)
```

```
## [1] 44950
```

```
alldocs <- merge(germany, alldocs, by = "id")
nrow(germany)
```

issue	1	2	3	4	5	6	7	9	10	12	15	16	17	19.1	19.2	20	99
n	2529	3083	1933	1681	2806	2457	3675	2318	3469	3321	3100	2266	1126	5595	2430	1298	1863

```
## [1] 44950
```

```
# Table of predicted issues
table(alldocs$issue_r1) %>%
    as.data.frame() %>%
    dplyr::rename(issue = Var1, n = Freq) %>%
    t() %>%
    kbl(booktabs = T) %>%
    kable_styling(latex_options = "scale_down")
```

```
table(alldocs$issue_r1)/nrow(alldocs)
```

```
##
##          1          2          3          4          5          6          7
## 0.05626251 0.06858732 0.04300334 0.03739711 0.06242492 0.05466073 0.08175751
##          9         10         12         15         16         17       19.1
## 0.05156841 0.07717464 0.07388209 0.06896552 0.05041157 0.02505006 0.12447164
##       19.2         20         99
## 0.05406007 0.02887653 0.04144605
```

```
# Clean party names
party_names <- data.frame(party = c("90gruene_fraktion", "afd_bundesverband", "afd_fraktion",
    "fdp_bundesverband", "fdp_fraktion", "linke_fraktion", "spd_fraktion", "union_fraktion"),
    party_name = c("B'90/Die Grünen", "AfD", "AfD", "FDP", "FDP", "DIE LINKE", "SPD",
        "CDU/CSU"))
alldocs <- merge(alldocs, party_names, by = "party")
nrow(alldocs)
```

```
## [1] 44950
```

```
# Tables for samples of press releases Environment
sample7 <- select(alldocs, c("party_name", "date", "header", "issue_r1")) %>%
    filter(issue_r1 == 7)
(sample7 <- sample7[sample(1:nrow(sample7), 10), ])
```

```
##           party_name       date
## 844  B'90/Die Grünen 2010-03-01
## 2317       DIE LINKE 2010-03-03
## 2399       DIE LINKE 2010-11-30
## 727  B'90/Die Grünen 2011-05-18
## 990  B'90/Die Grünen 2010-10-27
## 2883             SPD 2011-05-19
## 1985       DIE LINKE 2014-06-04
## 382  B'90/Die Grünen 2014-06-12
## 1946       DIE LINKE 2015-01-16
## 2958             SPD 2010-10-01
##                                                                                                  heade
## 844             Artenschutztag: Roten Thunfisch vor dem Aussterben schützen, Elfenbeinhandel verbiet
## 2317                                   Kürzungen der Solarförderung hemmt Ausbau erneuerbarer Energi
## 2399                  Subventionsforderungen der Automobilindustrie in Milliardenhöhe sind unverschä
## 727                         EEG-Novelle: Röttgen gegen beschleunigten Ausbau der erneuerbaren Energi
## 990                        Bundesregierung fördert Mietenexplosion und gefährdet sozialen Zusammenhal
## 2883                                      Die Energiewende fällt aus: Schwarz-Gelb täuscht und tricks
## 1985                                      Ex-Umweltminister Gabriel plant Einfallstor für Gas-Frackin
## 382                                                     Atomkraft: Hermesbürgschaften endlich gestopp
## 1946 Brunsbüttel-Urteil macht Entsorgungsnachweis für alle Atommeiler obsolet und erzwingt Abschaltun
## 2958                                    Gorleben: Merkel und Röttgen machen weiter wie Kohl und Merke
##      issue_r1
## 844         7
## 2317        7
## 2399        7
## 727         7
## 990         7
## 2883        7
## 1985        7
## 382         7
## 1946        7
## 2958        7
```

```
latex_out <- capture.output(sample7 %>%
    dplyr::rename(party = party_name, title = header) %>%
    stargazer(type = "latex", summary = F, rownames = F, title = "Sample of press releases classified a
        label = "tab:7-document-samples"))

latex_out <- capture.output(latex_out %>%
    str_replace_all("tabular", "tabularx") %>%
    str_replace_all("\\@\\{\\\\extracolsep\\{5pt\\}\\} ccc", "\\\\textwidth\\}\\{stX") %>%
    cat(sep = "\n"), file = "tables/7-document-samples.tex")

