Pre-Registration Report: Facing Election

Mathias Rask[*] Stig Hebbelstrup[†] Robert Klemmensen[‡]

November 16, 2021

[*] Aarhus University, Political Science. Email: mathiasrask@ps.au.dk.

[†] Aarhus University, Political Science. Email: stighebbelstrup@ps.au.dk.

[‡] University of Southern Denmark, Political Science. Email: rkl@sam.sdu.dk.

1 — Introduction

The core of most definitions of liberal democracy holds that voters elect representatives to represent their political orientations. Consequently, the legitimacy of electoral liberal democracy hinges on voters knowing the differences between parties and their policies in order to choose the party which best serves their preferences. This description of representation is based on voters rationally assessing their interests and that they are able to map their preferences to policies proposed by parties or candidates.

This model of representation has been challenged by at least three streams of literature showing that we need to take into account how nonverbal cues affect voters and that individual differences are important and somewhat overlooked features of representation. The first stream documents that voters are not equally interested in politics and that differences in political interest and sophistication moderate the extent to which individuals make ideologically informed vote choices (Carpini and Keeter 1996; Goren 1997; Lau and Redlawsk 2006; Neuman and Neuman 1986). Voters are not as intrinsically interested in politics as normative theories of representation assume. Therefore, the ideological orientations of disinterested voters may not be represented as effectively and fully as those of voters interested in politics. This literature suggest that some voters might only access low cost information such as campaign material during elections etc.

Secondly, an important literature suggest that politics have become increasingly personalized. Caprara and Zimbardo (2004) and Pedersen and Rahat (2021) argued that contemporary politics have become more personalised and less ideological (Hayes 2009; McAllister et al. 2007). With increasing individualisation, less rigid social identities, and a more mediatized political arena, voters may focus more on the perceived personal characteristics of candidates than the party platform on which they run. Thus, this stream suggest that how candidates present themselves is increasingly important relative to programmatic statements.

Finally, there is an emerging consensus that voters do not evaluate politicians and policies as rationally as most models of representation suggest. For instance, Lodge and Taber (2013) argue that voters evaluate political candidates and policies based on emotional cues. Numerous studies suggest motivated political reasoning bias voters' evaluation of facts as (Bisgaard 2019; Gaines et al. 2007; Kahan 2012) and their ability to correctly attribute responsibility for political actions (Achen et al. 2017; Tilley and Hobolt 2011).

The common denominator of these processes is the emotions that candidates and policies evoke in voters. Consequently, we expect that concrete candidate images and impressions may spur stronger reactions than abstract party programs which are mainly digested by politically interested voters and which requires a more cognitively more challenging effort. Voters do not need to be highly politically interested or cognitively sophisticated to form an opinion of the likeability, warmth or masculinity of a candidate from a picture which is part of some campaign material (Caprara and Zimbardo 2004). Furthermore – as noted – we know that voters form opinions fast based on nonverbal cues.

To answer some of the issues that these three streams of literature raise Todorov et al. (2005, 2015) proposed voters evaluate candidates visually rather than through ideological and policy attitudes. Interestingly, Todorov and co-authors found that voters are capable of detecting signals from faces on how competent candidates are perceived to be and that this perception competence predicts electoral success. Furthermore the evaluations of faces is surprisingly fast indicating that the perceived signal is not the result of a rational and deliberate process (Willis and Todorov 2006).

Thus, voters are fast in making *subjective* evaluations of candidates (Todorov et al. 2005) and these impressions have electoral consequences. However, surprisingly no study has, to our knowledge, investigated if *objective* facial features predict if some candidates are more likely to be elected than others. This is exactly what we investigate in this study: *Do objective facial features predict electoral success?*

This report proceeds as follows. In the next section we develop our hypotheses. Then we describe our data and the models that we trained and we plan to validate on the 2021 material.

