Replication materials for

“Distinctive Voices: Political Speech, Rhetoric and the Substantive Representation of Women in European Parliaments”

forthcoming in *Legislative Studies Quarterly*

by Jens Wäckerle and Bruno Castanho Silva

This README provides an overview of the replication materials for the article. The **Data** section describes the datasets required to reproduce the figures and tables in the paper and in the Online Appendix. The **Analysis** section summarises the purpose of the R scripts to reproduce the analyses in the article and the Online Appendix. All analyses were conducted with R 4.2.2 running on macOS Monterey 12.4. The following versions of R packages were used:

car 3.1-1
effects 4.2-2
grid 4.2.2
gridExtra 2.3
lme4 1.1-31
magrittr 2.0.3
quanteda 3.2.3
stargazer 5.2.3
texreg 1.38.6
tidyverse 1.3.2
yardstick 1.1.0

Data

All data required for the analyses in this paper is contained in the file **df_analysis.RData**. When opening this file in R, the following objects will be present:

- **df.all –** dataset to replicate all figures and tables in the manuscript and Online Appendix, with the script Analysis.Rmd. More details of the variables contained in it below.

- **dfm.de, dfm.es, dfm.ie, dfm.nl, dfm.se –** each one is a *document-feature-matrix (dfm)* produced from the corpora of speeches from Germany, Spain, Ireland, Netherlands, and Sweden, respectively. They are described in more detail below.

df.all

This is a dataset with 240010 rows and 51 columns, listed below. Each row corresponds to one speech. The variables are the following:

docid: document ID, unique for each speech within country and can be matched to the respective dfm.

policy_area_ml: policy area of that speech, coded based on the title of the parliamentary debate in which it was given. For details on how policy areas were coded, please refer to Online Appendix A.

date: date the speech was given, in YYYY-MM-DD format.

leg: legislative session in which the speech was given.

gender: gender of the speaker

party: party of the speaker

pred.glmnet: predicted probability of speech being given by a woman, from the Ridge regression.

pred.boost: predicted probability of speech being given by a woman, from the Xgboost model.

pred.oos: out-of-sample predicted probability of speech being given by a woman, from the Ridge regression.

pred.boost.oos: out-of-sample predicted probability of speech being given by a woman, from the Xgboost model.

speech.length: number of words in the speech

pron.r: proportion of speech that is composed of pronouns

verb.r: proportion of speech that is composed of verbs

adj.r: proportion of speech that is composed of adjectives

noun.r: proportion of speech that is composed of nouns

adv.r: proportion of speech that is composed of adverbs

Tentat.r: proportion of speech that is composed of tentative words

Social.r: proportion of speech that is composed of social processes words

Posemo.r: proportion of speech that is composed of positive emotion words

Negemo.r: proportion of speech that is composed of negative emotion words

Informal.r: proportion of speech that is composed of informal words

government_policy_area: combination of **leg** and **policy_area_ml**

value_Female: For that policy area and time, how many women mention it as one of the most important problems in the country in Eurobarometer

value_Male: For that policy area and time, how many men mention it as one of the most important problems in the country in Eurobarometer

gender_diff: value_Female minus **value_Male**

country_label: country

first_entered_parliament: year when MP who gave the speech first entered parliament

family: family of the speaker's party

name: speaker's name

position: which position in cabinet the speaker holds, if any (from the WhoGov dataset, Nyrup and Bramwell, 2020)

minister: binary, whether the speaker of that speech held a cabinet position at the time (1) or not (0)

classification: what kind of ministry the person held, from the WhoGov dataset, Nyrup and Bramwell (2020)

portfolio_1: first portfolio the speaker held at the time, if any (from the WhoGov dataset, Nyrup and Bramwell, 2020)

portfolio_2: second portfolio the speaker held at the time, if any (from the WhoGov dataset, Nyrup and Bramwell, 2020)

