Coalition Differentiation at Scale: Measuring Intra-Coalition Conflict with Supervised Text Classification

Note: This is a first draft. Please do not cite or circulate without permission.

Christine Sheldon*

October 20, 2025

Abstract

Coalition parties face a unity–differentiation dilemma: they must uphold compromise to govern effectively whilst maintaining distinct identities for electoral success. Existing measures of coalition differentiation are largely static, and the few dynamic metrics cover a limited sample of governments. I introduce a comparative, dynamic measure based on supervised classification of legislative speeches by party label. Accurate classification indicates parties defending distinct preferences; misclassification signals a united front. The measure provides monthly scores for parties in nearly 50 coalition governments across six European democracies, enabling disaggregation of coalition governance to the party level behaviour. This offers a flexible, empirically grounded tool for examining intra-coalition conflict, legislative responsiveness, and strategic behaviour, expanding the scope of coalition research beyond formation-stage proxies.

*Research Fellow in Computational Social Science, Centre for AI in Government, University of Birmingham. c.r.sheldon@bham.ac.uk.

1 Introduction

Political parties joined together in government are subject to two competing forces. They are aware that to enjoy incumbency benefits, it is imperative to maintain the coalition compromise. Yet simultaneously, once the coalition terminates, planned or otherwise, each party is accountable to the electorate as a distinct entity. This, in turn, is an incentive to publicly distinguish oneself during the incumbency instead. Boston and Bullock (2010) identify these orthogonal incentives as the unity-differentiation dilemma: parties seek 'how best to balance the desire for unity or cohesion (and hence effectiveness) with the equally important electoral imperative to maintain party distinctiveness or differentiation' (p.41). In navigating this dilemma, parties are restricted in the extent that they can differentiate. Throughout incumbency, coalition parties are strictly bound by collective cabinet responsibility (Fortunato 2021). This responsibility 'stipulates that the government present a united front on all policy matters, *particularly* after the policy has passed the parliament' (Fortunato 2021, p.93). As such, the only real opportunity parties have to differentiate without violating this responsibility, is during the legislative review period. Yet even during this time, differentiation comes with a price. Effective, united coalition governance and differentiation are mutually exclusive. Choosing to emphasise differences over the coalition compromise is therefore an inherently obstructive behaviour. As stated by Fortunato (2021), differentiation, at its core, 'is conflict. It is heated debate in the plenary, pointed questions of ministers in subpoenaed testimony in committee hearings, and bold amendments in legislative review' (p. 27). The extent to which differentiation occurs can, therefore, be interpreted as a widely, and publicly, observable indication of coalitional conflict throughout incumbency. This enables valuable insights into the otherwise opaque dynamics of coalition cooperation.

By far the most common practice for capturing how conflictual or differentiated a coalition government is, is to measure the likelihood conflict will occur during incumbency. This can be influenced by aspects of institutional design which allow ministers to submit legislative proposals closer to their preferences, rather than the coalition consensus, referred to as

ministerial drift(Martin and Vanberg 2011); The number of parties in a coalition, as an indication of bargaining complexity (Zubek 2015); Or the ideological distances between parties in their electoral manifestos signifying the extent of compromise required for cooperation (Carroll and Cox 2012; Dörrenbächer, Mastenbroek and Toshkov 2015; Klüver and Bäck 2019; Klüver and Spoon 2017, 2020; König et al. 2022; Meyer, Sieberer and Schmuck 2023). These approaches assume that the incentive, and therefore the propensity for differentiation is determined apriori and remains stable throughout the incumbency. Lupia and Strøm (2010), however, emphasise that coalition party behaviour during incumbency is dynamic. The utility of different behaviour is continuously assessed. The choice to differentiate and publicly signal division is no exception. A party's decision to act on either of the mutually exclusive incentives of the unity-differentiation dilemma is the result of a careful weigh-off between the costs and benefits of each option. Whether it makes sense to prioritise boosting electoral prospect over generating incumbency benefits depends on different variables, many of which likely to vary over time (such as polling performance (Fortunato 2019)). The prevailing option is therefore not static, or consistent throughout a government's incumbency and instead, is likely to vary throughout. This highlights a clear mismatch between popular measurement and the actual process of coalition differentiation: Static measures solely capture differentiation at the start of a coalition government, and therefore fail to unveil exactly what differentiation dynamics are at play during incumbency.

Some studies have set out to dynamically measure the degree of differentiation and conflict in a coalition. This includes capturing differences between coalition parties in parliamentary applause (Imre et al. 2022), parliamentary speech (Fortunato 2021; Martin and Vanberg 2008), and tabled amendments (Fortunato 2019, 2021). Yet either limitations in data availability, or the intensive nature of coding the measures, has restricted these efforts to small samples of governments. Considering many outcomes of interest related to coalition conflict are rare, such as terminations or cabinet reshuffles, inferences made on such small samples do not generalise well.

This paper presents a novel dynamic measure of coalition differentiation, which is easy and consistently applicable across large samples of coalition governments. Using supervised machine classification of parliamentary speeches delivered by members of coalition parties, I generate monthly differentiation scores for six Western European democracies. The classification algorithm is first trained to identify the party label of a speech, based on the words used by members of parliament (MPs) and ministers from each respective coalition party. The trained model is then used to make out of sample predictions. The resulting performance indicator of this exercise signifies how well the algorithm is able to distinguish parties based on their word use in parliament. If it does so successfully, parties are differentiating by speaking in a distinct fashion from one another, whereas in instances of poor performance when classifying speeches, coalition parties are presenting more of a united front. This classification pipeline generates monthly differentiation scores for all parties in the coalition, for 48 governments from six European countries, over the course of 20 years. This method thus generates a dynamic coalition differentiation metric for a substantive sample of governments, and can also easily be extended to any other corpus of party-labelled coalition text data.

Below, I will elaborate on past measures of coalition differentiation and their limitations, the selection of supervised classification as a suitable similarity measure for this purpose, and the classification pipeline. In turn, a number of validation tests indicate that this classification based measure is effectively capturing coalition party-based linguistic differences.

2 Measuring Coalition Differentiation Dynamically

A handful of measures attempt to gauge the extent to which coalition parties behave differently during incumbency. One of the earliest examples of measuring coalition differentiation is from Martin and Vanberg (2008). These authors code the length of legislative speeches from German and Dutch coalition governments, as an indication of differentiation: the longer one speaks, the more opportunity for differentiation. This naturally is a strong assumption

to make, and as such, other measures of differentiation consider more explicit party behaviour. Fortunato (2019), for instance, considers legislative amendments made by coalition parties to government bills to be a form of differentiation. To capture this process, he codes all amendments tabled by coalition parties to alter legislative proposals submitted by their coalition partners. Imre et al. (2022) consider different legislative behaviour in their measure of coalition mood. This measure infers the extent to which a coalition is cooperative or competitive, not through amendments, but through the frequency with which coalition parties applaud each other's speeches in parliament. Plescia and Kritzinger (2022), in turn, generate a comparative measure of coalition conflict based on reporting on the state of coalition governments in both national and international media.

What all these measures have in common, is that they are only produced for relatively small samples. Coding amendments is very time intensive, and hence Fortunato's differentiation measure is only available for three European countries. Applause, in turn, is rarely consistently codified in debate transcripts and as such Imre et al. (2022) are restricted solely to German and Austrian coalitions. And Plescia and Kritzinger (2022) stress how 'given the relatively low number of events available for each month in each country, the ICEWS data set did not allow us to take into consideration the dynamic aspect of intra-coalition conflict' (Plescia and Kritzinger 2022, p.40).

One avenue to combat these challenges of coding and data limitation is to turn to computational text analysis, which has opened the door to efficient large-scale measurement of abundant political text data. Sagarzazu and Klüver (2017), for instance, conceptualise coalition differentiation as occurring through variations in policy attention. Coalition parties distinguish themselves by addressing distinct topics from their coalition partners. In measuring this, they use a Bayesian hierarchical topic model, developed by Grimmer (2010), to classify German coalition party press releases into policy areas, and then interpret differences in policy attention devoted to each of these areas as a strategic coalition differentiation behaviour. Another example comes from Fortunato (2021), who uses Slapin and Proksch (2008)

Wordfish model to generate daily ideological positioning scores from parliamentary speech for the parties of the 2010-2015 United Kingdom (UK) coalition. The distance between these generated scores is then interpreted as the degree of differentiation.

Both topic modelling and Wordfish are forms of unsupervised machine classification. This means these methods are trained to recognise patterns, or groupings, in data without being given prior instructions about the nature of these groupings. By locating and interpreting these clusters, one can make inferences about how these are grouped in this space, and what the distance between these clusters indicates. The one key challenge with unsupervised classification is best summarised by Goet (2019), when he states that: 'the scores that these models produce, identify which speakers tend to use similar words to one another [...].These estimates may or may not prove to have anything to do with the party, or indeed to have any stable structure over different debates or sessions' (p.7). This implies that the clusters found by an unsupervised classification algorithm do not necessarily correspond with what the researcher sets out to capture. The possibility of inconsistencies between results from different applications means careful, case-by-case post-hoc validation to assess the nature of the clusters is essential. This implies that the application of unsupervised classification to measure differentiation across time and countries, could result in wildly inconsistent, or at least incomparable, results, and would be subject to time-intensive validation. It is, therefore, unsurprising that both unsupervised classification measurements introduced here are still restricted to a single country and/or government. Thus, although these computational methods are able to process vast quantities of data with relative ease, the thorough post-hoc validation to ensure it is, in fact, differentiation between parties, as opposed to some other dimension of difference, requires significant time investment.