1.1. Hypotheses

The current literature suggests that *subjective* evaluations of faces are important in predicting electoral outcomes (Todorov 2017; Todorov et al. 2005). However, in a series of papers Kosinski and co-authors show demonstrate that *objective* facial features can predict the sexual orientation of an individual (Wang and Kosinski 2018) and more recently the ideological leanings of voters (Kosinski 2021). Boussalis and Coan (2021) have demonstrated that objectively measured facial expressions are related to voter's perceptions of candidates and that politicians of different genders display different emotions. Hence there is a burgeoning literature arguing that objective facial features are important if we want to gain a deeper understanding of the consequences of nonverbal communication in the politics (Carpinella et al. 2016; Everitt et al. 2016; Stewart et al. 2009). Given the most recent advances in the literature we expect that:

> H1: We can predict the electoral success of candidates above random chance using objective measures of faces

While we do expect that faces are an essential part of the nonverbal communication taking place between candidates and voters, we do not expect that facial morphology is responsible for our prediction of electoral success. Todorov (2017) convincingly argues that there is no evidence that facial morphology can explain variation in political ideology.

> H2: Electoral success is correlated with facial expressions of emotions (facial expressions)

However, Boussalis and Coan (2021) show that there are significant differences in the which emotions candidates for political office display depending on they gender. Building on role congruence theory they confirm their hypotheses that women display more happy facial emotions compared to men and that men display more emotion

> H3: There is a gender bias such that we are better at predicting the electoral success for female than for male candidates.

Our two next hypotheses are concerned with which are gender specific that links facial traits to electoral success. Berggren et al. (2010) and Lawson et al. (2010) identify a so called beauty bonus for candidate for office in as diverse contexts as Finland, Mexico and Brazil. In these countries there are more candidate beauty predicts electoral success and for women the effect is particularly pronounced, which leads to hypothesis 4a below. Another facial trait that has gender specific traits is masculinity. A series of papers show that candidates how are more

> H4a: Electoral success is correlated with the beauty for female candidates.

> H4b: Electoral success is correlated with masculinity for male candidates.

Our final hypothesis examines the interplay between facial expressions and electoral success. Parties can use to different types of lists when taking part in elections. When parties use a closed list format, partisan determine the order of candidates such that the top candidate has higher chances of gaining representation. When parties employ open lists only voters determine who gets elected and the candidate's position on the list is, theoretically, inconsequential for who gain representation. Candidates representing parties using open lists have not been selected by partisans and hence the way voters form opinions on who to vote for is more consequential for these candidates compared to candidates who face the voters from a closed list. Therefore we expect:

> H5: We are better at predicting the electoral success in open-list than in closed-list municipalities.

Our two final hypotheses serve as robustness check on our model. The above hypotheses are based an a dichotomous measure of eithe gaining representation or not. If the expectations are true we should also expect that the hypotheses above hold when we substitute the last layer in our model with a layer including votes shares won by individual candidates.

> H6a: We expect to reproduce predictions based in hypotheses 1 to 5 using the vote shares of candidates.

The final robustness check concerns comparing candidates who ran for office in 2017 and 2021 to new candidates. For our predictions to pick up a true signal we expect that there is not significant difference across the predictors identified in hypotheses 1 through 5

> H6b: We expect to reproduce predictions based in hypotheses 1 to 5 when comparing candidates re-running for office compared to candidates running for office for the first time.

2 — Methods

2.1. Pilot data (training and validation data)

To test our initial expectations, we used pictures of candidates and personal votes obtained from the 2017 Danish municipal election. A total of $9,556$ candidates ran in the 2017 elections and we obtained pictures for $5,184$ of the candidates from the so called *candidate test* hosted by the Danish Broadcasting Corporation (DR). We manually reviewed each of the images to judge whether they were of sufficient quality to analyze using computer vision techniques. For instance, we dropped images that contains multiple persons, dogs, and other objects not related to our focus. After reviewing the images, we were left with $4,699$ pictures. We used the $4,699$ images to train a set of models, one for male candidates and one for female candidates, to test our hypotheses on a validation. We then seek to replicate our findings (whether its a null finding or not) on the images of candidates running in the 2021 local elections in order to test if our expectations generalize. We pre-register both of our models on GitHub. We apply a set image preprocessing steps for both the 2017 and 2021 data. See Section 2.4.2 for details or our `.py`-scripts available on our pre-registered GitHub for further details.

