Measuring Rhetorical Similarity with Supervised Machine Learning *

Nicolai Berk

November 10, 2020

Abstract

Recent advances in the application of supervised learning have shown how the method can be employed to measure polarization by assessing classifiers' accuracy. Building on these contributions, I propose a reconceptualisation to enable extended utilization of this approach. Focusing on predicted probabilites as a measure of rhetorical similarity, I validate supervised learning for the measurement of accommodation to radical right parties through established parties and politicians in Austria, Germany, and the Netherlands. Results indicate that the method produces valid estimates of parties' and speakers' rhetorical similarities to respective radical-right parties and outperforms existing similarity measures and scaling methods. I discuss possible further applications and limitations.

Keywords: measurement, rhetorical similarity, parliamentary debate, text analysis, supervised learning, machine learning, party competition, radical-right, political rhetoric.

1 Introduction

Political scientists are often interested in the rhetorical similarity of different sets of texts. For example, scholars of representation might be interested whether certain groups communicate distinctively in parliament (Pitkin 1967) or whether legislation is similar to demands voiced by interest groups (Gilens and Page 2014). Students of government formation and termination want to know which parties communicate most similar and are hence most likely to govern with each other (Gamson 1961), and when these coalitions are likely to break down (Grofman and Van Roozendaal 1994). Public opinion scholars evaluate how the media or political groups left their mark on the current public discourse (Zaller 1992). Scientists interested in party competition might assess historical similarities between parties to judge whether one party filled a place abandoned by another (Kitschelt 1986), or whether one party moved towards another (Downs

*Replication materials are available via https://github.com/nicolaiberk/SMLSE

1957). Knowing how similar MPs communicate allows to judge whether certain MPs are likely to leave their party (Hirschman 1970).

Despite the centrality of rhetorical similarities to political science research, existing approaches to measure it are limited. Scaling methods such as wordscore (Laver, Benoit, and Garry 2003) and wordfish (Slapin and Proksch 2008) place labels[1] in a one- (wordfish) or multi-dimensional space (wordscore), based on their word usage. These measures were designed for measuring ideological positions, not similarities, as they cannot precisely estimate the similarity of a given document *with* a corpus. As such, they measure a latent dimension to be interpreted by the researcher, not a stable, pre-defined construct (Goet 2019). Simpler similarity measures, such as cosine similarity, are designed to assess literal identity of single documents, however perform badly for meaningful assessments (Prasetya, Wibawa, and Hirashima 2018). To my knowledge, no specified method to estimate the rhetorical similarity of groups exists at the time of writing.

To address this methodological gap, I propose the use supervised learning to measure the similarity of a given document with a specified corpus. This approach builds on an existing method to assess the similarity between groups using classifier accuracy (Peterson and Spirling 2018), but moves moves the focus to predicted probabilities as a substantive similarity measure. This enables to provide a precise similarity estimate for each document, based on its word use. In the next section, I briefly discuss the relevance of rhetorical similarity and the capability of existing methods to capture it. Afterwards, I explain how supervised learning can be used to estimate rhetorical similarity step by step. Using parliamentary speeches from Austria, Germany, and the Netherlands, I then provide evidence indicating that the estimates are able to distinguish radical-right parties in line with theoretical expectations and produce valid estimates of similarity to the radical-right for established parties and speakers. The last section compares these estimates from the supervised method with wordfish and cosine similarity estimates and shows that the method returns more meaningful and precise results.

2 Estimating rhetorical similarity

The logic of my approach builds on the idea that language is indicative of the way humans perceive and understand their surrounding world. In this view, rhetorical similarities, i.e. the shared dictionaries and figures of speech to describe the world, are symptoms of similarities in mental representations and cognitive processes (Pennebaker, Mehl, and Niederhoffer 2003; Lakoff and Johnson 1980). This extends to the political realm: people who talk about political topics in similar ways should have similar perceptions of which issues are at stake and share common mental representations of these issues.

[1] I define 'labels' as meta-information about documents (such as authorship), which allows to group said documents into corpora. A corpus (plural: corpora) is a group of documents.

A measure of such rhetorical similarity should have several properties to be considered valid and useful. First, such an indicator should reflect a given document's similarity to other documents. The resulting estimate should then be informative about which corpus a document is likely to belong to. As textual data sets are often very large, the method should be scalable and fast to apply. Ideally, it's application would not presuppose an understanding of the language through the researcher, but allow for 'language-blind' application in different contexts. However, *if* the researcher is able to understand the language, meaningful information about what drives similarities and distinctiveness of documents is desirable.

2.1 Existing similarity measures

Several existing methods could be used to measure rhetorical similarity. One of the most popular methods - cosine similarity - places documents in a multidimensional space. Every word constitutes a dimension ('feature') and the number of times that word was used defines the placement of a document on that dimension. Each document is then represented by a vector in high-dimensional (word) space. The cosine of the angle between those two vectors indicates how similar the word use in two given documents is, independent of the documents' length[2] (Egesdal, Gill, and Rotemberg n.d.). Similarly simple, Jaccard similarity compares the words used in two given documents and returns the share of common words (independent of how often they are used) relative to all words used in the documents (Jaccard 1912).

Methods like cosine and Jaccard similarity are designed for literal comparison of single documents. This kind of 'lexical' similarity is interesting in several domains of social sciences, for example whether the release of specific documents affected political speech (Egesdal, Gill, and Rotemberg n.d.; Hager and Hilbig 2020). However, political scientists are usually interested in the similarity of groups, such as political parties or identities such as gender. Additionally, each feature (word) has the same weight - the differing importance of words cannot be modelled. As a result, these measures are well designed to measure literal similarity, but do not perform well when similarity in meaning is concerned (Prasetya, Wibawa, and Hirashima 2018).

Rhetorical similarity could also be assessed using scaling methods. Well proliferated in political science for the measurement of the ideological placement of texts, these methods estimate underlying dimension(s) explaining differences in word use and scale the documents on this dimension. This happens either using labelled reference texts (wordscore; Laver, Benoit, and Garry 2003), or using certain assumptions about the distribution of words (wordfish; Slapin and Proksch 2008). These methods even allow researchers to assess the placement of individual terms, enabling to detect more meaningful differences between corpora inductively.

[2] That is, two documents containing words a and b are considered identical, as long as they contain them in equal shares.

Wordscores are highly dependent on the reference documents selected and labelled by the researcher (Lowe 2008). This not only reduces the amount of processed information to the reference texts, but possibly introduces subjective bias. While careful pre-processing can address some of these problems, this adds to the already laborious tasks of selection and labelling.

Less tedious is the application of an unsupervised method for scaling: Wordfish extracts the primary underlying dimension explaining word use by estimating it as a function of the length of a document, the frequency of a word in all documents, and estimated weights for words and groups, such as parties. The method then re-estimates this model - starting with arbitrary values - until the model fit is maximised (Slapin and Proksch 2008). It also fits the model on the entire data, which substantially increases the variation taken into account.

Nevertheless, it also builds on the assumption that the differences in word use between the documents are indicative of the main (ideological) differences between labels (e.g. parties). However, the factors extracted are not necessarily indicative of those differences, which might sometimes result in failure to distinguish labels of interest (Grimmer and Stewart 2013, 292f). Most importantly, both methods are designed to extract the primary dimension structuring differences in word use, not similarity to a corpus. This means that differences in scaling are indicative of placements on an often hard to interpret underlying dimension, not a clearly defined and easy-to-interpret construct.

All discussed approaches reflect similarities in word use, are highly scalable, and (with the exception of wordscores) allow an application without knowledge of the particular language texts are written in. However, these methods have major shortcomings for the assessment of the similarity of documents to corpora, either because they do not account for meaningful differences or are not designed to estimate similarities. An indicator that mitigates these problems should focus on the differences between labels. Moreover, without relying on subjective judgement or hard-to-interpret factors, it should estimate precise similarity scores while maintaining simplicity of application, scalability, and language-blindness. Such an indicator is presented in the next section.

2.2 Using supervised machine learning to measure rhetorical similarity

Instead of estimating similarities of single documents or underlying factors, differences in word use can also be exploited through machine learning to predict the likelihood that a document belongs to a certain corpus. Classic machine learning applications such as authorship studies use this approach to estimate the likelihood that an unlabelled document belongs to a given label (e.g. is authored by a specific person; Mosteller and Wallace 1963). Peterson and Spirling (2018) invert this logic: instead of using labelled documents to extrapolate to unlabelled documents, they measure how similar two groups producing texts are, using only labelled data. That is, for each observed period (in their case, the legislative period), they obtain a classifier accuracy, which is informative of how well the groups of documents (e.g. speeches from two parties) can be

distinguished, based on their word use. From that, they infer how similar the groups are to assess how polarised the British parliament was in a given legislature (for similar applications, see Gentzkow, Shapiro, and Taddy 2019; Goet 2019).

I extend this more descriptive application of supervised machine learning to generate an indicator of similarity to a given corpus for each document. Rather than assessing classifier accuracy and obtaining one data point per period, I focus on predicted probabilities as a substantive quantity of interest. These allow the precise estimation of the similarity of each single document to a given corpus (e.g. speeches by a specific party). The method is language-blind (as it only needs to know the frequencies of words) and highly scalable (ten thousand or one million speeches need the same amount of code and human labour). If researchers speak the language, they can assess the best predictor words to learn about the quality of the rhetorical differences. Most importantly, it measures an easy-to-interpret dimension, defined by the researcher.

Summarised, the method I propose works as follows: a classifier is trained on the full data[3] to estimate whether a text carries a certain label (e.g. is authored by a certain party). This results in a model where each word is assigned a correlation coefficient, indicating it's association this label. Based on this model, the classifier 'predicts' the likelihood that a text belongs to label of interest. This predicted probability is estimated *based on the similarity of word use* to the texts with this specific label. Thus, the likelihood that a text belongs to the category of interest is itself a similarity measure. Even though the researcher knows whether a document was authored by a member of the radical-right party, the specific word use can still resemble that of a 'typical' radical-right speech more or less. The predicted probabilities of the bag-of-words model reflect this variation.