Existing measures of differentiation thus offer only limited and inconsistent coverage across time and cases. A measure which dynamically captures differentiation and is generated for a sizable sample of coalition governments, still remains absent. In the rest of this paper I will argue that the challenges preventing such robust measurement can be overcome by

using supervised classification instead.

3 Supervised Classification to Capture Differences

The application of supervised classification algorithms to predict the party label of political text, most notably parliamentary speech, is an increasingly popular method to measure latent variables in text (Goet 2019; Høyland et al. 2014; Ishima 2024; Peterson and Spirling 2018; Wäckerle and Silva 2023; Yu, Kaufmann and Diermeier 2008). In this method, a supervised learning classifier is trained to assign a text to a political party on the basis of the words used in the speech. The trained classifier is then applied to an out of sample set of texts to predict their party label. The resulting performance of this classification exercise can be considered an indication of distinctiveness: the better a classifier is able to recognise party affiliation of a speech, the more distinct one party's speech behaviour is from the other parties in the sample. Yet, the worse it performs, the more alike speeches are, and thus the more similar parties speak. This method, therefore, overcomes the challenge to the reliability of results of unsupervised methods by imposing the dimension of difference: party label. As a result, 'these estimates are guaranteed to be driven by the "party factor", regardless of the number of topics that are addressed, or other sources of variation in word use' (Goet 2019, p.7), e.g. the differences supervised classification picks up on, will always be differences based on party groupings. Such a consistent result allows for interpretation of results across time and cases with relative ease.

Previous application of this method have predominantly been used to measure ideological polarisation, in the US congress (Yu, Kaufmann and Diermeier 2008), UK parliament (Goet 2019; Peterson and Spirling 2018), and the European Parliament (Høyland et al. 2014). For the purpose of this paper, I merge this literature on measuring ideological polarisation with theories on coalition differentiation. By applying a supervised classification algorithm to a subset of solely coalition party speeches, one can interpret the performance metric of the

classifier not as ideological polarisation, but as differentiation. More precisely, interpreting the fine-grained differences in word use between parties as ideological polarisation, relies on the assumption that all speakers in parliament are transparent about their ideological positions in their parliamentary addresses. Yet, theories on coalition differentiation stress how members of coalition parties are not always at liberty to do so, and instead are required to publicly defend a coalition consensus instead. From this perspective, one can consider the differences in speech as the extent distinct party preferences actually shine through. When a supervised classifier is solely fed speeches by government parties, its resulting performance is, therefore, more closely related to a strategic decision made to emphasise or understate differences,than of an overall degree of ideological polarisation.

The differences between parties, which supervised classification algorithms identify, are those in word use. These can be driven by a number of different factors, such as addressing different topics (Sagarzazu and Klüver 2017), discussing the same topic with different vocabularies (Goet 2019), or even explicitly choosing to (de)emphasise their party membership over coalition membership. This measure does not distinguish which source of variation is driving the results, and instead considers all of the above as meaningful and contributing to the extent of overall coalition differentiation. [1]

A core assumption of this methods is that coalition parties behave as unitary actors in the legislative arena. Empirical evidence does suggests that access to the parliamentary floor is regulated by party elites, and thus that there is a central strategy to party speeching behaviour. Proksch and Slapin (2015), for instance, find that party elites 'still have a choice over which backbenchers should engage in monitoring through debate and could employ loyalist backbenchers for the task' (p.44). This is also reflected in the findings of Bäck et al. (2019) who find that speechmaking is restricted to fewer, more high-ranked legislators as incumbency draws to a close and communication of a consistent party label becomes

[1] Another driver of differences is the gender composition of parties, with women speaking differently to men (Wäckerle and Silva 2023), or similarly regional dialects. Since the classifiers are trained on data from a full cabinet, as described below, I assume these differences remain relatively stable over time within cabinets and, as a result, should not introduce significant amounts of noise in the measure.

more important. All of which suggest that the extent a coalition party differentiates in parliamentary speech, is a strategic decision made by the party as a whole. That being said, legislative discipline is included as a scope condition in the sample selection to follow, and the validity of this unitary actor assumption is assessed Appendix C.

4 Method

Unlike traditional supervised classification approaches, the pipeline executing this measurement is not designed to optimise classifier performance. On the contrary, the metric of interest is grounded in missclassification. The objective is to find consistent error driven by linguistic similarities, as opposed to achieving the highest accuracy. The pipeline is designed to do so, and to be easily extended to other parliamentary arenas not included in this sample, as well as applied to party-labelled coalition corpora from other sources.

4.1 Data and Pre-processing

The measure is generated for a sample of ten countries: Austria, Denmark, Germany, Ireland, Netherlands, and Sweden. These countries are not only characterised by consistent coalition governance, but also have a practice of parliamentary party discipline, which is a scope condition of the sample selection. the speech data from these countries comes predominantly from ParlSpeech V2 (Rauh and Schwalbach 2020), with the exception of Ireland, for which the speech data is from the Database of parliamentary speeches in Ireland, 1919-2013 (Herzog and Mikhaylov 2017). In total, there are 48 coalition governments from the last 20 year included in the selection. Metadata on coalition membership and duration are taken from the Party Government in Europe Database (PAGED) (Bergman, Bäck and Hellström 2021).

The speech data is grouped per cabinet, and excludes speeches by the speaker of the house, those of the opposition, and any speeches of fewer than 40 words. The remaining corpus includes speeches by both legislators and members of the executive. Conflating these two

types of speakers, and whether that is a violation of the parties as unitary actors assumption is dissected in Appendix C. The corpus is vectorised using a bag-of-words approach, with removal of stop words during TF-IDF (Term Frequency-Inverse Document Frequency) vectorization using country-specific stopword dictionaries from the Natural Language Toolkit (NLTK) (Bird, Klein and Loper 2009). TF-IDF vectorising considers how important a word is to a document relative to the full cabinet corpus. This measure of importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the entire corpus, helping to downweight common words and highlight more distinctive terms. This process improves the representation of meaningful for the subsequent classification. Though simple, this bag-of-words method of text vectorisation is efficient, allows for the investigation of specific words influencing classification, and moreover, performs equally when compared to more advanced techniques like document embeddings as displayed in Appendix A.

One thing to note is the challenge of vocabulary size influencing classifier performance, as noted by Peterson and Spirling (2018) and Gentzkow, Shapiro and Taddy (2019). specifically, the size of the vocabulary of unique words found within the text for a classification (in this case one cabinet), can influence the subsequent performance metric of choice. This would add noise to the final measure, and obscure the signal of coalition-differentiating behavior. This would make it difficult to capture meaningful variation across cabinets. To account for this, I generate fixed vocabularies of unique words appearing in all the (cleaned) speech data used per country, as per Peterson and Spirling (2018). Using these vocabularies to subsequently generate the TF-IDF matrices for individual cabinets of the same country means every classification feature space has the same number of columns, enabling cross-cabinet comparison. Differences in vocabulary size and structural factors *between* countries remain, yet this can be mitigated by including country fixed effects in multi-country models.

4.2 Classification

4.2.1 Classifier Selection

I employ three distinct classifiers for the supervised classification task, each chosen for their suitability for both binary and multi-class settings, as well as their compatibility with the sparse feature matrices typical of text data. Crucially, each classifier represents a different optimization strategy for parameter estimation, which allows for a triangulation of results across fundamentally distinct modeling paradigms. This makes it possible to assess whether consistent errors are being captured across methods, reducing the risk that results are driven by model-specific characteristics.

The first classifier is Logistic Regression, implemented with a SAGA solver. Logistic regression is a linear model particularly well-suited to high-dimensional, sparse input data. The SAGA solver is efficient for large-scale problems and handles multinomial classification directly. In this application, I apply L2 regularization to prevent overfitting and to the inevitable correlated features of text data. Importantly, this model produces probabilistic outputs, which I employ to calculate the subsequent performance metrics.

The second classifier is a Linear Support Vector Classifier (LinearSVC), which is a margin-based model that optimizes a hinge-loss function. LinearSVCs often perform well on text classification tasks due to their robustness in high-dimensional feature spaces. Because they do not inherently output predicted probabilities, which are required for subsequent score calculation, I calibrate this classifier using CalibratedClassifierCV, which fits logistic models to the decision scores in a second stage. This produces comparable probabilistic outputs, while offering a learning dynamic distinct from logistic regression.

The third classifier is a Stochastic Gradient Descent (SGD) classifier with a logistic loss, meaning it optimizes the same objective function as logistic regression with L2 regularization. The distinction lies in the optimization strategy: rather than solving the problem in batch form, as with the SAGA solver, the SGD classifier updates model weights incrementally using

individual or small batches of training examples. All three classifiers are executed using the Python module Scikit Learn (Pedregosa et al. 2011).

Taken together, these three classifiers were selected to capture a broad spectrum of supervised learning strategies: probabilistic linear models, stochastic gradient–based optimization, and margin-based classifiers. Additionally, all three produce predicted probabilities, which I employ to generate the final metric. Applying all three allows for a robust test of model independence. The results in Appendix B, indeed, corroborate that there is no specific model dependency, as the results of all three classifiers converge. This indicates that they capture the same latent variable, which I interpret as coalition differentiation. The final predicted probabilities are averaged across all three classifiers to reduce model-specific noise and emphasize only those errors that are consistently detected.