2.2. Design

We apply *Convolutional Neural Networks* (Goodfellow et al. 2016) to predict the electoral success of political candidates using only their facial images. Instead of detecting specific facial landmarks (e.g. the position of the nose) or estimates of the head pose, CNNs learn millions of abstract features that in combination increases or decreases the probability that the candidate is elected or not – if facial features matter for electoral success. Thus, we assume that no single features like the size of the nose or how much a candidate smiles discriminates whether a candidate is elected or not. Rather, the link between the faces of candidates and electoral success are a function of a combination of multiple abstract facial features learned by the CNNs.

It is generally known that deep learning requires large sample sizes to achieve proper generalization. In particular, this is the case if the data is noisy (low signal-to-noise ratio), the target function $f : \mathcal{X} \to \mathcal{Y}$ ($\mathcal{X} =$ faces of candidates and $\mathcal{Y} =$ electoral success) is difficult to learn, or if the target function f if $\mathcal{X} \to \mathcal{Y}$ is only weakly related. In our application, we expect both difficulties to be present. We bot expect that the signal-to-noise ratio is low *and* that the target function f is difficult to learn as the link between the faces of candidates and electoral success is presumably weak compared to other explanatory factors of vote choice. Further, we have a fairly small data set of a total of $4,699$ images for both male and female candidates. As we describe in a second, we train separate models for male and female candidates meaning that our effective training data is reduced even further. Together, this makes our application a difficult learning task.

To accommodate the difficult learning task due to the combination of a small training set and a weak related target function f, we apply transfer learning shown to achieve state-of-the-art results using CNNs (Yang et al. 2020). The idea is to take an existing network and its weights (parameters) and either use the weights directly (i.e. as a *feature extractor*) or to adapt the weights to learning task at hand (i.e. *fine-tuning*). We use the widely used pre-trained network `VGG16` proposed by Simonyan and Zisserman (2014) and available in `Keras` as `keras.applications.VGG16` as our main convolutional base. We use grid search to select whether to use the `VGG16` as a feature extractor or to fine-tune the weights. See Section 2.5 for details. We remove the existing fully connected layers of the `VGG16` network and instead adapt it to our application. We flatten the output of the pre-trained network and then applies a dropout layer and a dense layer with 512 weights with a sigmoid activated output layer with a single node to finalize the network. We demonstrate our application in a pre-registered `.py`-script on GitHub.

Overfitting is likely to a larger problem in applications with a small training data set. To combat this, we use three techniques. First, and as already mentioned, we use a dropout layer randomly deleting weights from the flattened layer[1]. Second, we employ image augmentation randomly changing the representation of the images. This serves dual purposes at the same time both reducing the likelihood of overfitting; we make the network less sensitive to specific representations of images, and we effectively increases the number of training samples. Third, we use early stopping to terminate the network after the validation accuracy does not improve after t patience steps. This means that our network returns the weights from

[1] See Section 2.5 for details about how to set the dropout rate.

the epoch with the highest validation accuracy instead of the weights after the final epoch.

We train separate models for male and female candidates as we expect that the features relating faces of candidates to electoral success differ between the two genders. We use the same overall network for both male and female candidates except for tuning of the following hyperparameters: dropout rate, transfer learning (feature extractor vs. fine-tuner), learning rate, and batch size. See Section 2.5 for details.