# Immigration
sample9 <- select(alldocs, c("party_name", "date", "header", "issue_r1")) %>%
    filter(issue_r1 == 9)
(sample9 <- sample9[sample(1:nrow(sample9), 10), ])
```

```
##           party_name       date
## 1051             FDP 2018-10-04
## 1507       DIE LINKE 2015-09-16
## 1127             FDP 2012-04-24
## 974              FDP 2015-09-17
## 83   B'90/Die Grünen 2016-12-14
## 591              AfD 2018-01-15
## 2290         CDU/CSU 2011-09-05
## 969              FDP 2015-04-16
## 1608       DIE LINKE 2014-06-11
## 1300       DIE LINKE 2017-03-22
##                                                                                                    hea
## 1051                                                        Eine sprachliche Einigung auf unterem Niv
## 1507                                            Legale Wege für Flüchtlinge statt Soldaten im Mitteln
## 1127                                            Fachkräfte-Zuwanderung verbessert Lage bei Mangelberu
## 974                                                                                 Alte Fehler vermei
## 83                                                                      Einwanderung nachhaltig gestal
## 591  Bernd Baumann: Die anderen Parteien laufen der AfD hinterher - Rückkehrprozess von Syrern beginn
## 2290                                                              Vorgehen von Pro Asyl nicht akzepta
## 969   ZIMMERMANN an die Mitglieder des Deutschen Bundestages: Soforthilfe-Fonds für Flüchtlinge aufle
## 1608                       Großzügige Aufnahmeregelung für syrische Flüchtlinge ist das Gebot der Stu
## 1300                                                Schluss mit der Kriminalisierung von Schutzsuchen
##      issue_r1
## 1051        9
## 1507        9
## 1127        9
## 974         9
## 83          9
## 591         9
## 2290        9
## 969         9
## 1608        9
## 1300        9
```

```
latex_out <- capture.output(sample9 %>%
    dplyr::rename(party = party_name, title = header) %>%
    stargazer(type = "latex", summary = F, rownames = F, title = "Sample of press releases classified a
        label = "tab:9-document-samples"))

latex_out <- capture.output(latex_out %>%
    str_replace_all("tabular", "tabularx") %>%
    str_replace_all("\\@\\{\\\\extracolsep\\{5pt\\}\\} ccc", "\\\\textwidth\\}\\{stX") %>%
    cat(sep = "\n"), file = "tables/9-document-samples.tex")
```

3.2 Aggregation of the issues categories over time and party

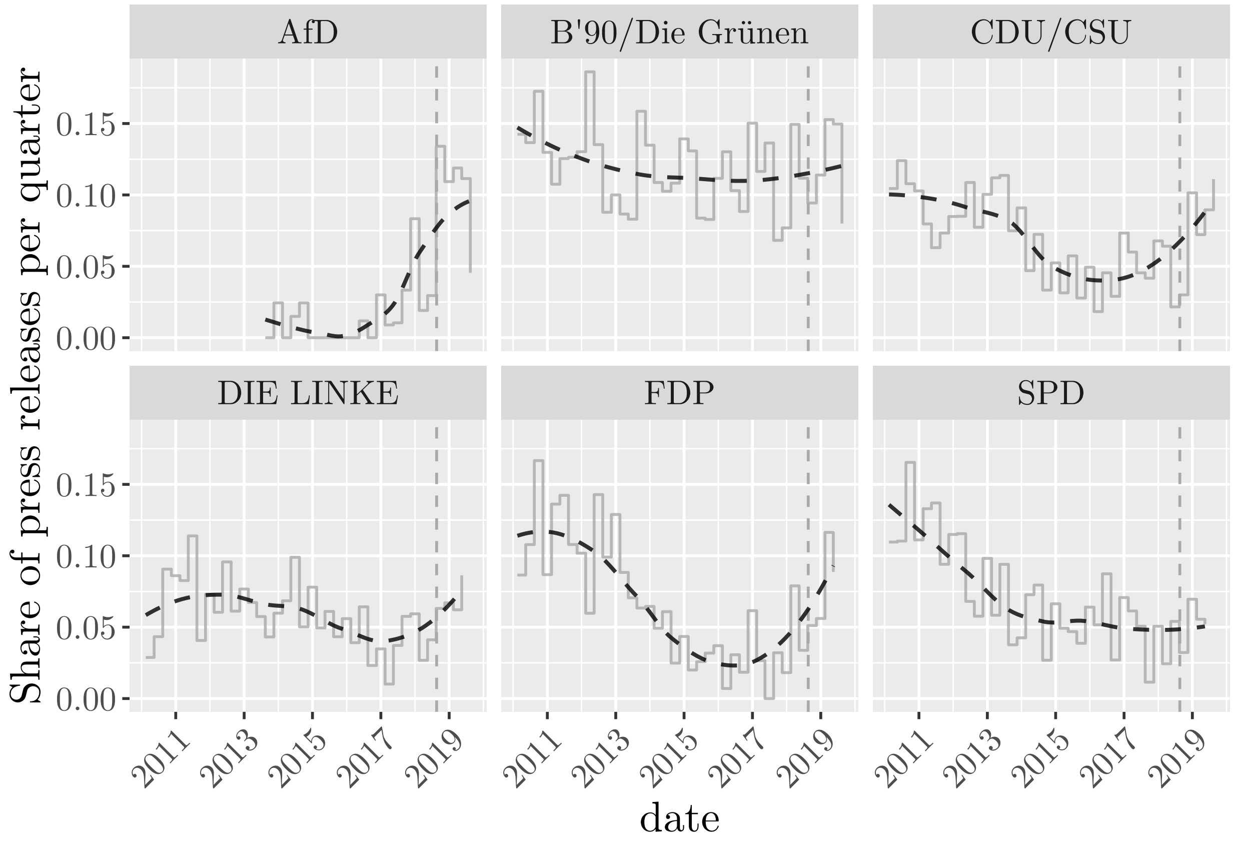
To measure parties' evolving issue agendas, we aggregate the category counts over time.