4.2.2 Classification Pipeline

To obtain performance metric scores for each month of incumbency, I employ a cross-validation technique. More precisely, for each cabinet the data is randomly split into ten stratified samples, thus including speeches by all parties in the coalition. Each of the classifiers mentioned above is then trained and tested ten times, each time training the data of nine of the sections and applying the trained classifier to the remaining section to assess performance and generate predicted probabilities. As it stands, the training data is rarely balanced. This implies that the number of speeches per party will differ. To ensure the classifier is trained on balanced data, meaning it is fed roughly equal amounts of information per party, I apply Synthetic Minority Oversampling Technique (SMOTE) to each of the fold's training sets. This oversampling technique duplicates observations in the minority class, or classes, to match the number of observations in the majority class. By replicating observations based on the k-nearest neighbours of observations in the minority class, the class is expanded without adding new information to the model. As a result, all parties in

each cabinet will have the same number of speeches fed to the classifier. [2]

With this pipeline, for each of the folds, the model is trained to recognise the average party-identity of each coalition party for the duration of incumbency. This trained model is then applied to a random (yet stratified) sample of speeches from the incumbency, on which the model was not trained, to asses how these speeches diverge from this average identity. By doing this ten times using a cross validation split, each speech in the corpus is included in a test set once, and will be assigned predicted probabilities of belonging to each of the classes, or coalition parties.

4.2.3 Score Calculation

From the averaged predicted probabilities monthly party recall scores are generated by aggregating the probabilities per month for each of the parties comprising the coalition. Previous studies employing this method have typically evaluated classifier performance using overall accuracy, reflecting aggregate distinguishability across all parties in parliament. To capture individual party behaviour, I focus on recall instead, which reflects the proportion of a party's speeches that are correctly classified as such. High recall for a given party thus indicates that its rhetorical profile is sufficiently distinct for the classifier to reliably identify it, while low recall suggests lower distinguishability. In this context, false negatives (speeches from a party that are misclassified as belonging to another) are more substantively meaningful than false positives, as they directly represent moments when a party's rhetoric blends into that of its coalition partners. This party-level recall therefore provides a more nuanced measure of coalition differentiation as an individual party behaviour, allowing analysis of how clearly each party distinguishes itself over time.

Additionally, existing applications of this method have almost exclusively been applied to two-party systems. These applications, therefore, did not compare results across classifi-

[2] This does not account for structural differences in the length of speeches, i.e. if smaller coalition parties systematically speak more succinctly, they would remain underrepresented in the training data even with the same number of speeches. To account for this, I include speech length as an additional feature in the classification.

cations with different numbers of classes. Yet the number of parties, and thus the number of classes in the classification, will affect the baseline recall, or the level of performance at which a classifier is as good as random. In a binary classification this is at 50%, whereas in a four-class classification this is 25%. Comparison between coalitions of different sizes is impossible without correcting for this. I do so by only considering the surplus recall as my final measure of distinctiveness signalling. This is the rate superseding the baseline score of each month. The formula for this measure is:

$$\text{Recall Surplus} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} - \frac{1}{n}$$

Using recall surplus thus indicates how the classifiers pipeline is performing beyond randomness and regularises the results across cabinets of different sizes. Finally, since some scores are quite noisy for months with very few speeches (such as summer and winter recess), I drop the scores/months in the bottom 5% of speech frequency. I employ a relative cut-off point, as each country will differ in the number of speeches its parliamentary speech rules will allow for on average.

4.3 Interpretation of results

The differentiation scores are generated for each individual coalition party, for every month of incumbency. Recall surplus as coalition differentiation scores indicate whether a party is distinguishable, or if it frequently confused for another party. The score thus illustrates the extent a party has assimilated or united with one or more of its coalition partners, or whether it is publicly defending a unique party identity instead. What constitutes a low, or a high score, will always be relative to context. The overall performance of the classifier, and thus the size of the scores, will be influenced by linguistic characteristics such as vocabulary size,

Table 1. Dashboard Summary of Scores and Sample by Country

Country	Scores	Governments	Unique Parties	Year Range	Original Dataset
AT	447	9	3	1996–2018	ParlSpeech V2
DE	563	8	4	1991–2018	ParlSpeech V2
DK	423	9	6	1997–2018	ParlSpeech V2
IE	491	8	6	1992–2013	DPI 1919-2013
NL	638	10	6	1994–2019	ParlSpeech V2
SE	504	4	6	1991–2018	ParlSpeech V2
Total	**3066**	**48**	**31**		

but also by the distinct parliamentary rules and the culture of a country's parliament. These will be relatively stable within cases but vary across countries. For example, a country may have the parliamentary practice to address fellow politicians by their full name as opposed to directly addressing the speaker. As names will only be associated with a single party, this is a very easy distinguishable feature for a classifier to identify. As such, their coalition differentiation scores will be systematically higher than countries who do not share this practice. This is not an indication that these coalitions are on average more differentiated. Instead, this is a product of institutional customs. With this in mind, I recommend using this measure to capture within-case variation, as cross-country variation is not necessarily meaningful.

5 Results

The descriptive details of the dataset generated by the aforementioned pipeline is detailed in Table 1. Across six countries, this dataset encompasses a total of 3,066 monthly coalition party differentiation scores collected from 48 distinct government cabinets, representing 31 unique parties over time periods ranging from 1991 to 2019 depending on the country. The generation method described is straightforward to expand to any party-labeled coalition text corpus. This can include parliamentary speech from countries not covered in this dataset, or other coalition-relevant text such as press-releases or interviews.

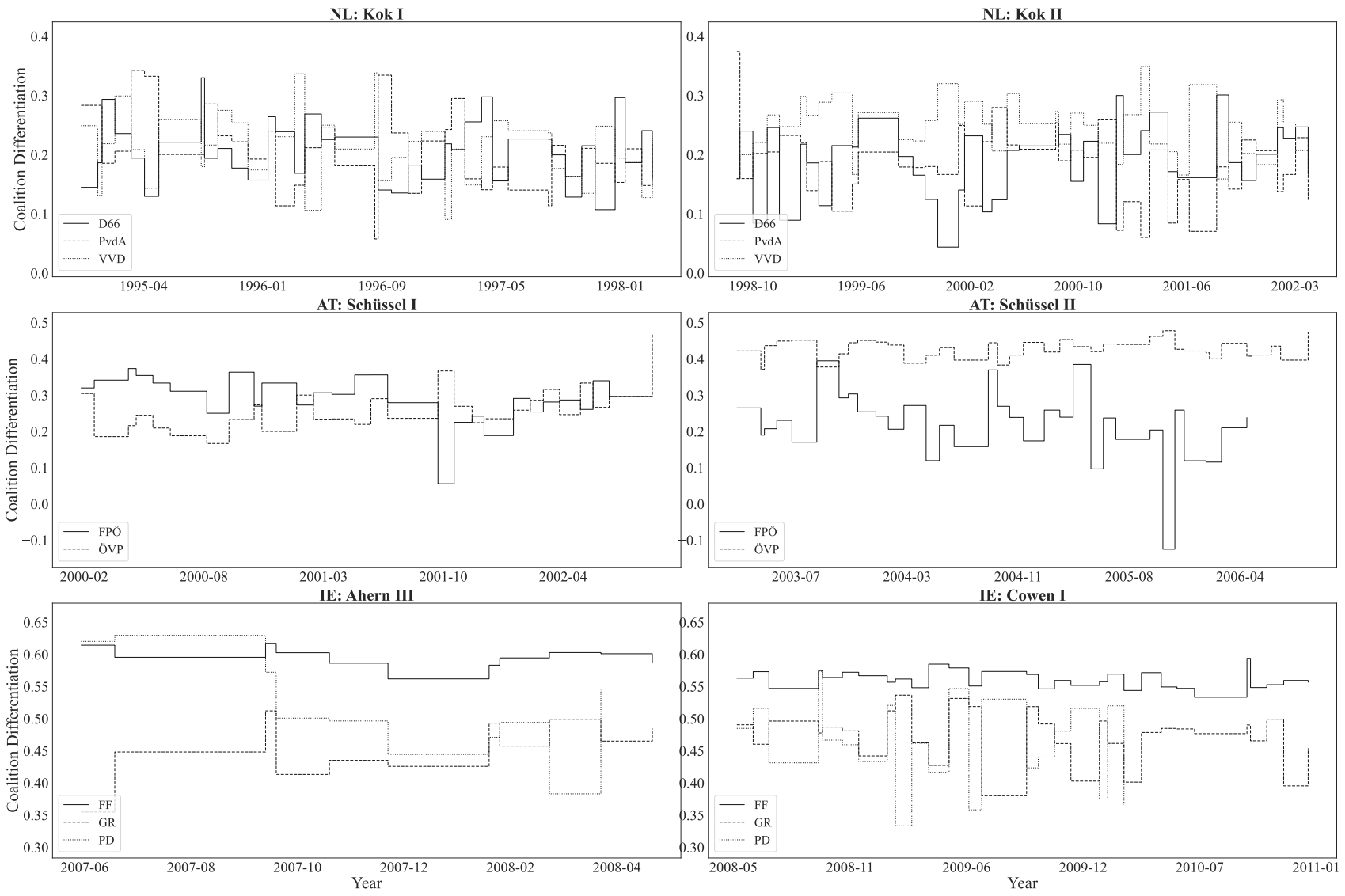

Figure 1. Coalition Differentiation for three pairs of consecutive Cabinets

The distinctive attribute of this pipeline and dataset, are the generation of dynamic, time-varying scores, for every party in cabinet. To illustrate the value of this granularity, Figure 1 plots all scores for a subsection of cabinet pairs from my sample. These pairs are successive governments from the same country, comprising the exact same parties. In this subset, therefore, the number of parties, as well as the ideological distance between the parties in office, is kept approximately constant. As mentioned previously, static cabinet characteristics such as the size and ideological distance, are popular measures indicating the state of a coalition government. Nevertheless, when these remain relatively stable for successive governments, and therefore existing measures indicate little to no change in coalition dynamic, this novel measure captures very distinct patterns of differentiation behaviour from one government to the next. For example, the Democrats'66 (D66) display a lower level of differentiation in the Kok II cabinet when compared to their behaviour in the preceding government. The same can be said about the Freedom Party of Austria (FPÖ) in the Schüssel II government. Figure 1 thus shows differentiation is an inherently dynamic behaviour across cabinets, but also within cabinets. Such fine-grained variation is not even closely captured by pre-existing static variables.