2.3. Model Selection

As already mentioned, we expect that we face a difficult learning task due to the combination of a small training set and a weak related target function f. We find clear evidence of this expectation in our initial model training as we find substantial variation in the performance of the same network with the same set of hyperparameters. Although we use a different training/validation-split each time (such that the training data varies) this can not alone cause the instability.

To combat this, we use a combination of grid search and "soft selection" of a model. Instead of training only a single model with the best set of hyperparameters (i.e. "hard selection"), we train our best model, as returned from the grid search, 20 times and select the fit on the 75th percentile. Note that this distribution is over the balanced training accuracy as we use all our images in the 2017 data to train our final model. We call this "soft selection" as it is based on a distribution rather than a single fit. We risk that this approach introduces overfitting, but on the other hand, we discovered substantial underfitting in our initial training as we ended up with models that predicted the majority class for every image.

In addition to our "soft selection" approach, we report two additional approaches to model selection:

1. "Hard selection": Train a model a single time and apply it on the 2021 data

2. PAC-Bayesian: We define a distribution over the 20 models and compute the optimal posterior and use the posterior to weight the predictions of an ensemble (i.e. majority vote) of the 20 models. Thus, we assign higher weights to models that returned higher balanced validation accuracy. See our supplementary information to our pre-registration for further information.

2.4. Sampling plan

2.4.1 Getting Images from the 2021 election

We plan to scrape pictures of candidates running in the Danish local elections of 2021 from so called *candidate tests*. Multiple Danish news media provide online candidate tests where the running candidates are invited to answer a series of policy questions. In doing so, the candidates are encouraged to upload an image of themselves so that the voters can recognize them. These self-uploaded images constitute our base sample. The voters can then take the same test and answer the same questions as the candidates to see with whom the voter agrees and disagrees the most.

Since it is voluntary for the candidates to answer the candidate tests, there is inevitably some selection bias. *Ceteris paribus*, we expect candidates with a better chance of election to be more likely to answer the candidate test as they have more at stake and more resources to complete the test. In our 2017 pilot data, candidates that answered the test received an average of 359 personal votes compared with 119 votes for candidates that did not answer the test. We expect the same difference to be present in the 2021 elections.

We scrape the pictures of the candidates who have have completed the test and uploaded an image hosted by DR shortly after the election on November 16 2021. This data collection has been approved by the ethics board of the Aarhus University.

2.4.2 Image Preprocessing

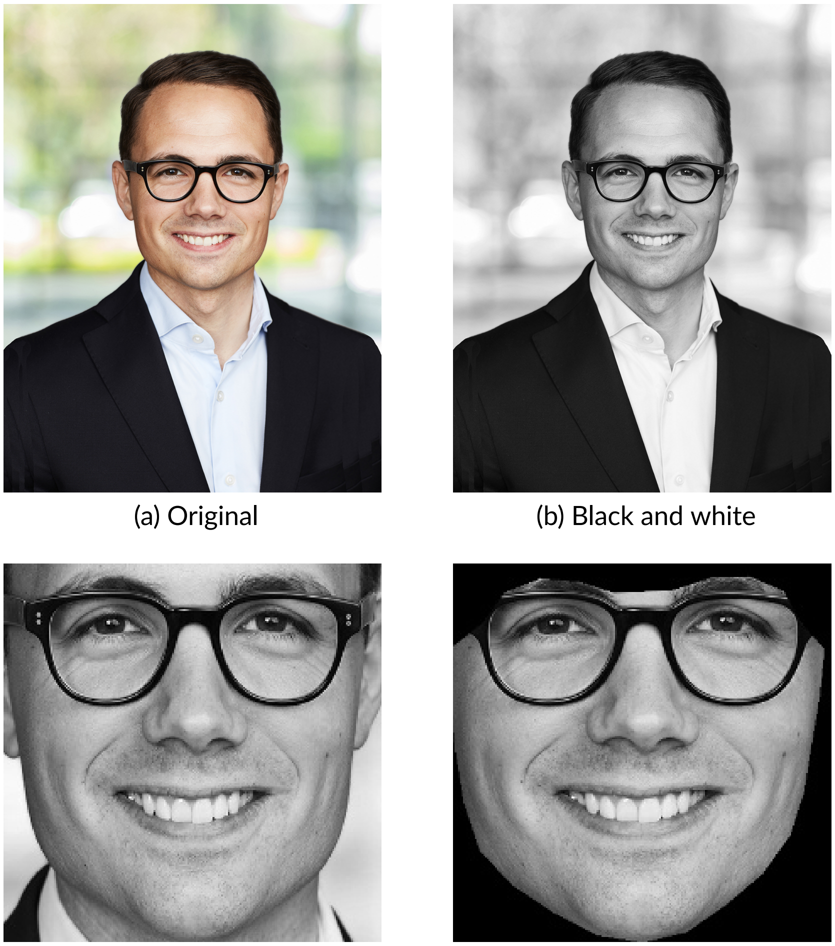
As for our 2017 pilot data, we manually review the pictures we scrape from the 2021 candidate test to evaluate whether they are of sufficient quality to analyze using computer vision techniques. To ensure that our images only contain the faces of the candidates and do not pick up on other features, we apply a set of preprocessing steps. First, we convert each image to black and white to eliminate color effects. Second, we crop the faces from each image using pre-trained landmark detection from `Dlib` using the `Python` API. Third, we remove the background of the cropped facial image to get rid of any background effects not already removed by the cropping. After these steps, we computed an estimate of the blurriness of the image and dropped images below a certain threshold (100). We provide an example of our preprocessing steps in Figure 1.