The list below outlines the application procedure. It should help readers understand and evaluate the estimates, but also guide researchers in their efforts to employ this indicator for their own purposes.

1. **Select a Classifier**
   Although the similarities are only interpretable relative to other estimates, it is desirable to use a classifier that is able to distinguish well between the labels in order to extract meaningful information. To that end, one should compare the performance of a number of pre-processing steps and classifiers before deciding which to use. Following Peterson and Spirling, I do not remove stopwords from a pre-defined list but very frequent as well as rare terms to avoid overfitting 2018. This is more context-specific and should thus also be more neutral in its impact (see Schoonvelde, Schumacher, and Bakker 2019 for a discussion of this issue). Beyond that,

[3] I diverge from the applications by Peterson and Spirling 2018 and Goet 2019 in that I train the classifiers on the full data, reflecting the descriptiveness of the application (rather than inferring to unlabelled texts) and taking into account the entire available data. The section on comparative performance compares estimates from the full data to those trained on a subset of the data

the researcher should compare the performance of classifiers when using stemmed and unstemmed text, raw term counts and counts weighted by overall document frequency (tf-idf), as well as different classification algorithms (Denny and Spirling 2018).

Training classifiers several times on subsets of the data and assessing average performance[4] (cross-validation) minimises the risk of overfitting in the selection process (Breiman and Spector 1989). This is especially relevant as textual data suffers from finite-sample bias, meaning the speakers can only choose so many words from a huge dictionary of possible words. This bias might result in the over-estimation of partisan differences simply due to chance[5] (Gentzkow, Shapiro, and Taddy 2019).

2. **Balance the data**
   Classifiers perform sub-optimal on imbalanced data. This could result in better performance for larger parties and worse performance for smaller parties, rendering the latter more similar than they actually are. To avoid this, I use a synthetic minority over-sampling technique (SMOTE) to train the model on a balanced set of speeches from each party. SMOTE uses a 'nearest-neighbor' approach to generate additional, synthetic cases to balance the data without information loss (Chawla et al. 2002). The final training set hence contains an identical number of speeches from each party.

3. (if necessary) **Divide the data into time-periods**
   Sometimes, researchers might want to assess similarities across time. However, the use and meaning of words change across time. To address this issue, the data can be divided into subsets across time (e.g. a legislative session or presidential term). Training one classifier per time-period controls for changing language.

4. **Fit the classifier(s) and estimate predicted probabilities**
   The selected classifier is fitted to the over-sampled data (think of a regression where each word is an independent variable predicting the label, e.g. party membership of the speaker), returning coefficients[6] for each word based on the correlation with the outcome category of interest (e.g.

[4] Note that I deviate from Peterson and Spirling by training one classifier per country instead of per session to maintain maximal comparability between the estimates (2018). This is also in line with the idea that pre-processing steps such as stemming should be either relevant or not relevant, depending on the language of the speeches, but not the particular period assessed.

[5] Three steps are taken to reduce this problem: first, a bag-of-words approach is used (whereas Gentzkow, Shapiro, and Taddy used bi-grams), severely de-creasing the feature space. Second, the five-fold cross-validation makes sure a classifier that performs well on different subsets of data is selected, thus reducing the chance to overfit. Third, very rare terms (being used in less than five speeches within a period) are excluded, thus again reducing the likelihood that certain rare words discriminate between the parties just by chance.

[6] Note that many classifiers do not return coefficients nor use regression. This example was chosen for its perspicuity.

radical-right authorship). Using this model, predicted probabilities are estimated for each speech. The higher the likelihood estimated, the more similar the word use in the text to the word use in the outcome category. Hence the estimated likelihood that a text belongs to the corpus is inherently a similarity score. For those speeches in the corpus of interest, the estimates can be interpreted as a distinctiveness-score. The higher the probability for a text to be part of this group, based on its word use, the more distinctive the text must be from the word use of other groups.

3 Empirical data

To demonstrate the application and validity of SMLSE, I employ the ParlSpeech dataset of parliamentary speeches (Rauh and Schwalbach 2020) to study the rhetorical similarities of parties and speakers with the radical-right. The accommodation of radical-right parties by mainstream parties has inspired a vast literature and constitutes a major area of research on party competition (see e.g. Arzheimer 2009; Bale et al. 2010; Dahlström and Sundell 2012; Harmel and Svåsand 1997; Krause, Cohen, and Abou-Chadi n.d.; Meguid 2005; Schumacher and Kersbergen 2014; Spoon and Klüver 2020; Van der Brug, Fennema, and Tillie 2005; Van Spanje 2010; Wagner and Meyer 2017). I selected the cases of the Austrian *Nationalrat*, Germany's *Bundestag*, and the *Tweede Kamer* of the Netherlands. The Dutch and Austrian case allow me to make a historical assessment of the similarity of mainstream parties and speakers to the radical-right, as their party systems include radical-right parties for at least the past twenty years. Germany, on the other hand, has only recently witnessed the rise of the radical-right, but represents a much-studied case (see Arzheimer and Berning 2019: 2-5 for a review).

Using python's scikit-learn-package (Pedregosa et al. 2011), I develop a classifier to estimate whether a text is authored by a radical-right party (Germany: AfD; Austria: FPÖ, BZÖ; Netherlands: LPF, PVV, FvD). I use the German data for classifier selection. After removing very infrequent and very frequent terms as well as punctuation from the texts, I train three different models (Logistic Regression, Multinomial Naïve Bayes, Support Vector Machine) on stemmed and un-stemmed text, quantified as either raw document counts, or weighted by the terms' inverse frequency across documents. This results in 3x2x2 = 12 different classifiers. I use five-fold cross-validation to choose a classifier that performs well. The logistic regression on inverse-document-frequency-weighted un-stemmed texts produces the most desirable results (accuracy: 0.89, recall: 0.59, precision: 0.62) and is hence chosen as the classifier of interest for all applications (see table 2 in appendix A for a comparison of the classifier performances). Bi- and tri-grams did not substantially affect these estimates nor improve interpretability of the best predictors and were hence excluded of the analysis. The training data was over-sampled using the SMOTE-algorithm from the `imblearn`-package (Lemaitre, Nogueira, and Aridas 2017). For the classification of Austrian and Dutch speeches, I train one classifier per legislative

session, thus controlling for changing language across time. The German parliament saw the re-emergence of the radical-right only in 2017, therefore only one classifier is trained for this data (which ends in December 2018).

4 Validity

The measure displays criterion-related validity by identifying the outcome category (radical-right authorship) with 89% accuracy. I will address two other types of validity in the next section. First, I assess the best predictor words of the classifiers in the most recent parliamentary session in Austria, Germany, and the Netherlands. I highlight the content validity of the measure by showing which features the classifier discriminates on. Construct validity, i.e. the "extent to which a measure performs according to theoretical expectations" (Carmines and Woods 2004), is shown presenting four cases where the method provides valid estimates of parties' and speakers' similarity to the radical-right.

4.1 Content validity: best predictor words

Radical-right parties are usually defined based on their distinctive policy positions. These parties are nationalist and hold especially conservative cultural and authoritarian positions. Additionally, they often communicate with populist rhetoric, including anti-elitism and people-centrism (Mudde 2007). Table 1 shows the words most positively (red) and negatively (blue) correlated with radical-right authorship. These words are the best predictors to distinguish the radical-right parties (FPÖ; AfD; PVV/FvD) in the most recent legislative period in the data (ending in Dec 2018 in DE & AT, Jul 2019 in NL), out of all words in the model.

Considerable differences exist in the rhetoric distinguishing radical-right parties in these three countries. The AfD in Germany is foremost distinguished by its aversion to gender-inclusive language, which the party criticises as 'gender-madness'[7]. It also displays heightened reference to itself, Germany (in line with its nationalist ideology), and the government (being the largest opposition party), compared to other parties. Its members address the parliament with 'ladies and gentlemen', not the male and female forms of 'colleagues'. Using causal adverbs ('therefore', 'hence') and referring to 'democrats' is also less found in AfD-authored speeches. These findings also caution against a one-size-fits-all approach to pre-processing, as the rejection of gendered nouns would have remained hidden had I stemmed the words.

As the Austrian FPÖ was in government at the time, it makes sense that it is distinguished from other parties by reference to the opposition (the social democrats of the SPÖ are the main opposition party) rather than the government, including the reference to Vienna, a stronghold of the Social Democrats much criticised by the radical and centre-right. It also prefers to address the parliament as 'ladies' and 'gentlemen' rather than 'colleagues'. One of the major

[7] See `https://afdkompakt.de/tag/genderwahn/`

Germany	Austria	The Netherlands
citizen/s [m]	SPÖ	immigration
Merkel	[Intercept]	and so on
german	honoured	Islam
AfD	gentlemen	PVV
Germany	ladies	Islamic
old-parties	Vienna	Brussels
here	to	possible
government	opposition	immigration policy
much	once	illegal immigrants
thanks	sector	hate-speech paragraph
ladies	patients [m]	sign
thank you	take	animal police
gentlemen	just/exactly	house-keeping
employees	colleague [f]	gigantic
soldiers [f]	patients [f]	discount
state secretary	Stöger (SPÖ)	bene
warmly	must	status holders
important	Kern (SPÖ)	heaven's sake
coalition agreement	yes	Kops (PVV)
last	believe	consideration
therefore	minister [f]	At the same time
the left	good	agreements
colleagues	colleagues [m]	diligent
hence	many	advised against
need	commission	mister
citizens [f]	motion	look
democrats	minister [f]	Agema (PVV)
say	humans	bright
think	colleagues [f]	predominant
colleagues [f]	dear	Baudet (FvD)

Table 1: Thirty best predictor words increasing (red) and decreasing (turquoise) likelihood of radical-right authorship for Germany, Austria, and the Netherlands. Squared brackets indicate gender, round brackets note the party affiliation for specific names.

projects of the FPÖ-led health ministry was a reform of the health sector, which explains the reference to 'patients'. Referring to members of the government in female form decreases the likelihood of radical-right authorship (similar to the German AfD), as does reference to 'humans' and addressing the house 'warmly'.