Naturally, the question remains whether this measure successfully captures coalition differentiation. The following section sets out a number of validation test to assess the extent these scores effectively capture meaningful distinctions between coalition parties over time.

6 Validation of Differentiation Scores

6.1 Feature Importance

The combination of TF-IDF vectorisation, and traditional machine learning algorithms allows for a certain degree of transparency into the classification, specifically by investigating feature weights. More precisely, each word, or feature, is ascribed a weight or importance in the classification process, every time such a pipeline is run. This means that, retrospec-

tively, one can assess which words in particular were driving classification for each cabinet, and whether this is consistent with the expectations derived from the theory on coalition differentiation. In a two-party coalition, the resulting weights indicate the importance of a word in determining the dominant class. As such the high-scoring words are associated with this one coalition party, and the low-scoring words with the other. For coalitions of more than two parties the resulting feature weights are an average of binary classifications across all combinations of classes. This means that unlike with binary classification, the feature weights cannot be interpreted as predicting a single party. As a result, assessing the importance of these features to validate these coalition differentiation scores can only be done for two-party cabinets. Figure 2 plots the top 20 words per class, for eight two-party coalitions from the sample.

From this figure one can deduce that the most important words per class tend to be words included in the party name (radikale, liberale, sozialdemokraten), or the party initialism (CDU, ÖVP, VVD). Such explicitly stating of the party name is also associated with differentiation: the more a party's name is stressed the more differentiated they are publicly. In contrast, mentioning the coalition, as German Liberal party (FDP) does frequently in Kohl VI, can be an indication of publicly uniting. Moreover, certain policy- specific words relevant to particular parties also score highly. For example. flygtninge (refugees) for the Danish radical left party (RV), arbeitnehmer (employee) for the Austrian Social Democratic party (SPÖ), and onderwijs (education) for the Dutch Christian Democrats (CDA). These results combined strongly suggest that the classification pipeline is picking up on party differences correctly and the classification is not driven by random linguistic patterns.

6.2 Convergent Validity: Static Variables

Most traditional indicators of coalition conflict are static, typically a single value assigned per cabinet. It is worth considering whether this dynamic measure of coalition differentiation, and thus indirectly of coalition conflict, corresponds with these established metrics in a way

Austria: Kern (2016-2017)

Tokens: SPÖ	Tokens: ÖVP
versuchen	daher
menschen	bauern
diskussion	geehrter
geschichte	natürlich
arbeitnehmerinnen	unternehmer
denke	glaube
heißt	endes
letztendlich	dank
überzeugt	kollege
davon	dr
bereits	familien
lohn	bürokratie
planstellen	finanzminister
verkehr	sicherheit
insofern	anlangt
frage	entsprechend
recht	wesentlich
schauen	grünen
öffentlichen	sinne
kern	ama

Austria: Klima (1997-1999)

Tokens: SPÖ	Tokens: ÖVP
sozialdemokraten	volkspartei
geschätzte	övp
daher	zweifellos
geehrte	bauern
fraktion	selbstverständlich
schon	etwa
danke	wabl
geehrten	minister
arbeitnehmer	zweitens
wichtig	verehrten
betrifft	familien
möchte	zwei
damen	unserer
arbeitnehmerinnen	ländlichen
menschen	ministerin
hoffe	brauchen
ressort	bürokratie
beschäftigten	jedenfalls
herren	frau
beschäftigung	gott

Germany: Merkel I (2005-2009)

Tokens: CDU-CSU	Tokens: SPD
unserer	spd
linken	sozialdemokraten
herzlichen	kolleginnen
deswegen	hoffe
cdu	gut
künast	verbraucherinnen
unionsfraktion	gute
verehrten	liebe
grünen	tat
union	herrn
erachtens	bildung
land	deutlich
angela	stelle
entscheidend	wissen
damen	fdp
csu	beschäftigten
zunächst	worden
kolb	walter
bundeskanzlerin	glaube
zeigt	tun

Germany: Kohl VI (1994-1998)

Tokens: CDU-CSU	Tokens: FDP
verehrten	fdp
cdu	rexrodt
csu	koalition
norbert	kolb
bundeskanzler	liberalen
peter	günter
merkel	kinkel
deshalb	heinrich
wahr	liberale
miteinander	walter
parl	auswärtigen
bereits	türkei
konkret	gegenwärtig
vergangenen	muß
wolfgang	westerwelle
verehrte	botschaft
kohl	otto
letztlich	ja
wirtschaftsstandort	richter
bayern	freien

Ireland: Kenny I (2011-2013)

Tokens: FG	Tokens: Lab
named	student
extent	supplement
protection	welfare
skills	schools
mayo	housing
informed	drugs
naturalisation	planning
disadvantaged	poverty
farmers	humanitarian
immigration	dublin
animal	determining
advised	withdrawn
conservation	vec
mail	teaching
hospitals	refuse
followed	prescribed
licensing	meath
military	statement
criminal	notified
prison	weekly

Sweden: Löfven I (2014-2018)

Tokens: MP	Tokens: S
miljöpartiet	socialdemokrater
miljöpartiets	socialdemokraterna
rödgröna	naturligtvis
naturvårdsverket	borgerliga
glad	tillväxt
yrkar	fundera
utsläppen	andersson
arbetar	trafikverket
bostäder	landsting
bistånd	jobb
såklart	arbetsmarknaden
riksdagsledamot	arbetslöshet
post	socialdemokratiska
liknande	tror
grön	konkurrenskraft
tv	dessutom
kultur	sjukvården
konsumenter	välfärd
erik	sagt
levande	både

Denmark: Nyrup Rasmussen III (1997-1998)

Tokens: RV	Tokens: S
radikale	socialdemokratiet
venstre	faktisk
radikal	socialdemokratiets
nato	således
støtte	forbindelse
pågældende	regeringen
derfor	mennesker
undervejs	bedre
fn	klart
vej	forhold
folkeafstemning	udtryk
ordførere	oplysninger
albrechtsen	arbejde
mener	konkrete
holger	optaget
altså	mere
asyl	kriminalitet
flygtninge	hensyn
synes	konservatives
søvndal	tidligere

Netherlands: Balkenende III (2006-2006)

Tokens: CDA	Tokens: VVD
cda	vvd
scholen	allerlei
debat	nederlandse
gebracht	medische
zorg	parlementaire
regeling	volgens
moment	wel
politieke	cpb
onderwijs	helemaal
zorgen	jaren
scherp	ieder
biedt	laten
leerlingen	bestuurlijke
partners	misschien
daarom	allemaal
richtlijn	laatste
hierover	schiphol
organisaties	voorzitter
voortgezet	krijgt
treden	gegevens

Figure 2. Features Predicting Class from Eight Two-Party Cabinets.

that aligns with theoretical expectations. Using metadata on coalition cabinets from the PAGED dataset (Bergman, Bäck and Hellström 2021), I average the coalition differentiation scores across parties and months to generate a single score for each cabinet. I then compare these averaged scores across several metrics of interest.

The most straightforward starting point is to examine how differentiation scores vary between cabinets that terminated due to conflict and those that did not. Figure 3 displays the distribution of differentiation scores for these two groups. Interestingly, the mean scores for the two groups are not significantly different. I attribute this to two factors. First, some cabinets may experience periods of high conflict without ultimately terminating for that reason. Second, cabinets that do end due to conflict may have undergone a sudden onset of tensions, preceded by a period of relatively stable cooperation. Because my measure is averaged over the entire cabinet's lifetime, these dynamics would not necessarily result in large differences in the overall scores. The bimodal distribution among conflicted cabinets reflects this pattern as well. Some cabinets exhibit consistently high levels of conflict, while others experience only a single bout of disagreement.