2.4.3 Class Imbalance and Oversampling

It is well-known that class imbalance can hamper generalization in machine and deep learning (Buda et al. 2018). In our pilot data (the 2017 election), our target $\mathcal{Y}$ is imbalanced with prior class probabilities 34.2% and 38.7% for male and female candidates respectively. To circumvent potential issues, we apply oversampling to reduce the imbalance. We use a oversampling ratio of 0.75 for both men and women meaning that we oversample the minority class $\{1 : \texttt{elected}\}$ until it contains 75% of the samples of the majority class $\{0 : \texttt{unelected}\}$. We demonstrate this in a pre-registered `.py`-script on GitHub.

2.4.4 Image Augmentation

Since we have a fairly small number of samples at our disposal, we employ image augmentation, i.e. randomly resizing, rotating, and adding noise to the existing images. This has been shown to be very effective in stabilizing model generalization in computer vision tasks as the model becomes less sensitive towards specific pixel



(a) Original

(b) Black and white

(c) Cropped

(d) Background removed

Figure 1 — Example of image preprocessing: Simon Aggesen

values if only using the original image. Furthermore, it effectively increases the number of samples, which can used to train the CNN. The augmentation is only used for the training samples. We use the following augmentations:

Augmentation	Value
Rotation range	1
Width shift range	0.1
Height shift range	0.1
Shear range	0.1
Zoom range	0.1
Fill mode	nearest[2]

The values and augmentation options are implemented using `ImageDataGenerator` from `keras.preprocessing.image` in `Python`. We demonstrate our use of generators in a pre-registered `.py`-script on GitHub.

2.4.5 Test of hypotheses

We don't use traditional hypothesis testing in the paper. In some occasions, we test the statistical difference between two numbers. To do so, we use the approach put forward in `scripts/validation/mean_differences.py`.

2.5. Analysis Plan

In this section, we outline our main workflow used in our analysis.