Lastly, and differing from their German and Austrian equivalents, the Dutch FvD and PVV are distinguished primarily by reference to immigration and Islam. These parties' rhetoric refers to 'immigration', 'Islam'/'islamic' and 'ille-

gal immigrants' more than other parties, in line with their distinctively anti-immigration positions. Reference to 'Brussels' underlines the Eurosceptic position of these parties (Rooduijn et al. 2019). Words like 'house-keeping', and 'discount' could refer to the Budget, where at least the FvD proposes tax cuts[8]. Additionally, FvD and PVV make use of more informal ('and so on', 'heaven's sake') and less nuancing language ('consideration', 'diligent', 'advised against', 'predominant'). As nouns are usually not gendered in Dutch, it is not surprising that gendered language is no predictor here.

The best predictor words are mainly in line with expectations about the political difference of radical-right parties in Germany, Austria, and the Netherlands. While the Dutch radical-right parties are primarily distinguished by their reference to immigration, as well as informal language, the German and Austrian radical-right parties resemble each other in their rejection of gendered language. The at-the-time governing Austrian FPÖ displays increased reference to the opposition, while the German AfD is more likely to address the government.

4.2 Construct validity I: coalition governments

This section presents data from government formations between radical-right and centre-right parties in Austria and the Netherlands. After a brief description of the cases, SMLSEs will be presented and their implications discussed. In line with expectations, the data shows an increased similarity between centre-right and radical-right parties when they govern together.

The Austrian case has seen the formation of three coalition governments between the conservative ÖVP and the radical-right FPÖ (and, briefly, the BZÖ) in 2000, 2003, and 2017. In 2000, the ÖVP managed to secure the chancellorship as the third-largest party by forming a coalition with the radical-right FPÖ. This broke the *cordon sanitaire* formed by the social-democrat SPÖ and ÖVP to isolate the FPÖ after the takeover of right-win populist Jörg Haider in 1986. The formation was met by strong internal and external criticism due to the FPÖ's radical xenophobic policy positions, resulting in diplomatic sanctions against the country and large-scale demonstrations. The coalition broke down in 2002 due to internal rifts within the FPÖ, forcing early elections. This time, the ÖVP emerged as the largest party and entered coalition talks with the Greens, which eventually broke down. A coalition government was again formed with the FPÖ. In April 2005, the FPÖ split, and the coalition continued with the split-off BZÖ until 2006 (Luther 2010).

The recent formation of the ÖVP-FPÖ coalition in 2017 was again met with domestic protests. After the experience of the so-called 'migration crisis' in 2015, immigration was the dominant topic in campaigns. ÖVP *Spitzenkandidat* Kurz attacked the FPÖ's issue ownership, also taking a restrictive position, especially surrounding the closure of the 'Balkan-route' for asylum seekers. This conveniently aligned these two parties' policies on this salient issue (Bodlos and Plescia 2018). The data under study ends before the collapse of the government

[8] See `https://www.fvd.nl/economie`

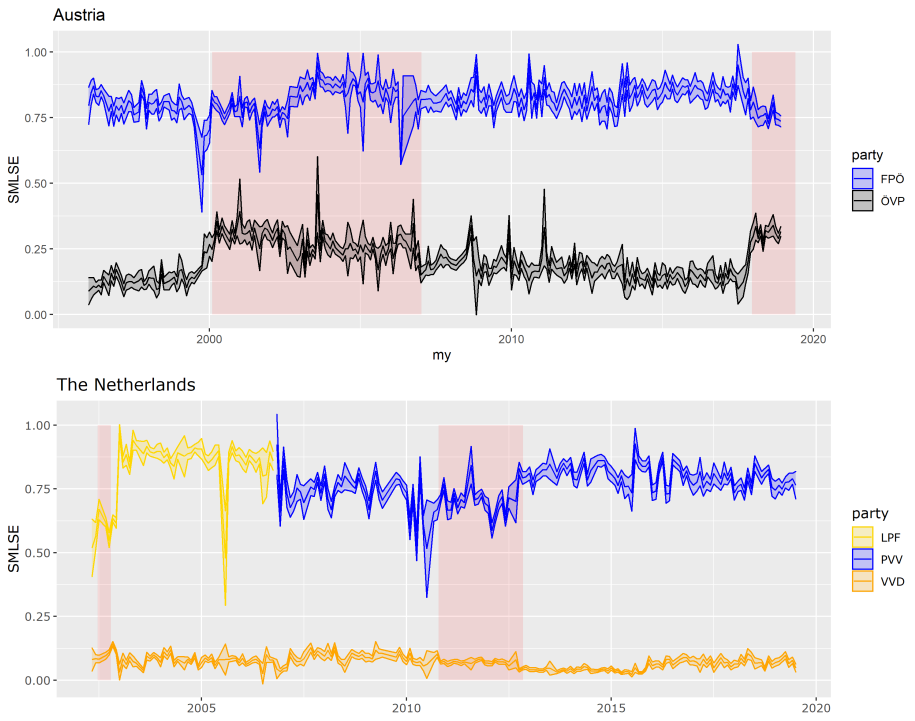

Figure 1: Monthly average similarity to radical-right parties for centre-right and radical-right parties in Austria and the Netherlands with 95% confidence intervals. Red shaded areas indicate coalition governments and cooperation in minority governments.

after the publication of the 'Ibiza-video' in 2019, showing FPÖ leaders Strache and Gudenus promising public contracts to a supposed Russian oligarch.

Summarised, I expect the FPÖ to be less distinctive and the ÖVP to be more alike the FPÖ while they govern together. This is especially true for the government formations in 2000 and 2017, where the ÖVP opted out of the mainstream coalition with the social democrats.

The top row in figure 1 shows the monthly SMLSEs for ÖVP and FPÖ across time. The BZÖ was excluded for ease of interpretation[9], but the estimate contains the classifier estimate for similarity to either radical-right party, i.e. the likelihood that the speaker belongs to *either* BZÖ or FPÖ[10]. The red shaded areas are the periods in which the ÖVP governed together with either of the two parties. A first observation is that the parties seem to move closer together when governing. This is confirmed by t-tests, which show that the distance between the two parties is significantly reduced ($t=-10.2$, $p<0.001$). Interestingly, this

[9]See figure 5 in appendix B1 for a graph including the BZÖ.
[10]See appendix B2, figure 6 for estimates of 'FPÖ-ness' only.

seems to be driven largely by the ÖVP (t=12.4, p<0.001), while the FPÖ is only slightly less distinctive in government (t=-1.68, p<0.1).

More interesting patterns are visible. Preceding the formations in 2000 and 2018, the conservatives became more similar to the radical-right, possibly signalling interest in coalition formation with the party. In 2003 however, the party becomes less similar for a short period[11] - they were in coalition talks with the Greens at the time and either expected heavy concessions from the FPÖ (which was already asking to govern again), or wanted to change course (Luther 2010: 91). The FPÖ, on the other hand, became less distinctive before the first formation in 2000 - possibly related to Haider's efforts to calm worries about the parties' Nazi past (ibid: 84); but more distinctive before the formation in 2018, where the ÖVP challenged the FPÖ's issue ownership on immigration, while the FPÖ underlined its 'orginalness' (Bodlos and Plescia 2018). It is also visible how the FPÖ became more distinctive after the 'Knittelfeld crisis' in September 2002, where party members challenged their leadership in open opposition to their course (Luther 2003).

The Netherlands saw only one formal coalition government between radical-right and centre-right, but a additional minority government supported by the radical-right PVV. In 2002 the *Lijst Pim Fortuyn* (LPF) brought unprecedented volatility and polarisation into Dutch politics (Bischof and Wagner 2019; Van der Brug 2003). Only formed about three months preceding the election, the party gained wide attraction due to its charismatic leader, a strong media presence and its distinctive anti-immigration stance (Koopmans and Muis 2009). On election day, nine days after the party leader Pim Fortuyn was assassinated, the party came in second with 17% of the vote and formed a coalition with the christian-democrat CDA and the centre-right VVD. Having lost its founding father and as a result of its rapid success, the party was unprepared for government and was turmoiled by internal power struggles. This peaked when two LPF ministers stopped talking to each other. As a result, CDA and VVD broke the coalition only 87 days after its formation (Heinisch 2003; Lucardie and Voerman 2007). I expect especially the LPF to become less distinctive here, as the party members decide to enter a coalition government. Similarly, CDA and VVD should become more alike the LPF in government.

The co-operation of VVD, CDA and PVV in 2010 did not result in a formal majority coalition (partly due to internal conflicts in the CDA regarding a coalition with the radical-right PVV), but a 'supported minority government'. This was the first minority government formed in the Netherlands since 1922. VVD and CDA staffed the cabinet, while the PVV only promised support on major issues. As this support was settled in a formal agreement, this government became basically a 'majority government in disguise' (Strøm 1990; Van Holsteyn 2011). In fact, the government cooperated very little with the opposition even compared to full majority governments - this is likely an outcome of the sorting of parties into a right-wing government and a left-wing opposition

[11] This is more pronounced in the FPÖ-only estimate excluding the BZÖ, see figure 6 appendix B2.

but underlines the similarity of this government to classic majority coalitions (Otjes and Louwerse 2014). As a result, I expect an increased similarity of the coalition partners, slightly less than in a majority coalition.

The lower row of figure 1 shows the monthly average SMLSE in the Netherlands for PVV, LPF, and VVD[12]. The LPF is rather similar to all other parties when in government, obtaining a distinctiveness estimate of only around 60%[13] throughout the time governing. Once the party leaves the government, the party becomes very distinctive again. For the VVD, we can observe a slightly increased score when governing with the LPF. For the supported minority government in 2010, things are far less clear. The VVD's estimate seems to remain rather stable, with possibly a slight *decrease*, while the PVV becomes much less distinctive just *before* signing the support agreement. Afterwards, the PVV stays somewhat indistinctive throughout the time supporting the government at around 60% - 70%. After the time in government, the party's communication becomes more distinctive again.