To explore the relationship between monthly coalition differentiation scores and coalition division further, I next consider ex-ante variables known to be associated with more complex bargaining environments and thus a higher propensity for conflict. Figure 4 shows the relationship between averaged differentiation scores and the duration of government formation negotiations, measured in days. The negative association indicates that as formation duration increases, cabinets tend to display less differentiation behaviour. This is consistent with theoretical expectations. A lengthy formation process often signals that substantial issues have been hashed out prior to taking office. The more settled in advance, the less opportunity there is for conflict while in government. A similar logic applies to the length of the coalition agreement. Figure 5 plots the relationship between differentiation scores and the number of words in the coalition agreement. The longer the agreement, the more policy detail is codified before governing begins, leaving less room for differentiation once in office.

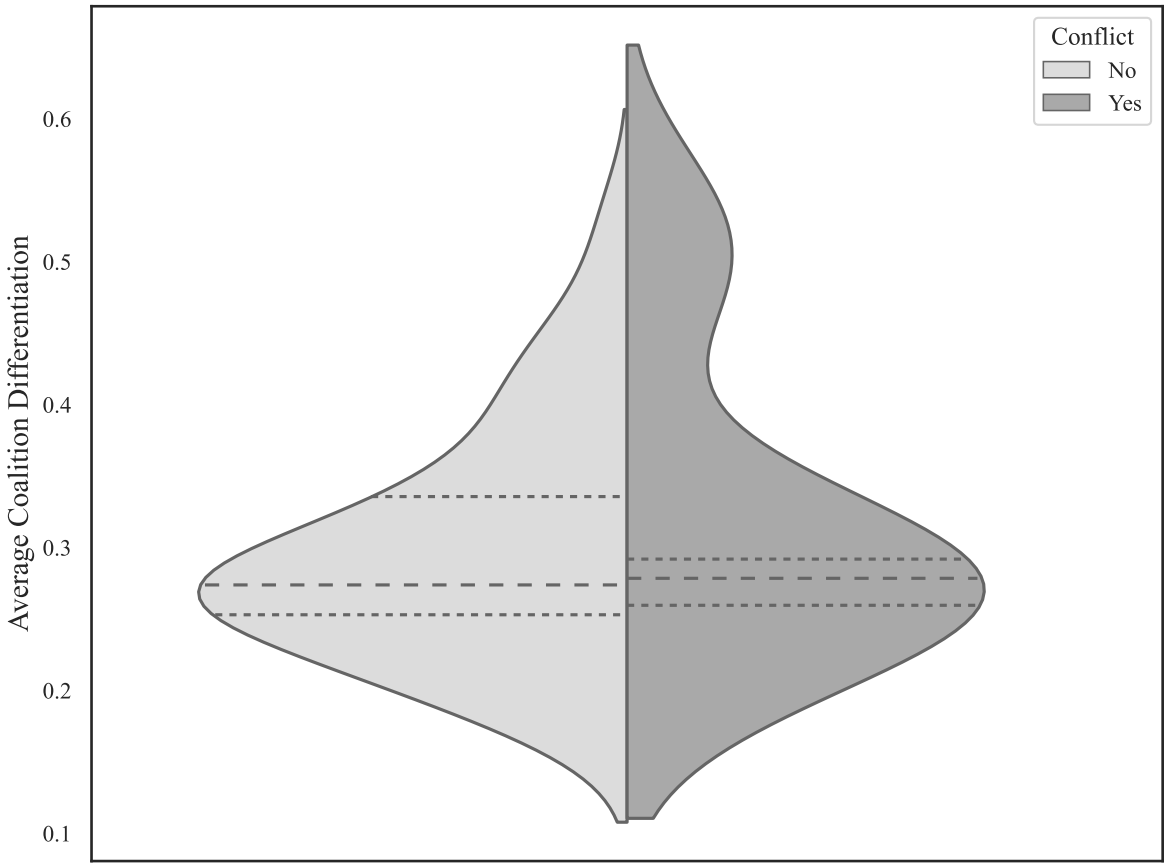

Figure 3. Average Coalition Differentiation for Cabinets Terminated by Conflict

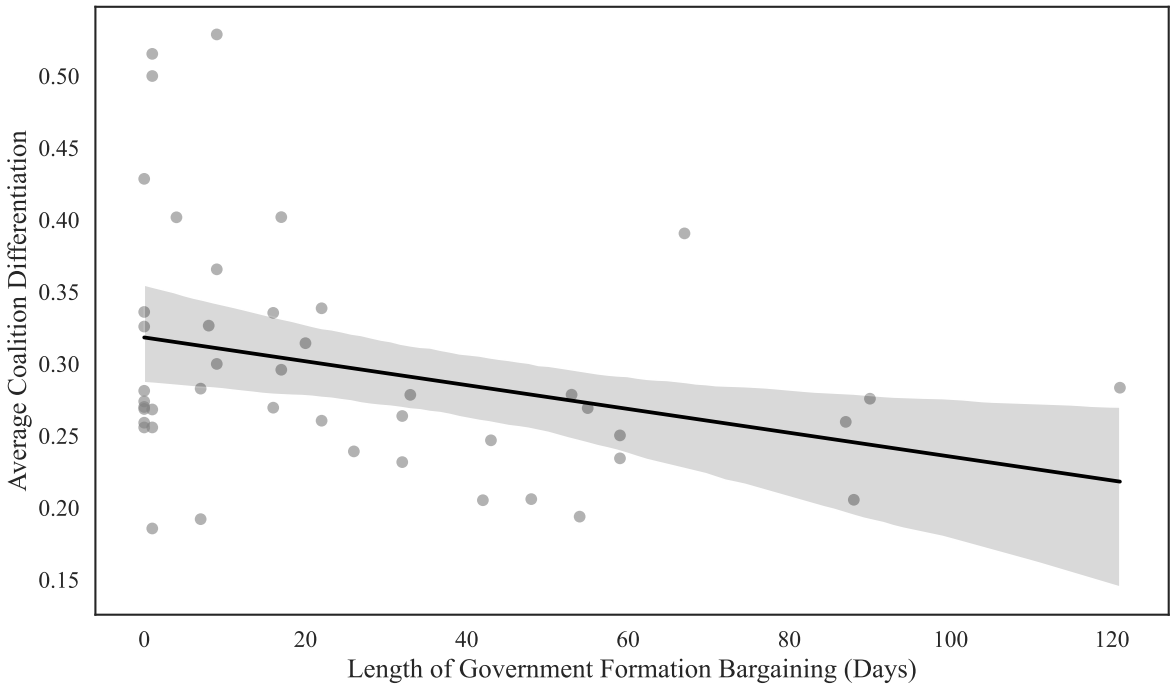

Figure 4. Average Coalition Differentiation Compared to Length of Coalition Agreement Document

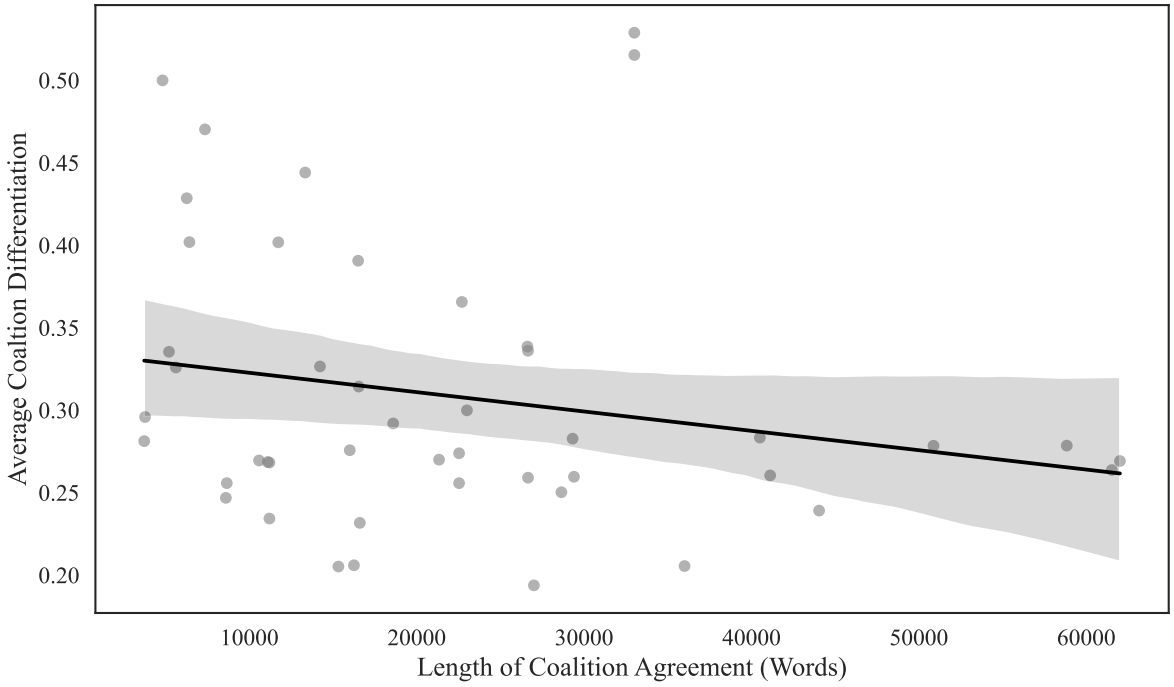

Figure 5. Average Coalition Differentiation Compared to Length of Coalition Agreement Document