2.5.1 Main workflow

We use the same workflow for male and female candidates:

1. Pre-processing of images to use as input for the CNNs. We follow the pre-processing steps layed out in Section 2.4.

2. To select our specification of our `VGG16` CNN, we used grid search over the following specifications and hyperparameters, iterating the procedure five times:

Hyperparameter	Search space
`transfer learning`	$\in$ {`feature extractor`, `fine-tune`}
`learning rate`	$\in \{1e{-}04^{*}, 1e{-}05, 1e{-}06\}$
`batch size`	$\in \{16, 32, 64\}$
`dropout`	$\in \{0.4, 0.5, 0.6\}$

* Only for female candidates

giving a total of 54 hypothesis models for women and 36 for men. For each model in each iteration, we used a stratified $0.7/0.3$-split. We used an iterative approach to maximize the variation in performance during the training and validation phase to get a sense of the distribution for each model defined over the search space. We stratified the training-validation-split according to the distribution of our target (electoral success) to ensure that we got the exact same number of training samples after oversampling the minority class to obtain a more balanced target.

3. After conducting the grid search, we ended up using the following set of hyperparameters for men and women:

Hyperparameter	Male	Female
`transfer learning`	`fine-tune`	`fine-tune`
`learning rate`	$1e{-}06$	$1e{-}05$
`batch size`	16	16
`dropout`	0.5	0.5

The results of the grid search is available on our pre-registered GitHub.

4. After selecting the best set of hyperparameters for each of the two models, we then used our iterative "soft selection" approach to select our final model as outlined in Section 2.3. We train the models as specified in the table above 20 times and then save the model at the 50th percentile. This amounts to the 10th best model and is the median model.

5. When we have the results and images from the 2021 election, we then use the pre-registered models to classify whether a candidate is elected or not and analyzes our hypotheses using the approaches described in the next section.

2.5.2 Test of hypotheses

Although we use oversampling, we still have a clear minority and majority class. Therefore, we use the the balanced accuracy as our main performance metrics throughout the paper. For each of the hypotheses, we explore whether there are ideological differences in our ability to predict the electoral success. Thus, we analyze whether our classification accuracy is different for left and right-wing candidates. We do not have any clear expectations about whether the accuracy is higher for the left or the right-wing.

- Hypothesis 1 (`H1`)
  - `H1`: We can predict the electoral success of candidates above random chance (50%).
  - *How?*: If the balanced test accuracy is above 50%, it supports our hypothesis.
- Hypothesis 2 (`H2`)
  - `H2`: Electoral success is correlated with facial expressions of emotions (facial expressions)
  - *How?*: We compare the probabilities with measures of facial expressions measured with Azure's API. We also measure facial expressions with the `Python` library `Py-Feat` (Cheong et al. 2021). We report the corresponding correlation coefficients and their p-value.
- Hypothesis 3 (`H3`)
  - `H3`: There are a gender bias such that we are better at predicting the electoral success for female than for male candidates.

  - *How?*: We compare the balanced test accuracy of our two CNNs models and investigates whether they are statistically different from each other using the approach outlined in the script `scripts/validation/mean_differences.py`.

- Hypothesis 4a (`H4a`)
  - `H4a`: Electoral success is correlated with the beauty for female candidates.
  - *How?*: We use transfer learning to label the beauty of female candidates in our sample using a pre-trained neural network[3] trained to predict the beauty of faces. We then use the beauty scores and test whether is mean differences across elected and unelected female candidates. Further, we also investigate whether the probabilities correlate with the beauty scores using Spearman correlation.

- Hypothesis 4b (`H4b`)
  - `H4b`: Electoral success is correlated with masculinity for male candidates.
  - *How?*: We use the so-called facial width heigh ratio (fWHR[4]) to compute masculinity scores for male candidates. We then use the masculinity scores and test whether is mean differences across elected and unelected male candidates. Further, we also investigate whether the probabilities correlate with the masculinity scores using Spearman correlation.

- Hypothesis 5 (`H5`)
  - `H5`: We are better at predicting the electoral success in open-list than in closed-list municipalities.
  - *How?*: We make subgroup analyses for both men and women and test whether the balanced test accuracy differs between open and closed-list municipali-

[3] `https://github.com/lucasxlu/TransFBP`

[4] `https://github.com/TiesdeKok/fWHR_calculator`

ties. We compute the statistical difference using the approach set forward in `scripts/validation/mean_differences.py`.