T-tests confirm these observations. The LPF does show a significantly decreased distinctiveness in government ($t=-6.15$, $p<0.01$), as does the PVV ($t=-5.36$, $p<0.001$). The VVD only shows slightly increased similarity to the respective radical-right party when governing with the LPF ($t=3.21$, $p<0.1$), but when supported by the PVV, this effect becomes virtually zero ($t=0.6$, $p=0.55$). The CDA is assigned similar SMLSEs, though more clearly reacting to the LPF ($t=7.1$, $p<0.01$), but neither to the PVV ($t=0.96$, $p=0.34$). Overall, the VVD is not significantly affected by governing with a radical-right party ($t = 1.64$, $p>0.1$), while the CDA is significantly changing its rhetoric ($t=2.01$, $p<0.05$).

Additionally, throughout the time in government, the LPF becomes less distinctive, while the VVD seems to become more similar. This is in line with an ideological shift in the VVD's manifesto, which took up a more anti-immigrant position (Pennings and Keman 2003), while the LPF dropped some of it's populist rhetoric (Lucardie and Voerman 2007). For the PVV, a shift towards the centre is visible before the formation of the minority support government - possibly signalling willingness to govern, similar to Haider's FPÖ preceding the Austrian elections in 2000.

The findings are in line with my expectations: governing parties are more similar to each other - the estimates display construct validity. Interesting variation can be observed across countries and different government formations. While it is mainly the centre-right which accommodates the radical-right in Austria, it is usually the radical-right which becomes more similar in the Netherlands, even when only supporting a minority government. This might have to do with the differing coalition logic in the two countries. The ÖVP held the ideological middle-ground of the three major parties and was able to choose with whom to form a coalition - both in 2000 and 2017, the party decided against a continuation of the coalition with the social democrats. In the Netherlands,

[12] CDA and FvD were excluded to maintain readability. The full plot can be found in figure 7, appendix C.

[13] This value can be interpreted as the confidence with which the classifier identifies LPF speeches as such.

coalitions often include many parties, which gives those in charge of forming a government many options. The moderation on behalf of the PVV may have been necessitated by the resistance from parts of the CDA against the formation of a coalition with the PVV. Interesting is also that radical-right parties often moderated before coalition formations, possibly signalling willingness to govern. Only the FPÖ reacted to the ÖVP's accommodation in 2017 with further distinction. These observations indicate possible starting points for future research with this method and underline the richness of the data.

4.3 Construct validity II: speaker estimates

Beyond parties, the similarity or distinctiveness of particular speakers can also be of interest. I use two cases from Germany and the Netherlands to further validate the method's precision. In both cases, I have strong expectations about the placement of individual speakers. In Germany, six members of the AfD have since left the party, mainly due to the increasing strength of the far-right faction within the radical-right party (Steffen 2020). Two of these members, one of whom was former party leader Frauke Petry left within the first ten days following the election (Göppfarth 2018). As they were members of the party themselves, these members should obtain higher similarity scores than members of other parties[14]. Similarly, four members of the AfD who left the party later (outside of the observed time-frame) should obtain lower similarity scores, compared to loyal members of the AfD.

Figure 2 shows the mean estimate for each speaker in the current Bundestag. Each speaker is represented by a dot. The dot's position on the vertical axis indicates the speaker's similarity to the AfD, according to the SMLSE, while its color and horizontal position indicate the speaker's party affiliation. The classifier assigns a probability greater than 0.5 for all AfD-members and a probability smaller than 0.5 to all speakers from established parties. In line with my expectation, the now independent former members of the AfD obtain rather high values (Frauke Petry: 52.6%, Mario Mieruch: 39.3%) and are located in between the AfD and the other parties. The mean estimate for all speeches by these two speakers is distinctive from all other parties. Despite being labelled not to be part of the AfD in the training data, the classifier correctly locates them in proximity of the AfD.

Uwe Kamann would leave the AfD five days after the observed period ends, on December 17th, 2018. His estimate is relatively low compared to his party, however several members obtain lower estimates. Lars Herrmann and Verena Hartmann, two other members that left the AfD, followed in winter 2019/2020, about one year after the data ends. In the data observed, the classifier still places them well inside their party (see two highlighted, unlabelled dots within the AfD). For these ex-members of the AfD, the expectations are not confirmed. It might be the case that they still move to the fringes of their party in the following

[14] Note that the data contains no speeches by these two independent members preceding their exit, excluding a longitudinal analysis comparing their similarity before and after exit.

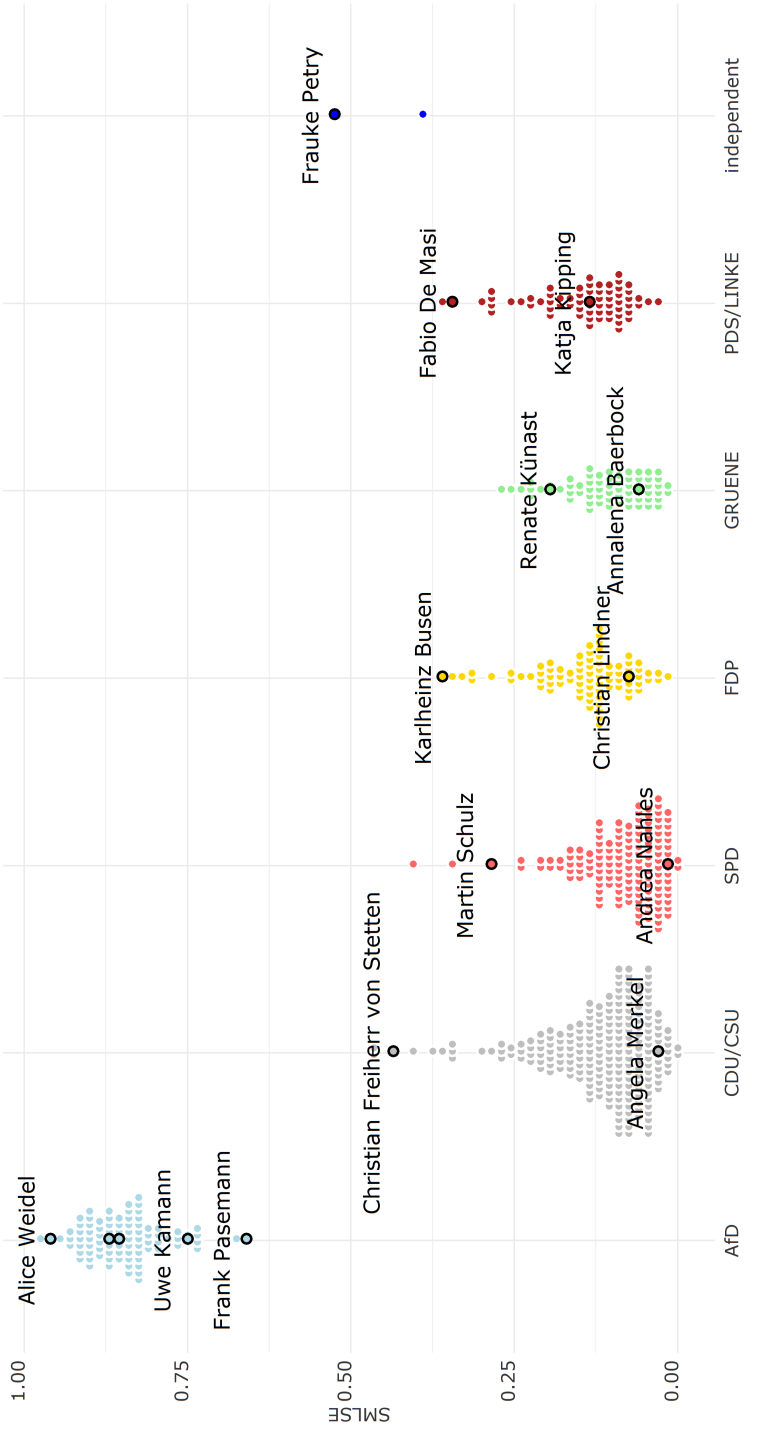

Figure 2: SMLSEs for speakers of the current German Bundestag.

year. The estimates for Frank Pasemann, the sixth member in the data to leave the party (and the only one to do so involuntarily), exemplify well how SMLSEs work. Pasemann was expelled from the AfD in August 2020[15], most likely as a result of anti-semitic statements and his proximity to members of neonazi- and radical-right associations[16]. As the concept being measured is 'AfD-ness' (specified as the outcome category in the training data), the classifier assesses how well a speech fits into the corpus of AfD-speeches. As a result, Pasemann, who is clearly located to the right of his party, obtains lower estimates than other members of his party.

Beyond these specific members, all parties show considerable variation. The overall distribution of the two governing mainstream parties is skewed towards the lower end of the scale, with only few members in the tails with higher probabilities. This tail is more pronounced for the CDU/CSU. Among the conservatives, the measure seems to have an ideological component: those with a high probability of being members of the AfD also seem to have more conservative positions. While Angela Merkel is assigned a low similarity score (2.9%), Christian Freiherr von Stetten (43.2%) and other more conservative members of the party are placed more closely to the AfD. Stetten is not very present in the media, he was a vocal advocate of a conservative turn of his party under the leadership of Friedrich Merz (Weinzierler 2019), and has been affiliated with the conservative WerteUnion, which works towards a right-wing turn in German politics[17]. Within the SPD, it is interesting that delegates from former industry regions in North Rhine-Westphalia and Eastern Germany like Dirk Vöpel (40.8%), Detlef Mueller (34.9%), Martin Schulz (28.9%), and Josip Juratovic (24.3%) are assigned higher values compared to their party. However note that e.g. Martin Schultz is a fierce advocate against the radical-right[18]. This might indicate rhetorical rather than ideological similarity - the classifier does not distinguish.

The estimates for the other opposition parties are somewhat more spread out than the two governing parties, but with similar tails towards the top. Within the FDP, Karlheinz Busen gets assigned the highest estimate of his party (35.3%). His political positions revolve around agricultural matters [19], which is an issue that the AfD is increasingly trying to mobilise on [20]. Among the Linke, several members obtain rather high estimates, most prominently Fabio de Masi (34.6%), a politician with a strong focus on EU-policy, where the party holds a Eurosceptic position[21] (as does the AfD; Rooduijn et al. 2019).