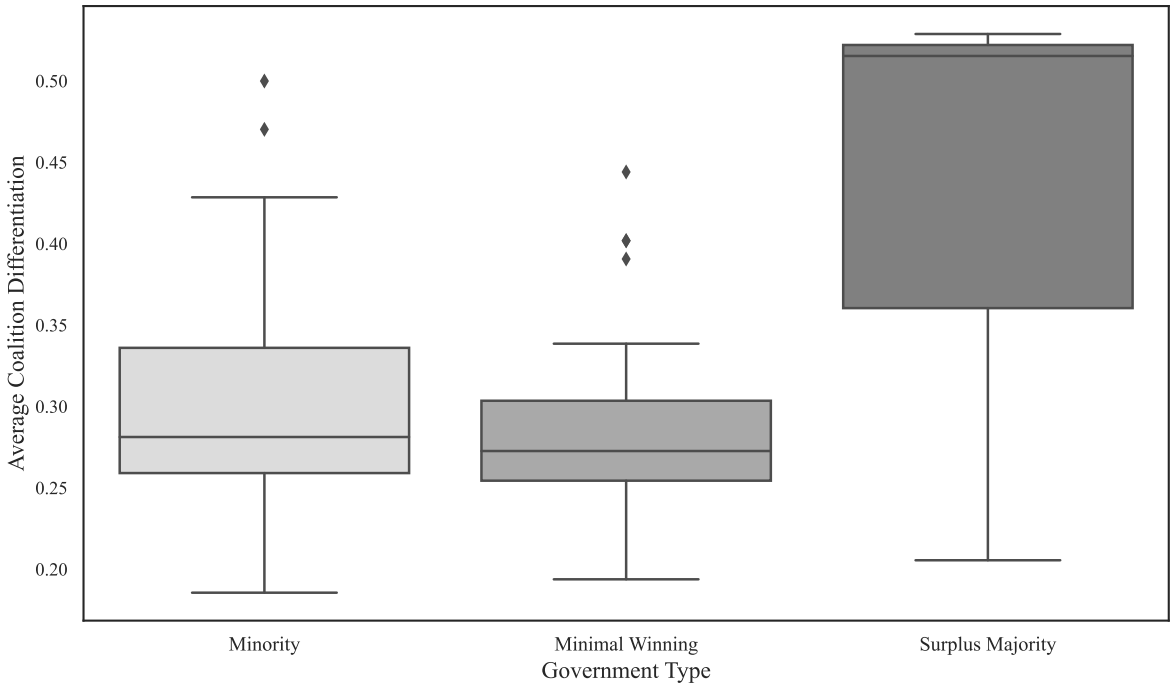

Figure 6. Average Coalition Differentiation for Different Cabinet Types

Figures 7 and 6 present average differentiation behaviour by cabinet size and cabinet type, respectively. As well established in the literature, larger coalitions face more complex bargaining environments and consequently a higher likelihood of conflict(Warwick 1992). Figure 7 reflects this pattern, with average coalition differentiation increasing alongside the number of parties in the cabinet. Cabinet type also reveals consistent differences in differentiation behaviour. Minority cabinets depend on external parliamentary partners to achieve a majority, which reduces the incentive to emphasise internal cabinet unity. Minimal winning coalitions, in contrast, rely entirely on their coalition partners to pass legislation, placing greater emphasis on internal consensus. Surplus majority cabinets, which include more parties than strictly necessary for a majority, can often form multiple majorities internally, creating more opportunities for internal division.

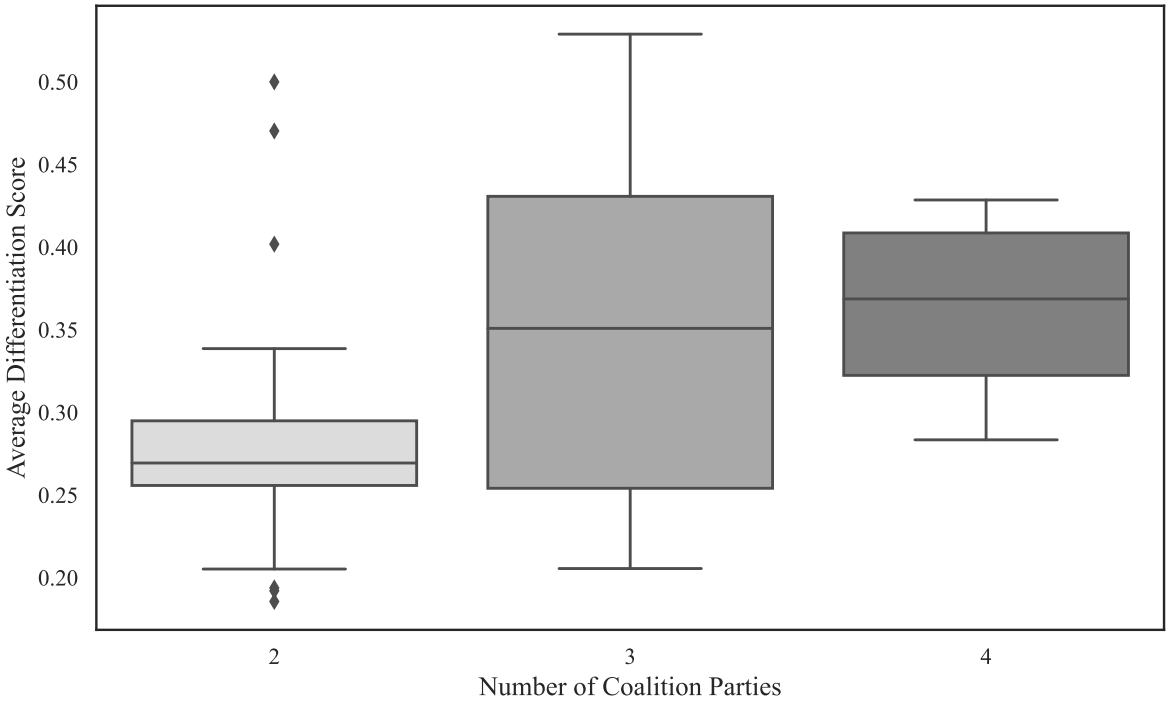

Figure 7. Average Coalition Differentiation for Different Coalition Sizes

6.3 Convergent Validity: Dynamic Variables

As mentioned prior, dynamic measures of coalition differentiation are rare, especially for sizable samples. Yet, considering the few dynamic measures which do document month-to-month fluctuations, allow me to assess whether the classification-based measure of coalition differentiation responds to temporal patterns, and not just fixed cabinet characteristics. To do so, I compare the monthly classification-based coalition differentiation scores against three such measures. The results of simple linear regression models, estimated separately for each of these comparisons and including country fixed effects where appropriate, are reported in Table 2.

The first comparison with an existing dynamic measure of coalition differentiation is to the average speech length per party per month, introduced by Martin and Vanberg (2008) as a proxy for differentiation: the longer a legislator speaks, the more opportunities there are to diverge from the coalition line. I compute monthly averages for each coalition party and correlate these with the classification-based differentiation scores in Model 1 of Table 2. It

should be noted that speech length is also included as one of the features in the supervised classification, meaning that the classification score already reflects some of the information speech length can provide about party identity. The fact that the coefficient in Model 1 is very small in substantive terms, though statistically significant and positive, suggests that speech length contributes only marginally to the overall differentiation signal, which is largely driven by other linguistic features. The significance is nonetheless noteworthy: even beyond the influence it had in the classification stage, monthly variation in speech length aligns with variation in the differentiation score, providing weak but consistent evidence of convergence.

The second comparison is with coalition mood, developed by Imre et al. (2022) from parliamentary applause data in Austria and Germany. Here, I aggregate the monthly differentiation scores to the cabinet level by taking their average, and regress these cabinet-level averages on the corresponding monthly coalition mood scores, including country fixed effects to account for unobserved heterogeneity between the two countries. The coefficient in Model 2 is significant and positive: as the mood of the coalition improves, differentiation also increases. This runs counter to the expectation that more supportive legislative applause behaviour would coincide with greater unity and less differentiation. One possible explanation is that applause may not purely reflect approval of coalition unity, but rather a broader positive response to active, engaging, or high-salience parliamentary debate. Closer inspection in Figure 8 shows that the relationship is driven primarily by Austria, suggesting that contextual or institutional factors specific to the Austrian parliamentary setting may explain this pattern. However, without further qualitative investigation, this remains speculative.

Finally, I compare my measure to the variance in issue attention across coalition party press releases, introduced by Sagarzazu and Klüver (2017) for a set of German cabinets. This captures the extent to which coalition partners prioritise different topics in their public communication. Model 3 of Table 2 shows a positive coefficient ($p = 0.07$), confirming that higher differentiation in press releases is associated with higher differentiation in parliamentary speech. This convergence is reassuring, given that the two measures are based

Table 2. Regression Results: Dynamic Measures Predicting Coalition Differentiation

	Party-Month	Cabinet-Month	Cabinet-Month
	(1)	(2)	(3)
Intercept	0.257***	0.196***	0.324***
	(0.010)	(0.015)	(0.049)
Average Speech Length	0.000**		
	(0.000)		
Coalition Mood		0.012***	
		(0.002)	
Issue Diversity (Logged)			0.023*
			(0.013)
Country Fixed Effects	✓	✓	
Number of Countries	6	2	1
Observations	3077	323	113
R^2	0.276	0.162	0.028
Adjusted R^2	0.274	0.156	0.019
Residual Std. Error	0.101 (df=3070)	0.044 (df=320)	0.046 (df=111)
F Statistic	194.745*** (df=6; 3070)	30.845*** (df=2; 320)	3.181* (df=1; 111)

Note: *p<0.1; **p<0.05; ***p<0.01
Standard errors in parentheses.