- Hypothesis 6a (`H6a`)
  - `H6a`: We expect to reproduce predictions based in hypotheses 1 to 5 using the vote shares of candidates.
  - *How?*: We use the same general approach as for `H1`. We make a categorical specification of vote shares and change the output layer of our CNNs to adapt the network to a multi-class classification problem. Consequently, we apply a softmax function rather than a sigmoid function in the output layer.
- Hypothesis 6b (`Hb`)
  - `H6b`: We expect to reproduce predictions based in hypotheses 1 to 5 when comparing candidates re-running for office compared to candidates running for office for the first time.
  - *How?*: We divide the 2021 candidates into two groups: the re-runners and the new-runners. We then compute the statistical difference using the procedure in `scripts/validation/mean_differences.py`

2.5.3 Validation

To validate our CNN approach, we use two main approaches:

1. **Grad-cam analysis**: We extract the gradients of the last layer of the CNN (before the classiciation layer) and computes a weighted total of the features most importance for classification (Selvaraju et al. 2017). We then superimpose the corresponding weights on the image to generate a heat map showing the activation of features. We demonstrate how we do so in the `scripts/validation/grad-cam.py`.

2. **Face averages**: Following Wang and Kosinski (2018), we use face averages to investigate whether there are any visual differences between elected and unelected candidates. We demonstrate how we do so in the `scripts/validation/face-averages.`

2.5.4 Supplementary Analysis

In addition to the main analyses where we test our hypotheses, we also report several supplementary analyses. These are most likely to figure in the Supplementary Information in the final manuscript.

- **Methods**:
  - In addition to our main model selection (i.e. selecting the 50th percentile across 20 iterations), we also report the results using a PAC-Bayesian Aggregation approach where we weight the predictions by ρ found by minimizing a PAC-bayes bound in an ensemble of all 20 models. Thus, we take a majority vote using all 20 models where their predictions are weighted by the posterior ρ. See the Supplementary Information for the pre-registered report for details.
  - We also report the results using the single best model of the 20 iterations.
  - We investigate whether our results differ if using two different CNN architectures: `MobileNetV2` and `InceptionResNetV2`.
  - We investigate the impact of class imbalance on the results. In our main approach, we oversample the minority class ($\{1 : \text{elected}\}$) to reduce class imbalance. We investigate how this matters for our results by removing the oversampling and by investigating the degree of oversampling.
- **Theory**:
  - We investigate whether the results differ if we only restrict the analysis to the “national” parties. Our main analysis includes all sorts of parties.

- We investigate whether a candidate’s position on the party list matters.

References

Achen, Christopher et al. (2017). *Democracy for realists*. Princeton University Press.

Berggren, Niclas, Henrik Jordahl, and Panu Poutvaara (2010). “The looks of a winner: Beauty and electoral success”. In: *Journal of public economics* 94.1-2, pp. 8–15.

Bisgaard, Martin (2019). “How getting the facts right can fuel partisan-motivated reasoning”. In: *American Journal of Political Science* 63.4, pp. 824–839.

Boussalis, Constantine and Travis G Coan (2021). “Facing the electorate: Computational approaches to the study of nonverbal communication and voter impression formation”. In: *Political Communication* 38.1-2, pp. 75–97.

Buda, Mateusz, Atsuto Maki, and Maciej A Mazurowski (2018). “A systematic study of the class imbalance problem in convolutional neural networks”. In: *Neural Networks* 106, pp. 249–259.

Caprara, Gian Vittorio and Philip G Zimbardo (2004). “Personalizing politics: a congruency model of political preference.” In: *American psychologist* 59.7, p. 581.

Carpinella, Colleen M et al. (2016). “The gendered face of partisan politics: Consequences of facial sex typicality for vote choice”. In: *Political Communication* 33.1, pp. 21–38.