In the Netherlands, the rise of the LPF and the subsequent politicisation of

[15] https://www.mdr.de/sachsen-anhalt/landespolitik/afd-politiker-pasemann-aus-partei-ausgeschlossen-100.html

[16] https://www.mdr.de/investigativ/afd-pasemann-ausschluss-extrem-rechte-100.html

[17] He is quoted as a supporter (including a picture of his face) on the movements website: `https://werteunion.net/`

[18] See e.g. https://www.tagesspiegel.de/politik/ex-spd-chef-attackiert-afd-vorsitzenden-schulz-wuenscht-gauland-auf-den-misthaufen-in-der-deutschen-geschichte/23057642.html

[19] `https://karlheinz-busen.de/`

[20] `https://www.sueddeutsche.de/politik/afd-bauern-landwirte-1.4764413`

[21] `https://www.fabio-de-masi.de/de/topic/15.eurokrise.html`

immigration forced other parties to react to the issue (De Vries and Hobolt 2012; Pennings and Keman 2003). This proved a wedge issue for the VVD, pitting the social-liberal faction around state secretaries Mark Rutte and Melanie Schultz against party members demanding a shift to the right, most famously Geert Wilders (Van de Wardt, De Vries, and Hobolt 2014; Vossen 2011). After being described as a likely candidate for party leadership of the VVD in the early 2000s (Vossen 2011, 182), Wilders developed an increasingly extreme anti-islamism from 2003 onwards (Vossen 2010, 26). This alienated him from the social-liberal wing of the VVD and the party's leadership, especially parliamentary group leader van Aartsen. In 2004, Wilders published a position paper together with his fellow VVD MP Gert-Jan Oplaat calling for a right-wing turn in the VVD. It was speculated whether Wilders would join the LPF, which held similar anti-muslim and anti-immigration positions (Soetenhorst 2004; Stokmans 2008). After several calls by the leadership to follow the party line, he left the VVD to form the radical-right PVV, while the VVD took a more centrist course (Vossen 2010; Vossen 2011).

As Wilders develops more anti-muslim positions, I expect him to become more similar to the LPF in the time preceding his exit, compared to other members of the VVD. This should especially be the case when compared to proponents of the social-liberal course of the VVD. Immigration-critical members of the VVD like Gert-Jan Oplaat (who co-authored the position paper with Wilders), Ayaan Hirsi Ali (who co-authored an op-ed with Wilders calling for a 'liberal Jihad'; Vossen 2011, 183) and Rita Verdonk (who later formed her own populist party *Trots*) should be placed in proximity of Wilders.

The left graph in figure 3 shows the quarterly mean SMLSE of Geert Wilders and (at the time) more socially-liberal oriented Mark Rutte, from the first entry of the LPF until Wilders' exit. It is visible that in 2002 and 2003, both estimates overlap at low levels. The classifier gives both speakers a 0-10% chance to belong to the LPF. In line with my expectation, Wilders' estimate moves upwards from mid-2003, indicating higher similarity to the radical-right LPF, until mid-2004 (the last estimate before Wilders left the party), when the classifier assigns a likelihood over 20% that this speaker might belong to the LPF. This is especially extreme compared to the significantly distinctive estimate for Rutte at below 5%. It seems that after the election, an issue sorting takes place, as Wilders embraces the LPF's position, while Rutte further distances himself (Carmines and Stimson 1986). After his exit, Wilders' language changes again, becoming less alike the LPF.

Based on the scores alone, it is unclear what drives the increased similarity of Wilders' language to the LPF. The expectation is that the distinction should revolve around Islam and immigration, the major issue that Wilders' party mobilized on later (Van Holsteyn 2011). Disagreement on these issues was also reflected in the support agreement between VVD, CDA and PVV in 2010 (Otjes and Louwerse 2014, 350). I calculate each word's influence on Wilder's SMLSE by multiplying the model coefficient with the (tfidf-weighted) word count from all his speeches between the second CDA-VVD-LPF government formation until his exit from the VVD in 2004. The 30 words with the strongest positive

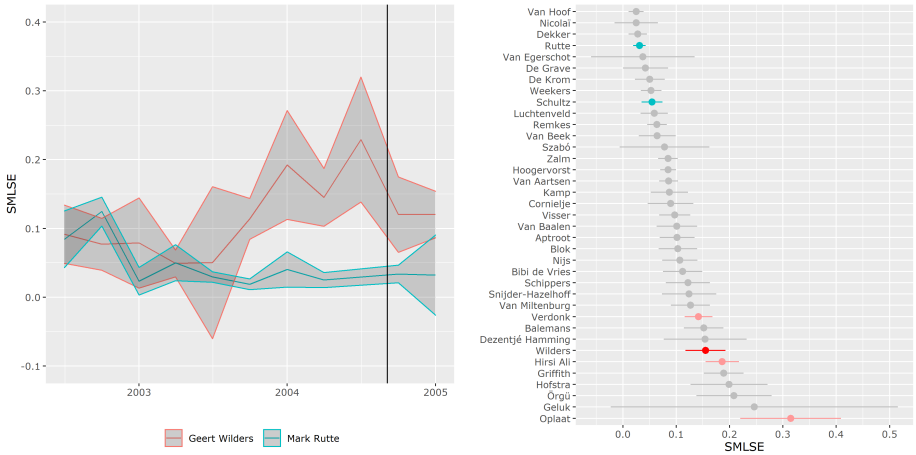

Figure 3: **Left**: quarterly SMLSE estimates for Geert Wilders and Mark Rutte in 2004. The vertical line indicates the date Wilders' left the party. **Right**: Placement of Wilders (red), compared to social-liberal (turquoise) and right-wing members (light-red) within the VVD in 2004, preceding Wilders' exit. Both graphs show the estimates with 95% confidence intervals.

influence[22] contain several terms related to Islam ('imam', 'mosques'), security ('AIVD'[23], 'Defense'), terrorism ('terrorists'), and muslim countries ('Algeria', 'Saudi', 'Arabia'), as well as 'Europe'. It seems that Wilders' increased similarity in this period is indeed related to his increased attention to Islam.

The right side of figure 3 shows Wilders' placement compared to other VVD members. In general, Wilders is placed relatively similar to the LPF, obtaining the 7th-highest estimate (20%) of the 37 members covered by the data. Likewise, other members opposed to Islam and immigration (Hirsi Ali, Oplaat, Verdonk) obtain similarly high estimates, while the proponents of a social-liberal course (Rutte, Schultz) are assigned relatively low estimates.

Although far more detailed, the speaker estimates are in line with expectations, underlining the precision of the method. The former members of the AfD were placed in between the AfD and all other parties. The member leaving the party shortly after the observed period is placed relatively low within the party. Surprisingly, the two members who would leave one year later are placed relatively central within the party - at the time, there was no indication that they might leave. Assessing the VVD in 2004, the estimates correctly place social-liberal and more right-wing members within the party. Even more impressive, Geert Wilders' divergence from the VVD from 2003 onwards is significant in the estimates.

[22] See table 3 in appendix D.
[23] AIVD is the Dutch intelligence service.

5 Comparative Performance

The SMLSE estimates conform to expectations in all described cases. Nevertheless, existing methods could be used to estimate rhetorical similarities. This section compares SMLSEs for the current German legislature to cosine similarity and wordfish estimates. Cosine similarity is a relatively simple measure, which has seen political science applications (see e.g. Egesdal, Gill, and Rotemberg n.d.; Hager and Hilbig 2020). It takes the cosine of the angle of two document vectors to assess their similarity. It equals one when the two compared documents are identical in their relative word usage and zero when the documents share no common words. Scaling methods like wordscore (Laver, Benoit, and Garry 2003) and wordfish (Slapin and Proksch 2008) have become popular to estimate differences in party communication, mainly to place texts on an ideological scale.

5.1 Cosine similarity

Cosine similarity is calculated for each document, comparing it to all AfD-authored speeches, then taking the mean similarity score. This results in very low overall similarity scores, ranging between 0.04 and 0.05. Figure 4 shows the mean speech estimates for SMLSE, cosine similarity, and wordfish for each party in the current German legislative period. While the AfD is clearly distinguished by the SMLSE, there are only slight party differences for the cosine measure, with comparatively large confidence intervals. Cosine similarity is unable to distinguish AfD speeches, with the AfD only ranked *third* in similarity to AfD speeches. This is in line with weak performance of this measure when meaning is relevant (as opposed to e.g. the detection of plagiarism, where literal similarity is relevant; Prasetya, Wibawa, and Hirashima 2018).

The weak performance becomes more obvious once the measure is correlated with speech length: over 75% of the variance in cosine similarity are explained by the length of the speech (Pearson's r = 0.87). It seems that the increased likelihood of a word to be included is the main driver of increased similarity here. The correlation of cosine similarity with the SMLSE is in fact weakly *negative* (-0.17)[24]. AfD authorship is statistically unrelated to cosine similarity (correlation of -0.02). Calculating cosine similarity of these groups differently by first generating the average word count of AfD speeches and then calculating cosine similarity results in virtually identical results. I also calculated Jaccard[25] similarity with similarly weak performance.

5.2 Wordfish

To assess how parties and speakers differ in word use and whether they are more or less similar, a researcher could also use scaling methods. As wordscore

[24] When using the logged SMLSE to normalise the distribution, this decreases to -0.36.

[25] Jaccard similarity assesses the share of common words among all words used (it is unrelated to how often the words are used).