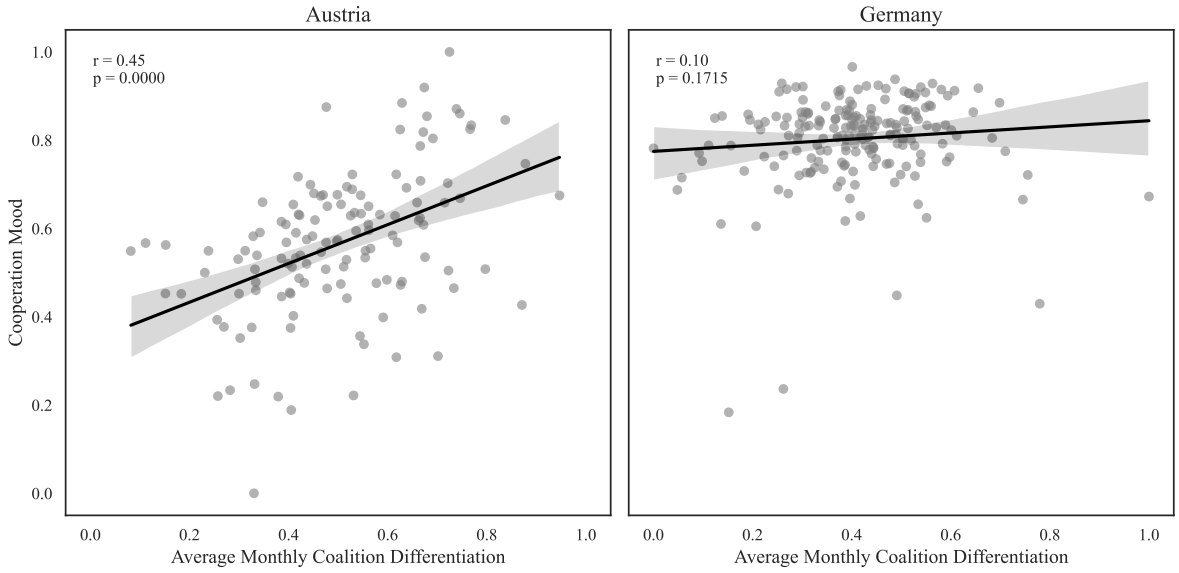

Figure 8. Average Monthly Coalition Differentiation and Coalition Mood

on entirely separate corpora (press releases versus parliamentary debate) yet show similar patterns in coalition differentiation. It is important to note that the coefficient of 0.023 is relatively small in magnitude. This is likely due to a combination of factors. First, the two measures are on different scales and constructed from entirely separate corpora. so a direct one-to-one relationship is neither expected nor realistic. Second, even theoretically, we would not anticipate a very strong relationship: coalition parties may differentiate differently in public communications versus parliamentary debate, depending on strategic considerations, audience, and timing. Therefore, the modest size of the coefficient is consistent with expectations and does not undermine the validity of the association; it merely reflects that the two measures capture overlapping, but not identical, dimensions of coalition differentiation.

Overall, the results of these comparisons provide cautious but meaningful evidence of convergent validity for the classification-based measure of coalition differentiation. Across all three dynamic measures the measure exhibits positive associations, even if some coefficients are small or contextually nuanced. The alignment with multiple, independently constructed dynamic measures suggests that the monthly differentiation scores are sensitive to temporal variation in coalition behavior, capturing shifts in party positioning beyond static charac-

teristics. While some patterns, such as the positive relationship with coalition mood in Austria, remain puzzling, the broader consistency across measures reinforces confidence that the classification-based approach effectively tracks coalition differentiation over time.

7 Conclusion

Using supervised classification, the measure presented in this paper captures the degree to which coalition parties present distinct preferences or a united front in parliament for every month a coalition is in office. With 3,066 monthly party scores covering close to 50 governments across six European democracies, this approach enables detailed, time-varying examinations of intra-coalition interactions. It allows researchers not only to observe coalition-level behavior but also to disaggregate these dynamics to the party level, providing insight into each party's role, responses, and strategic positioning within the coalition.

Because of this granularity, future research can employ this measure to assess key questions surrounding coalition governance with much more variation than previously possible. For example, the process of coalition legislating can be unpacked by assessing how a differentiated coalition affects legislative output, and under which conditions legislating becomes protracted. In turn, differentiation scores allow for the study of coalition responsiveness: is a coalition government's sensitivity to public demand a united response across all parties, or does it induce coalition conflict? In a similar vein, one can assess how voters might reward or penalise publicly made concessions. At the cabinet level, differentiation scores can serve as indicators of internal divisions that may precipitate preterm coalition termination or affect strategic decision-making.

There are, however, important limitations to consider. The performance of the supervised classification is influenced by country-specific characteristics, including the size and composition of parliamentary speech corpora and legislative procedures that affect speech quantity. As a result, the measure is not suited for direct cross-country comparison without

accounting for these factors; the inclusion of country fixed effects is recommended when analyzing multiple countries. The measure's comparative value lies primarily in cross-cabinet and intra-coalition party comparisons within countries. Additionally, the measure assumes a degree of party discipline in the legislature. In political systems with weakly institutionalized parties or less centralized legislative behavior, the measure is likely to capture more noise than signal, as indicated in Appendix C. Even in disciplined legislatures, exceptions exist, such as coalitions involving populist parties, where prolonged, public intra-party divisions may reduce measurement precision. Consequently, the measure is less reliable in assessing coalition governance under such conditions.

Overall, this classification-based approach offers a robust and flexible tool for researchers seeking to understand the temporal dynamics of coalition behavior, the role of individual parties within coalitions, and the broader implications of differentiation for legislative performance and democratic accountability. Moreover, the method is readily extendable to any other party-labeled coalition corpora, making it a versatile tool for comparative research across different countries, time periods, or legislative contexts. By bridging the gap between static and dynamic measures of coalition differentiation, it provides a foundation for richer empirical investigation into the strategic and institutional mechanisms that shape coalition governance.

References

Bergman, Torbjörn, Hanna Bäck and Johan Hellström. 2021. *Coalition Governance in Western Europe*. Oxford University Press.

Bird, Steven, Ewan Klein and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Boston, Jonathan and David Bullock. 2010. "Multi-party governance: Managing the unity-distinctiveness dilemma in executive coalitions." *Party politics* 18:349–368.

Bäck, Hanna, Markus Baumann, Marc Debus and Jochen Müller. 2019. "The Unequal Distribution of Speaking Time in Parliamentary-Party Groups." *Legislative Studies Quarterly* 44:163–193.

Carroll, Royce and Gary W. Cox. 2012. "Shadowing Ministers." *Comparative political studies* 45:220–236.

Dörrenbächer, Nora, Ellen Mastenbroek and Dimiter D. Toshkov. 2015. "National Parliaments and Transposition of EU Law: A Matter of Coalition Conflict?" *Journal of Common Market Studies* 53:1010–1026.

Fortunato, David. 2019. "Legislative Review and Party Differentiation in Coalition Governments." *American Political Science Review* 113:242–247.

Fortunato, David. 2021. *The cycle of coalition : how parties and voters interact under coalition governance*. Cambridge University Press.

Gentzkow, Matthew, Jesse M. Shapiro and Matt Taddy. 2019. "Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech." *Econometrica* 87:1307–1340.

Goet, Niels D. 2019. "Measuring Polarization with Text Analysis: Evidence from the UK House of Commons, 1811-2015." *Political Analysis* 27:518–539.

Grimmer, Justin. 2010. "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases." *Political analysis* 18:1–35.

Herzog, Alexander and Slava Mikhaylov. 2017. "Database of Parliamentary Speeches in Ireland, 1919-2013.".
URL: `https://doi.org/10.7910/DVN/6MZN76`

Høyland, Bjørn, Jean-François Godbout, Emanuele Lapponi and Erik Velldal. 2014. "Predicting Party Affiliations from European Parliament Debates.".
URL: `http://emanuel.at.ifi.uio.`

Imre, Michael, Alejandro Ecker, Thomas M. Meyer and Wolfgang C. Müller. 2022. "Coalition Mood in European Parliamentary Democracies." *British Journal of Political Science* pp. 1–18.

Ishima, Hideo. 2024. "Talking Like Opposition Parties? Electoral Proximity and Language Styles Employed by Coalition Partners in a Mixed Member Majoritarian System." *Legislative studies quarterly* 49(3):721–740.

Klüver, Heike and Hanna Bäck. 2019. "Coalition Agreements, Issue Attention, and Cabinet Governance." *Comparative Political Studies* 52:1995–2031.

Klüver, Heike and Jae-Jae Spoon. 2017. "Challenges to multiparty governments: How governing in coalitions affects coalition parties' responsiveness to voters.".

Klüver, Heike and Jae-Jae Spoon. 2020. "Helping or Hurting? How Governing as a Junior Coalition Partner Influences Electoral Outcomes." *The Journal of politics* 82:1231–1242.

König, Thomas, Nick Lin, Xiao Lu, Thiago N. Silva, Nikoleta Yordanova and Galina Zudenkova. 2022. "Agenda Control and Timing of Bill Initiation: A Temporal Perspective on Coalition Governance in Parliamentary Democracies." *American Political Science Review* 116:231–248.