```
# Create dataframe with only necessary vars
issue_agendas <- alldocs %>%
    select(c(date, issue_r1, party_name)) %>%
    dplyr::rename(party = party_name)

# Make date quarterly
issue_agendas$date <- as.character(issue_agendas$date) %>%
    substr(1, 8) %>%
    str_c("15") %>%
    str_replace_all(c(`-01-` = "-02-", `-03-` = "-02-", `-04-` = "-05-", `-06-` = "-05-",
        `-07-` = "-08-", `-09-` = "-08-", `-10-` = "-11-", `-12-` = "-11-")) %>%
    ymd()

# Add variable for counting
issue_agendas$freq <- 1

# Aggregate by party, date and issue
issue_agendas <- aggregate(freq ~ party + date + issue_r1, issue_agendas, sum)

# Add observations with zero documents
for (thisparty in unique(issue_agendas$party)) {
    for (thisdate in unique(issue_agendas$date[issue_agendas$party == thisparty])) {
        for (thisissue in unique(issue_agendas$issue_r1)) {
            if (nrow(issue_agendas[issue_agendas$party == thisparty & issue_agendas$date ==
                thisdate & issue_agendas$issue_r1 == thisissue, ]) == 0 & nrow(issue_agendas[issue_agend
                thisparty & issue_agendas$date == thisdate, ]) != 0) {
                issue_agendas <- data.frame(party = thisparty, date = thisdate, issue_r1 = thisissue,
                  freq = 0) %>%
                  rbind.fill(issue_agendas)
            }
        }
    }
}

# Add var for total press releases per party and month
issue_agendas$party_sum <- ave(issue_agendas$freq, issue_agendas$date, issue_agendas$party,
    FUN = sum)

issue_agendas$attention <- issue_agendas$freq/issue_agendas$party_sum

# Add issue descriptions
issue_categories <- data.frame(issue_r1 = c(1:7, 9:10, 12, 15:17, 191:192, 20, 99),
    issue_r1_descr = c("Macroeconomics", "Civil Rights", "Health", "Agriculture",
        "Labor", "Education", "Environment and Energy", "Immigration", "Welfare",
        "Law and Crime", "Commerce", "Defense", "Technology", "International Affairs",
        "EU", "Government Operations", "Other"))