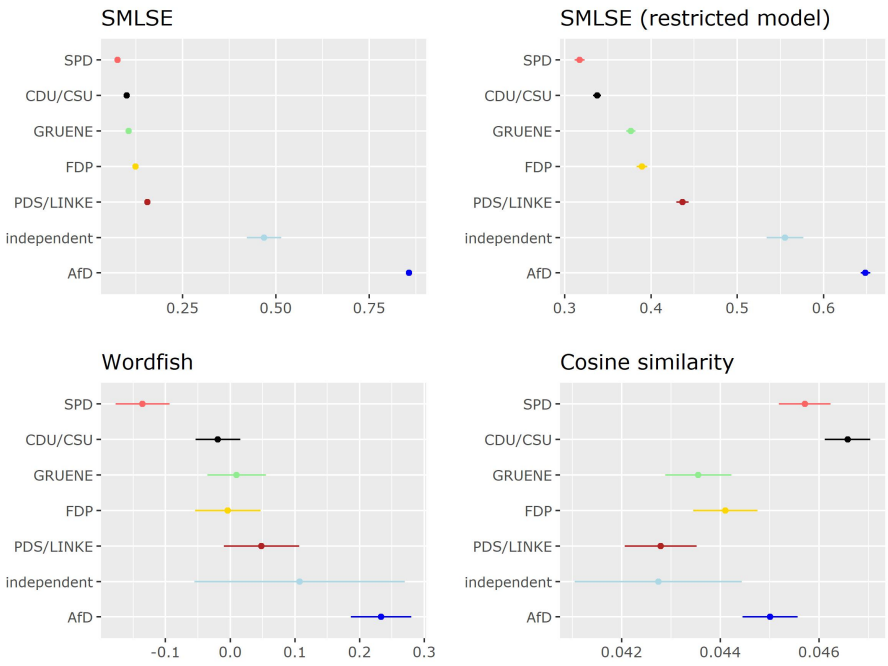

Figure 4: Mean party estimates with 95% confidence intervals for SMLSE, cosine similarity and wordfish estimates. Both wordfish and the restricted SMLSE model were trained on a subset of 1000 speeches, which then estimated the positions of all speeches.

(Laver, Benoit, and Garry 2003) requires the selection and labelling of anchor documents (and hence the introduction of subjective bias), I compare the estimates from the supervised algorithm to wordfish-scores. Wordfish extracts a major underlying dimension which explains differences in word use (Slapin and Proksch 2008). The SMLSE is conceptually different from wordfish estimates. Instead of estimating an underlying dimension explaining differences in word use, the SMLSE estimates a statistical model where word use explains an outcome variable (e.g. radical-right authorship) and then employs this model to assess the similarity of documents to the group of interest.

The last row in figure 4 shows the wordfish estimates for each party. As the process was to computationally intense to be run with the full set of 11,419 German speeches[26], I ran the scaling estimates on a sample of 1,000 speeches with at least 50 words. To make a fair comparison, the upper right graph in 4 shows SMLSEs for such a subsample. A first observation is that the wordfish distributions show far more overlap than the SMLSEs[27]. Other than cosine similarity measures, wordfish places the AfD very distinctly. It also correctly places the independent speakers in proximity to the AfD. Similar to the SMLSEs, the distribution does not follow a left-right pattern, but places the governing parties and especially the SPD as most distinct, the opposition parties in between and the independent speakers in proximity to the AfD. However, these estimates also show large confidence intervals and most parties' mean estimates cannot be distinguished. It is only somewhat correlated with the SMLSE (0.17, logged 0.22), although the order of the parties' mean estimates is similar - with the exception of the FDP. It seems that on the main underlying dimension structuring the differences in word use, the FDP is a little closer to the governing parties, while when the dimension of interest is the similarity to the AfD, the FDP is somewhat closer to the AfD, compared to e.g. the Greens. Furthermore, the interpretation of these scores is harder, as it does not correspond to similarity towards a group (or distinctiveness of that group). Instead, the estimates provide the placement of groups on the best fitting underlying dimension to explain the differences in word use. The interpretation of this dimension is left to the researcher. In this example, there seems to be a mixture of left-right and government-opposition dynamics[28].

Lastly, the runtime of the estimations is also relevant, as researchers want fast analyses that save them time. I compare the runtime of wordfish estimates to that of the oversampling and model fitting for SMLSE, both for the subsample of 1,000 speeches from the most recent parliamentary session in Germany. Wordfish is estimated using Will Lowe's `austin`-package for R[29] on Windows; SMLSE's are estimated using `scikit-learn` for Python on Linux. Although

[26] On a machine with 16GB working memory.

[27] Note that the sign of the wordfish estimates has been inverted to simplify comparison.

[28] Also note that this example likely downplays the conceptual difference of wordfish and SMLSEs, as the AfD is an extreme case, where similarity seems rather correlated with the main underlying factor. This should be different if similarity towards a centrist party is investigated.

[29] `https://conjugateprior.github.io/austin/articles/austin.html`

wordfish scores could well be estimated with Python and SMLSEs in R (and speed differences partly be caused by these different applications), these platforms reflect the most likely use case for both applications. Model fit and prediction for each speech with wordfish took 44 minutes for a sample of 1000 cases (35 minutes for model fit alone). The runtime for oversampling, model fit and generation of SMLSEs was around half a *second* when fitting 1000 speeches, and 3.9 seconds on the full set of 11,419 speeches. Apart from taking far less memory and enabling the estimation of far larger datasets, SMLSEs are around 5,000 times faster than the standard implementation of wordfish.

6 Conclusion

The precise measurement of rhetorical similarities is at the heart of many political science questions. This paper presented a novel approach making use of machine learning for the precise and efficient estimation of the similarity of documents to corpora. It allows researchers to estimate how well a given document fits a category of their choosing, compared to a number of other texts. The cases presented here show that SMLSEs provide valid estimates even for longitudinal data with fewer observations. It has also been shown that the estimates produced outperform applications using scaling methods or measures of lexical similarity.

The approach is particularly well-suited to study the quality as well as drivers and effects of party accommodation and it was developed with this application in mind. Given that researchers can use virtually any classification of their texts to estimate similarity scores, a plethora of other applications is conceivable. Building on the preliminary findings presented here, a more systematic inquiry might try to predict coalition formations or breakdowns based on the 'signalling' in parties' rhetoric. Vice versa, one could assess coalition inclusion probabilities (Kayser, Orlowski, and Rehmert n.d.) and their conditioning effect on accommodation. The approach could also be exploited to explore the validity of party families by assessing the best predictor words and their underlying dimensions (e.g. through coupling it with factor analysis). That way, it could be established whether members of the same party family are distinguished by similar language. Scholars of representation might use SMLSEs to assess whether and how certain groups of MPs differ in their political rhetoric, and whether they adapt to their peers across time in office. Given the validity of the estimates even on the speaker-level across time, researchers might apply this to predict party movements in response to leadership changes by assessing a new leaders' position compared to the party preceding the change.

Further avenues might also include historical analyses assessing whether certain parties or movements managed to shift public discourse or whether parties have filled opportunity spaces abandoned by other parties. But applications are not restricted to party competition. Peace and conflict scholars might be concerned with insurgent's communication to assess willingness for peace agreements. Communication scholars could explore the similarity of newspapers.

Lastly, two notes of caution. I have shown that SMLSEs excel in comparison to other methods when assessing similarities of documents to corpora. In my application, evidence of the in- and decreasing similarity of established and radical-right parties has been shown. This might nevertheless 'mask' changes in the overall discourse, which might be more or less similar to a certain party. If, for example, a radical-right party's rhetoric has become common among all parties, but the radical-right party has further radicalised to a similar extent, the measure would only capture this within legislative terms, not for more long-term developments[30]. If a researcher is interested in how parties affected long-term changes in the political discourse, a similar method might be used, but without training different models for each time-unit.

Furthermore, my claim is not that this method is *per se* superior. Instead, practitioners need to select a tool suited to their research interest. The necessity to self-define the outcome of interest gives the researcher considerable flexibility, but also puts the responsibility to properly define the concept of interest in their hands alone. Additionally, established measures might be more advantageous with different research questions. If similarities to single documents are of interest, cosine similarities should be more useful. When instead of similarity, the researcher is concerned with the placement of labels on underlying dimensions explaining differences in word use (e.g. ideology), scaling methods are the way to go. If the researcher's curiosity however revolves around group similarities, SMLSE should indeed be the method of choice. In this way, the paper has contributed to extend the toolbox of empirical social scientists.

Acknowledgements

This paper was enriched by helpful comments from Anna Adendorf, Britt Anlar, João Areal Neto, Eelco Harteveld, Eva Hoxha, Hauke Licht, Philipp Mendoza, Thomas Meyer, Pieter Moens, and Gijs Schumacher.

References

Arzheimer, Kai (2009). "Contextual factors and the extreme right vote in Western Europe, 1980-2002". In: *American Journal of Political Science* 53.2, pp. 259–275. ISSN: 00925853. DOI: 10.1111/j.1540-5907.2009.00369.x.

Arzheimer, Kai and Carl C. Berning (2019). "How the Alternative for Germany (AfD) and their voters veered to the radical right, 2013–2017". In: *Electoral Studies* 60.November 2018, p. 102040. ISSN: 02613794. DOI: 10.1016/j.electstud.2019.04.004. URL: https://doi.org/10.1016/j.electstud.2019.04.004.

Bale, Tim et al. (2010). "If You Can't Beat Them, Join Them? Explaining Social Democratic Responses to the Challenge from the Populist Radical Right in Western Europe". In: *Political Studies* 58.3, pp. 410–426. ISSN: 00323217. DOI: 10.1111/j.1467-9248.2009.00783.x.

Bischof, Daniel and Markus Wagner (2019). "Do Voters Polarize When Radical Parties Enter Parliament?" In: *American Journal of Political Science* 63.4, pp. 888–904. ISSN: 15405907. DOI: 10.1111/ajps.12449.

[30] I thank Pieter Moens for raising this point.

Bodlos, Anita and Carolina Plescia (2018). "The 2017 Austrian snap election: a shift rightward". In: *West European Politics* 41.6, pp. 1354–1363. ISSN: 17439655. DOI: 10.1080/01402382.2018.1429057. URL: http://doi.org/10.1080/01402382.2018.1429057.

Breiman, Leo and Philip Spector (1989). *Submodel Selection and Evaluation in Regression - the X-Random Case*. Tech. rep. Berkeley: University of California.

Carmines, Edward G. and James A. Stimson (1986). "On the Structure and Sequence of Issue Evolution". In: *The American Political Science Review* 80.3, pp. 901–920.