portfolio_3: third portfolio the speaker held at the time, if any (from the WhoGov dataset, Nyrup and Bramwell, 2020)

female: binary gender variable, taking the value of 1 if speaker is a woman or 0 if a man

years_parl: how many years the speaker had been in parliament when delivering that speech

pred.glmnet100: **pred.glmnet** multiplied by 100, to make model convergence easier

pred.boost100: pred.boost multiplied by 100, to make model convergence easier

pred.glmnet.oos100: pred.oos multiplied by 100, to make model convergence easier

pred.boost.oos100: pred.boost.oos multiplied by 100, to make model convergence easier

gender_diff100: gender_diff multiplied by 100, to make model convergence easier

value_Female100: value_Female multiplied by 100, to make model convergence easier

speech.length10: speech.length divided by 1000, to make model convergence easier.

speech.length2: speech.length10 squared

prop.w: proportion of women in parliament in that legislative session

prop_area: proportion of women among all speeches in that policy area in that legislative session.

womens_minister: whether the MP held one of the following portfolios: Children & Family; Education, Training & Skills; Health & Social Welfare; Ageing and Elderly; Youth.

gender_diff_bi: binary transformation of **gender_diff**, whether it is higher than 0 (1) or lower than 0 (0)

male: binary whether speaker is a man (1) or a woman (0).

dfms

These are *document-feature-matrices*, meaning each contains a matrix where each row is one speech, and each column one word that appears in any of the speeches in that country. The cells are filled with the number of times that specific word appears in the specific speech. They are used in the models to obtain the predicted probabilities that each speech was given by a woman, the main dependent variable in the paper. The script for that is in the file **Running_ML_models.Rmd**, described in more detail below. They are also used for tables D.1-D.5 in the Online Appendix, reproduced in the file **Analysis.Rmd**, described in detail below.

dfm.de: Document-feature matrix of: 10,051 documents, 164,106 features, from German speeches

dfm.es: Document-feature matrix of: 8,443 documents, 70,188 features, from Spanish speeches

dfm.ie: Document-feature matrix of: 30,101 documents, 63,225 features, from Irish speeches

dfm.nl: Document-feature matrix of: 143,047 documents, 182,479 features, from Dutch speeches

dfm.se: Document-feature matrix of: 39,756 documents, 177,099 features, from Swedish speeches

All dfm's contain document-level variables, corresponding to each speech. All dfm's have the same set of 10 covariates:

member_name: speaker's name

leg: legislative session in which the speech was given

policy_area_ml: policy area of that speech, coded based on the title of the parliamentary debate in which it was given. For details on how policy areas were coded, please refer to Online Appendix A.

date: date the speech was given, in YYYY-MM-DD format.

leg: legislative session in which the speech was given.

gender: gender of the speaker

party: party of the speaker

speech.length: number of words in the speech

docid: document ID, unique for each speech within country and can be matched to the **docid** in **df.all**

debate_title: title of the parliamentary debate in which the speech was given

Analysis

The file "**Analysis.Rmd**" is an Rmarkdown file containing the script with the code to reproduce all Figures and Tables in the manuscript and Online Appendix. It produces the **"Analysis.html"**, which contains all Figures and Tables in their appropriate formatting as seen in the paper and Online Appendix.

If the user has both the **"Analysis.Rmd"** and the R file **"df_analysis.RData"** in the same folder, which is also their working directory, the script will run exactly as is.

The file "**Running_ML_models.Rmd"** can be used to re-run the machine learning models that produce the predicted probabilities used as dependent variable in the paper. It runs the models for each country, and shows that the results match exactly those in the **df.all** dataset already provided for the replication of Figures and Tables.

NB! This script **takes 48-72 hours to completely run!** It will run as is if it is in the same folder as **df_analysis.RData**, given that folder is also the working directory. It is only present to show how to obtain the estimates used in the paper, but **all figures and tables can be produced by running the Analysis.Rmd file alone.**