Carpini, Michael X Delli and Scott Keeter (1996). *What Americans know about politics and why it matters*. Yale University Press.

Cheong, Jin Hyun et al. (2021). “Py-Feat: Python Facial Expression Analysis Toolbox”. In: *arXiv preprint arXiv:2104.03509*.

Everitt, Joanna, Lisa A Best, and Derek Gaudet (2016). “Candidate gender, behavioral style, and willingness to vote: Support for female candidates depends on conformity to gender norms”. In: *American Behavioral Scientist* 60.14, pp. 1737–1755.

Gaines, Brian J et al. (2007). “Same facts, different interpretations: Partisan motivation and opinion on Iraq”. In: *The Journal of Politics* 69.4, pp. 957–974.

Goodfellow, Ian et al. (2016). *Deep learning*. Vol. 1. MIT press Cambridge.

Goren, Paul (1997). “Political expertise and issue voting in presidential elections”. In: *Political Research Quarterly* 50.2, pp. 387–412.

Hayes, Danny (2009). “Has television personalized voting behavior?” In: *Political Behavior* 31.2, pp. 231–260.

Kahan, Dan M (2012). “Ideology, motivated reasoning, and cognitive reflection: An experimental study”. In: *Judgment and Decision making* 8, pp. 407–24.

Kosinski, Michal (2021). “Facial recognition technology can expose political orientation from naturalistic facial images”. In: *Scientific reports* 11.1, pp. 1–7.

Lau, Richard R and David P Redlawsk (2006). *How voters decide: Information processing in election campaigns*. Cambridge University Press.

Lawson, Chappell et al. (2010). “Looking like a winner: Candidate appearance and electoral success in new democracies”. In: *World Politics* 62.4, pp. 561–593.

Lodge, Milton and Charles S Taber (2013). *The rationalizing voter*. Cambridge University Press.

McAllister, Ian et al. (2007). “The personalization of politics”. In: *The Oxford handbook of political behavior*. Oxford University Press.

Neuman, W Russell and WR Neuman (1986). *The paradox of mass politics: Knowledge and opinion in the American electorate*. Harvard University Press.

Pedersen, Helene Helboe and Gideon Rahat (2021). *Political personalization and personalized politics within and beyond the behavioural arena*.

Selvaraju, Ramprasaath R et al. (2017). “Grad-cam: Visual explanations from deep networks via gradient-based localization”. In: *Proceedings of the IEEE international conference on computer vision*, pp. 618–626.

Simonyan, Karen and Andrew Zisserman (2014). “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556*.

Stewart, Patrick A, Frank K Salter, and Marc Mehu (2009). “Taking leaders at face value: Ethology and the analysis of televised leader displays”. In: *Politics and the Life Sciences* 28.1, pp. 48–74.

Tilley, James and Sara B Hobolt (2011). “Is the government to blame? An experimental test of how partisanship shapes perceptions of performance and responsibility”. In: *The journal of politics* 73.2, pp. 316–330.

Todorov, Alexander (2017). *Face value: The irresistible influence of first impressions*. Princeton University Press.

Todorov, Alexander et al. (2005). “Inferences of competence from faces predict election outcomes”. In: *Science* 308.5728, pp. 1623–1626.

Todorov, Alexander et al. (2015). “Social attributions from faces: Determinants, consequences, accuracy, and functional significance”. In: *Annual review of psychology* 66, pp. 519–545.

Wang, Yilun and Michal Kosinski (2018). “Deep neural networks are more accurate than humans at detecting sexual orientation from facial images.” In: *Journal of personality and social psychology* 114.2, p. 246.

Willis, Janine and Alexander Todorov (2006). “First impressions: Making up your mind after a 100-ms exposure to a face”. In: *Psychological science* 17.7, pp. 592–598.

Yang, Qiang et al. (2020). *Transfer learning*. Cambridge University Press.