Carmines, Edward G. and James Woods (2004). "Validity". In: *The SAGE Encyclopedia of Social Science Research Methodology*. Ed. by Michael S. Lewis-Beck, Alan Bryman, and Tim Futing Liao. SAGE, pp. 1171–1172.

Chawla, Nitesh V. et al. (2002). "SMOTE: Synthetic Minority Over-sampling Technique". In: *Journal of Artificial Intelligence Research* 16, pp. 321–357. ISSN: 10769757. DOI: 10.1613/jair.953. arXiv: 1106.1813. URL: https://arxiv.org/pdf/1106.1813.pdf%7B%5C%%7D0Ahttp://www.snopes.com/horrors/insects/telamonia.asp.

Dahlström, Carl and Anders Sundell (2012). "A losing gamble. How mainstream parties facilitate anti-immigrant party success". In: *Electoral Studies* 31.2, pp. 353–363. ISSN: 02613794. DOI: 10.1016/j.electstud.2012.03.001.

De Vries, Catherine E. and Sara B. Hobolt (2012). "When dimensions collide: The electoral success of issue entrepreneurs". In: *European Union Politics* 13.2, pp. 246–268. ISSN: 14651165. DOI: 10.1177/1465116511434788.

Denny, Matthew J. and Arthur Spirling (2018). "Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It". In: *Political Analysis* 26.2, pp. 168–189. ISSN: 14764989. DOI: 10.1017/pan.2017.44.

Downs, Anthony (1957). "An Economic Theory of Political Action in a Democracy". In: *Journal of Political Economy* 65.2, pp. 135–150. ISSN: 0022-3808. DOI: 10.1086/257897.

Egesdal, Michael, Michael Gill, and Martin Rotemberg (n.d.). "How Federal Reserve Discussions Respond to Increased Transparency". URL: https://ssrn.com/abstract=2676429.

Gamson, William A. (1961). "A Theory of Coalition Formation". In: *American Sociological Review* 26.3, pp. 373–382.

Gentzkow, Matthew, Jesse M. Shapiro, and Matt Taddy (2019). "Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech". In: *Econometrica* 87.4, pp. 1307–1340. ISSN: 0012-9682. DOI: 10.3982/ecta16566.

Gilens, Martin and Benjamin I. Page (2014). "Testing theories of American politics: Elites, interest groups, and average citizens". In: *Perspectives on Politics* 12.3, pp. 564–581. ISSN: 15375927. DOI: 10.1017/S1537592714001595.

Goet, Niels D. (2019). "Measuring Polarization with Text Analysis: Evidence from the UK House of Commons, 1811-2015". In: *Political Analysis* 27.4, pp. 518–539. ISSN: 14764989. DOI: 10.1017/pan.2019.2.

Göppfarth, Julian (Sept. 2018). "Heading into the mainstream? Reviewing a year of the AfD in the German parliament". In: *LSE European Politics and Policy (EUROPP) Blog*, pp. 1–4. URL: http://blogs.lse.ac.uk/europpblog/2018/09/05/heading-into-the-mainstream-reviewing-a-year-of-the-afd-in-the-german-parliament/.

Grimmer, Justin and Brandon M. Stewart (2013). "Text as data: The promise and pitfalls of automatic content analysis methods for political texts". In: *Political Analysis* 21.3, pp. 267–297. ISSN: 14764989. DOI: 10.1093/pan/mps028.

Grofman, Bernard and Peter Van Roozendaal (1994). "Toward a theoretical explanation of premature cabinet termination: With application to post-war cabinets in the Netherlands". In: *European Journal of Political Research* 26.2, pp. 155–170. ISSN: 14756765. DOI: 10.1111/j.1475-6765.1994.tb00438.x.

Hager, Anselm and Hanno Hilbig (2020). "Does Public Opinion Affect Political Speech?" In: *American Journal of Political Science* 00.00, pp. 1–17. ISSN: 15405907. DOI: 10.1111/ajps.12516.

Harmel, Robert and Lars Svåsand (1997). "The influence of new parties on old parties' platforms: The cases of the progress parties and conservative parties of Denmark and Norway". In: *Party Politics* 3.3, pp. 315–340. ISSN: 13540688. DOI: 10.1177/1354068897003003003.

Heinisch, Reinhard (2003). "Success in opposition - Failure in government: Explaining the performamce of right-wing populist parties in public office". In: *West European Politics* 26.3, pp. 91–130. ISSN: 01402382. DOI: 10.1080/0140238031233128060 8.

Hirschman, Albert O. (1970). *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States*. Cambridge: Harvard University Press.

Jaccard, Paul (1912). "The Distribution of the Flora in the Alpine Zone." In: *New Phytologist* 11.2, pp. 37–50. ISSN: 14698137. DOI: 10.1111/j.1469-8137.1912.tb05611.x.

Kayser, Mark A, Matthias Orlowski, and Jochen Rehmert (n.d.). "Coalition Inclusion Probabilities: A Dynamic Measure of Party Competitiveness and Cabinet Leverage". URL: http://mark-kayser.com/papers/KOR%7B%5C_%7DCoalProbs%7B%5C_%7D190125.pdf.

Kitschelt, Herbert P. (1986). "Political Opportunity Structures and Political Protest : Anti-Nuclear Movements in Four Democracies". In: *British Journal of Political Science* 16.1, pp. 57–85.

Koopmans, Ruud and Jasper Muis (2009). "The rise of right-wing populist Pim Fortuyn in the Netherlands: A discursive opportunity approach". In: *European Journal of Political Research* 48.5, pp. 642–664. ISSN: 03044130. DOI: 10.1111/j.1475-6765.2009.00846.x.

Krause, Werner, Denis Cohen, and Tarik Abou-Chadi (n.d.). "Does Accommodation Work? Mainstream Party Strategies and the Success of Radical Right Parties".

Lakoff, George and Mark Johnson (1980). *Metaphors we live by*. The University of Chicago Press.

Laver, Michael, Kenneth Benoit, and John Garry (2003). "Extracting policy positions from political texts using words as data". In: *American Political Science Review* 97.2, pp. 311–331. ISSN: 00030554. DOI: 10.1017/S0003055403000698.

Lemaitre, Guillaume, Fernando Nogueira, and Christos K Aridas (2017). "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning". In: *Journal of Machine Learning Research* 18, pp. 1–5. ISSN: 15337928.

Lowe, Will (2008). "Understanding wordscores". In: *Political Analysis* 16.4 SPEC. ISS. Pp. 356–371. ISSN: 10471987. DOI: 10.1093/pan/mpn004.

Lucardie, Paul and Gerrit Voerman (2007). "The list Pim Fortuyn and the Government: a love-hate relationship". In: *The extreme right parties and power in Europe*. Ed. by Pascal Delwit and Philippe Poirier. Bruxelles: Editions de l'Universite de Bruxelles, pp. 247–263.

Luther, Kurt Richard (2003). "The self-destruction of a right-wing populist party? The Austrian parliamentary election of 2002". In: *West European Politics* 26.2, pp. 136–152. ISSN: 01402382. DOI: 10.1080/0140238051233134114 1.

— (2010). "Governing with Right-Wing Populists and Managing the Consequences: Schuessel and the FPOE". In: *The Schuessel Era in Austria*. Ed. by Günter Bischof and Fritz Plasser. New Orleans: University of New Orleans Press, pp. 79–103.

Meguid, Bonnie M. (2005). "Competition between unequals: The role of mainstream party strategy in niche party success". In: *American Political Science Review* 99.3, pp. 347–359. ISSN: 00030554. DOI: 10.1017/S0003055405051701. URL: http://www.journals.cambridge.org/abstract%7B%5C_%7DS0003055405051701.

Mosteller, Frederick and David L . Wallace (1963). "Inference in an Authorship Problem". In: *Journal of the American Statistical Association* 58.302, pp. 275–309.

Mudde, Cas (2007). *Populist Radical Right Parties in Europe*. Cambridge University Press.

Otjes, Simon and Tom Louwerse (2014). "A special majority cabinet? Supported minority governance and parliamentary behavior in the Netherlands". In: *World Political Science Review* 10.2, pp. 343–363. ISSN: 19356226. DOI: 10.1515/wpsr-2014-0016.

Pedregosa, Fabian et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830. ISSN: 2375-0529. DOI: 10.1145/2786984.2786995.

Pennebaker, James W., Matthias R. Mehl, and Kate G. Niederhoffer (2003). "Psychological Aspects of Natural Language Use: Our Words, Our Selves". In: *Annual Review of Psychology* 54, pp. 547–577. ISSN: 00664308. DOI: 10.1146/annurev.psych.54.101601.145041.

Pennings, Paul and Hans Keman (2003). "The Dutch Parliamentary Elections in 2002 and 2003: The Rise and Decline of the Fortuyn Movement". In: *Acta Politica* 38.1, pp. 51–68. ISSN: 0001-6810. DOI: 10.1057/palgrave.ap.5500001.

Peterson, Andrew and Arthur Spirling (2018). "Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems". In: *Political Analysis* 26.1, pp. 120–128. ISSN: 14764989. DOI: 10.1017/pan.2017.39.

Pitkin, Hanna F. (1967). *The Concept of Representation*. Berkeley and Los Angeles: University of California Press.

Prasetya, Didik Dwi, Aji Prasetya Wibawa, and Tsukasa Hirashima (2018). "The performance of text similarity algorithms". In: *International Journal of Advances in Intelligent Informatics* 4.1, pp. 63–69. ISSN: 25483161. DOI: 10.26555/ijain.v4i1.152.

Rauh, Christian and Jan Schwalbach (2020). "The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies". In: *Harvard Dataverse, V1*, pp. 1–14. DOI: https://doi.org/10.7910/DVN/L4OAKN,. URL: https://dataverse.harvard.edu/dataverse/ParlSpeech.

Rooduijn, Matthijs et al. (2019). *The PopuList: An Overview of Populist, Far Right, Far Left and Eurosceptic Parties in Europe*. URL: www.popu-list.org.