issue_agendas <- merge(issue_agendas, issue_categories, by = "issue_r1") %>%
    select(-c(freq))

issue_agendas$date <- issue_agendas$date %>%
    as.Date(origin = "1970-01-01")

save(issue_agendas, file = "data/issue_agendas.RData")
```

4 Visualize issue agendas

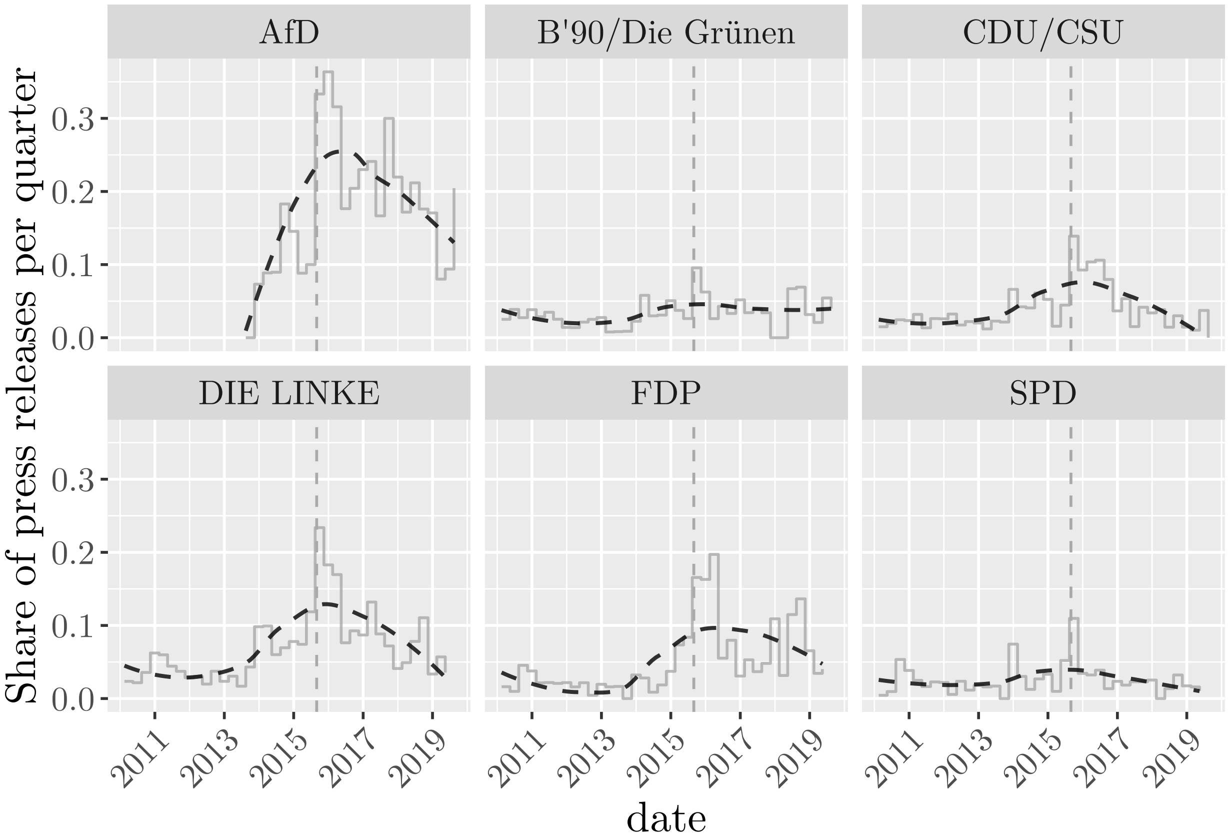
```
if (!dir.exists("plots")) dir.create("plots")

## Facet (all parties separate) Environment and Energy
plot_issue_agenda(issue_agendas, 7, unique(issue_agendas$party), T)
```

```
## [1] "7 - Environment and Energy_all-parties_facet"
```



```
# Immigration
plot_issue_agenda(issue_agendas, 9, unique(issue_agendas$party), T)
```

```
## [1] "9 - Immigration_all-parties_facet"
```



```r
# Time needed to run script
print(Sys.time() - start_time)
```

```
## Time difference of 36.57003 secs
```