Schoonvelde, Martijn, Gijs Schumacher, and Bert N. Bakker (2019). "Friends with text as data benefits: Assessing and extending the use of automated text analysis in political science and political psychology". In: *Journal of Social and Political Psychology* 7.1, pp. 124–143. ISSN: 21953325. DOI: 10.5964/jspp.v7i1.964.

Schumacher, Gijs and Kees van Kersbergen (2014). "Do mainstream parties adapt to the welfare chauvinism of populist parties?" In: *Party Politics* 22.3, pp. 300–312. ISSN: 14603683. DOI: 10.1177/1354068814549345.

Slapin, Jonathan B. and Sven Oliver Proksch (2008). "A scaling model for estimating time-series party positions from Texts". In: *American Journal of Political Science* 52.3, pp. 705–722. ISSN: 00925853. DOI: 10.1111/j.1540-5907.2008.00338.x.

Soetenhorst, Bas (Sept. 2004). *Van Aartsen van crisis tot crisis ; als geert Wilders vertrekt , Verliest fractievoorzitter*. URL: https://advance-lexis-com.proxy.uba.uva.nl:2443/api/document?collection=news%7B%5C&%7Did=urn:contentItem:4D73-VNP0-0151-02S6-00000-00%7B%5C&%7Dcontext=1516831.

Spoon, Jae Jae and Heike Klüver (2020). "Responding to far right challengers: does accommodation pay off?" In: *Journal of European Public Policy* 27.2, pp. 273–291. ISSN: 14664429. DOI: 10.1080/13501763.2019.1701530.

Steffen, Tilman (Jan. 2020). *AfD-Fraktion Bundestag : Politik auf verlorenem Posten*. URL: https://www.zeit.de/politik/deutschland/2020-01/afd-fraktion-verena-hartmann-abgeordnete-parteiaustritt.

Stokmans, Derk (2008). *Who is Geert Wilders ?* URL: https://web.archive.org/web/20090305110822/http://www.nrc.nl/international/article1876586.ece.

Strøm, Kaare (1990). *Minority Government and Majority Rule*. Cambridge University Press.

Van de Wardt, Marc, Catherine E. De Vries, and Sara B. Hobolt (2014). "Exploiting the cracks: Wedge issues in multiparty competition". In: *Journal of Politics* 76.4, pp. 986–999. ISSN: 14682508. DOI: 10.1017/S0022381614000565. URL: http://www.journals.uchicago.edu/doi/10.1017/S0022381614000565.

Van der Brug, Wouter (2003). "How the LPF Fuelled Discontent: Empirical tests of explanations of LPF support". In: *Acta Politica* 38.1, pp. 89–106. ISSN: 0001-6810. DOI: 10.1057/palgrave.ap.5500005.

Van der Brug, Wouter, Meindert Fennema, and Jean Tillie (2005). "Why some anti-immigrant parties fail and others succeed a two-step model of aggregate electoral support". In: *Comparative Political Studies* 38.5, pp. 537–573. ISSN: 00104140. DOI: 10.1177/0010414004273928.

Van Holsteyn, Joop J.M. (2011). "The Dutch parliamentary election of 2010". In: *West European Politics* 34.2, pp. 412–419. ISSN: 01402382. DOI: 10.1080/01402382.2011.546590.

Van Spanje, Joost (2010). "Contagious parties: Anti-immigration parties and their impact on other parties' immigration stances in contemporary western europe". In: *Party Politics* 16.5, pp. 563–586. ISSN: 13540688. DOI: 10.1177/1354068809346002.

Vossen, Koen (2010). "Populism in the Netherlands after Fortuyn: Rita Verdonk and Geert Wilders compared". In: *Perspectives on European Politics and Society* 11.1, pp. 22–38. ISSN: 15705854. DOI: 10.1080/15705850903553521.

Vossen, Koen (2011). "Classifying Wilders: The Ideological Development of Geert Wilders and His Party for Freedom". In: *Politics* 31.3, pp. 179–189. ISSN: 02633957. DOI: 10.1111/j.1467-9256.2011.01417.x.

Wagner, Markus and Thomas M. Meyer (2017). "The Radical Right as Niche Parties? The Ideological Landscape of Party Systems in Western Europe, 1980–2014". In: *Political Studies* 65.1_suppl, pp. 84–107. ISSN: 14679248. DOI: 10.1177/0032321716639065. URL: http://journals.sagepub.com/doi/10.1177/0032321716639065.

Weinzierler, Julia (2019). "Bei Lanz : Merkel-Kritiker aus der CDU zählt AKK an - TV-Star genervt von „ purem Machterhalt "". In: *Merkur*, pp. 1–3. URL: https://www.merkur.de/politik/markus-lanz-zdf-talk-hamburg-kritik-groko-merkel-cdu-parteitag-von-stetten-kritisiert-zr-13240423.html.

Zaller, John (1992). *The nature and origins of mass opinion.*

Appendix

Appendix A

Classifier	Vectorizer	Stemmed	Accuracy	Precision	Recall
Logistic Regression	tfidf	raw	0.87	0.55	0.64
Logistic Regression	tfidf	stems	0.83	0.45	0.69
Logistic Regression	count	raw	0.84	0.46	0.53
Logistic Regression	count	stems	0.83	0.43	0.54
Naive Bayes	tfidf	raw	0.85	0.48	0.64
Naive Bayes	tfidf	stems	0.84	0.46	0.62
Naive Bayes	count	raw	0.85	0.48	0.44
Naive Bayes	count	stems	0.84	0.47	0.43
Support Vector Machine	tfidf	raw	0.88	0.58	0.55
Support Vector Machine	tfidf	stems	0.84	0.47	0.62
Support Vector Machine	count	raw	0.82	0.41	0.48
Support Vector Machine	count	stems	0.82	0.4	0.51

Table 2: Performance of different classifiers. The cose classifier is in the first row.

Appendix B1

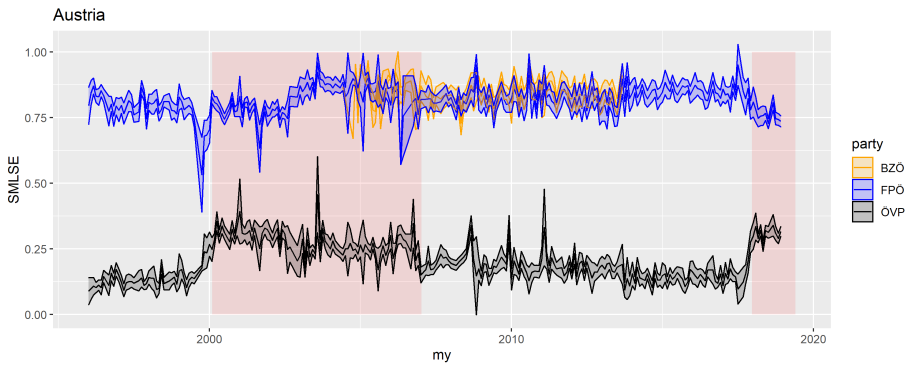

Figure 5: Estimates for Austria including BZÖ

Appendix B2

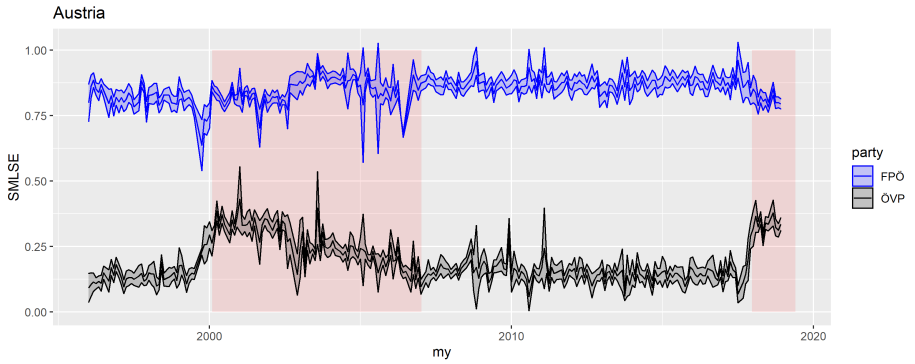

Figure 6: Estimates for Austria trained to detect FPÖ-speeches only.

Appendix C

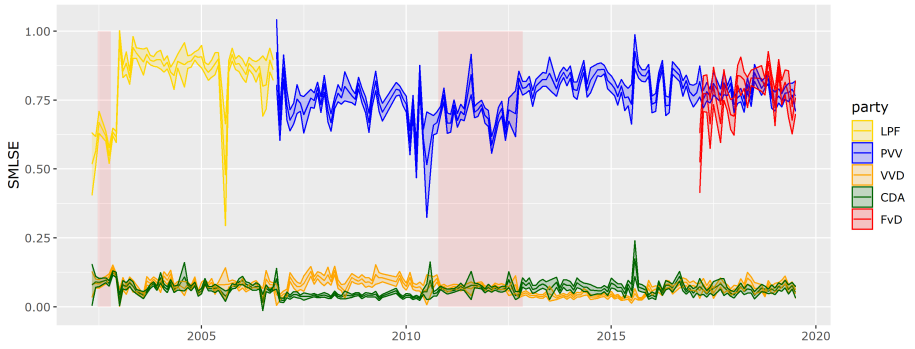

Figure 7: Estimates for the Netherlands including CDA and FvD.

Appendix D

Word	Contribution	Word	Contribution
minister	50.2	mij	8.4
kunnen	16.5	kabinet	8.2
vind	15.8	hij	7.8
Algerije	12	ministerie	7.3
ben	11.7	Het	7.1
motie	10.8	Arabië	7.1
LPF	10.7	imam	7
AIVD	9.9	terroristische	6.9
Wij	9.9	groot	6.9
zeker	9.6	Europa	6.8
mening	9.6	onderzoek	6.2
Defensie	9.5	Dutchbat	6.2
terroristen	9.4	man	6
Dat	9.3	moskeeën	5.6
Saoedi	8.6	moskee	5.5

Table 3: Thirty words with the strongest positive contribution in Wilders' speeches as a member of the VVD 2004.