Lupia, Arthur and Kaare Strøm. 2010. *Bargaining, Transaction Costs, and Coalition Governance*. Oxford University Press pp. 51–82.

Martin, Lanny W. and Georg Vanberg. 2008. "Coalition government and political communication." *Political Research Quarterly* 61:502–516.

Martin, Lanny W and Georg Vanberg. 2011. *Parliaments and coalitions : the role of legislative institutions in multiparty governance*. Oxford University Press.

Meyer, Thomas M., Ulrich Sieberer and David Schmuck. 2023. "Rebuilding the coalition ship at sea: how uncertainty and complexity drive the reform of portfolio design in coalition cabinets." *West European politics* ahead-of-print:1–22.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay. 2011. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research* 12:2825–2830.

Peterson, Andrew and Arthur Spirling. 2018. "Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems." *Political analysis* 26:120–128.

Plescia, Carolina and Sylvia Kritzinger. 2022. "When Marriage Gets Hard: Intra-Coalition Conflict and Electoral Accountability." *Comparative Political Studies* 55:32–59.

Proksch, Sven-Oliver and Jonathan B. Slapin. 2015. *The politics of parliamentary debate : parties, rebels and representation*. Cambridge University Press.

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.".
URL: `https://doi.org/10.7910/DVN/L4OAKN`

Sagarzazu, Iñaki and Heike Klüver. 2017. "Coalition Governments and Party Competition: Political Communication Strategies of Coalition Parties." *Political Science Research and Methods* 5:333–349.

Slapin, Jonathan B. and Sven-Oliver Proksch. 2008. "A Scaling Model for Estimating Time-Series Party Positions from Texts." *American journal of political science* 52:705–722.

Warwick, Paul. 1992. "Ideological Diversity and Government Survival in Western European Parliamentary Democracies." *Comparative political studies* 25:332–361.

Wäckerle, Jens and Bruno Castanho Silva. 2023. "Distinctive Voices: Political Speech, Rhetoric, and the Substantive Representation of Women in European Parliaments." *Legislative studies quarterly* 48:797–831.

Yu, Bei, Stefan Kaufmann and Daniel Diermeier. 2008. "Classifying Party Affiliation from Political Speech." *Journal of information technology politics* 5:33–48.

Zubek, Radoslaw. 2015. "Coalition Government and Committee Power." *West European Politics* 38:1020–1041.

Appendix

Contents

A Numerical Representation of Text

With a range of different methods available for creating document-feature matrices from text data, this appendix considers two different approaches to numerically represent coalition speech data for a subsample of cabinets from The Netherlands. Based on this comparison, I determine which numerical representation of text is most suitable for the generation of coalition party differentiation scores.

The most commonly employed methods for vectorising text are one hot-encoding, or bag-of-words approaches. The methods represent text data as word frequencies. This implies that one counts the frequency each word found in the corpus occurs in a document. In this particular application this would be the number of times a word from the set vocabulary appears in a coalition MP's speech, for all the coalition speeches performed during a cabinet's time in office. Since not all words are equally informative in this context, Term Frequency – Inverse Document Frequency (TF-IDF) weighting is often included to introduce a hierarchy amongst words in the corpus. TF-IDF weighting scales the absolute word counts by the occurrence of each word in the full corpus. Therefore, less frequent, but presumably more information-rich words, are weighed as more important. Whereas more frequently used filler are given less weight.

Yet, this bag of word approach only consider word frequency as a linguistic characteristic worth capturing. However, two speeches with no words in common and thus diametrically different as captured by one-hot encoding, might still be discussing the same topic or opinion. The second method employed, document embeddings, is a form of numerically representing text data which is more sensitive to semantics. Document embeddings are a distributed representation of words. This implies that instead of a vector of frequency, a document is

represented by a dense vector of a set number of averaged dimensions. Document embeddings average the embeddings for all the words found in a speech. These word embeddings represent the location of a word in a multidimensional space and thus indicate how similar words are to each other.[3] As a result, this method, therefore, can pick up semantic relationships between words, unlike one-hot encoding.

I use OpenAI's text-embedding-3-small model to turn each Dutch parliamentary speech into a single numerical representation with 1,536 values. The OpenAI embeddings model first tokenises the text, then processes them using a transformer-based neural network. It creates a set of numbers for each token, and finally combines these into one fixed-length vector for the whole speech by taking a weighted average. In the resulting matrix generated by this document embedding approach, the rows continue to represent all coalition speeches in a cabinet, yet the columns represent the 1,536 embedding dimensions instead. As such, using this method, one cannot investigate how precise words are driving classification. This degree of opacity into classification is a notable shortcoming when employing this method.

Using both the TF-IDF and embedding-based document-feature matrices for all Dutch cabinets, I apply the analytical pipeline outlined in the methods section to produce two distinct sets of coalition differentiation scores. Figure A.1 presents these scores in a scatterplot, revealing a strong positive association between the two measures. The Pearson correlation coefficient is 0.724 ($p < 0.001$), indicating that the scores derived from the two vectorisation techniques are highly similar.

Given this high degree of correspondence, it is unlikely that the additional investment required to generate text embeddings would be justified in this context. Producing embeddings involves processing the full corpus through a large neural network model, which substantially increases computation time compared to a TF-IDF transformation. It also requires greater memory and storage capacity to handle high-dimensional vectors, and, when using commercial APIs such as OpenAI's, it incurs direct financial costs proportional to the volume of text processed. Since the resulting coalition differentiation scores are not meaningfully different from those obtained with the far less resource-intensive TF-IDF method, the marginal gains in representational sophistication offered by embeddings do not translate into a measurable improvement for this specific application.

[3]Note that word embeddings therefore do not rely on the fixed vocabulary, as the chosen execution of the previous method does.

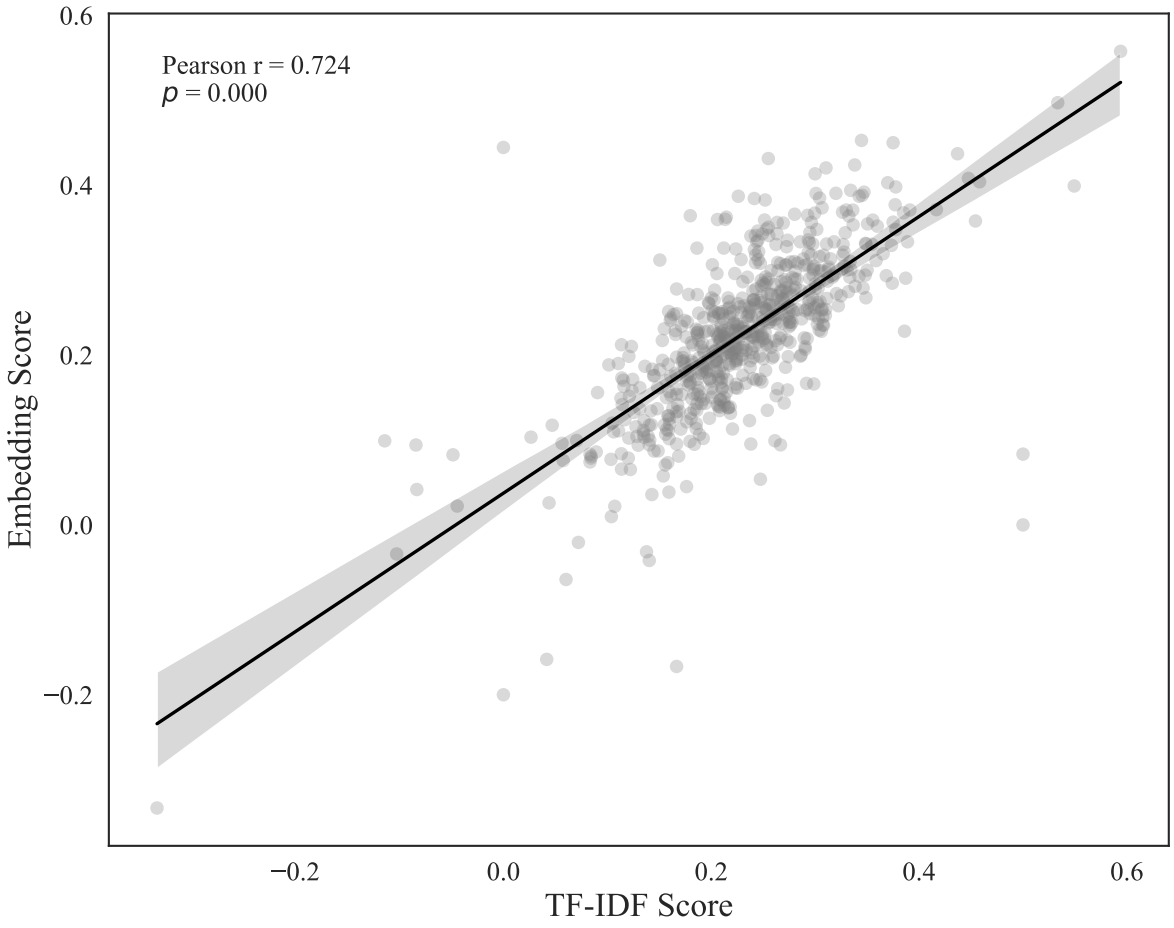

Figure A.1. Dutch Coalition Party Differentiation Scores Generated by TF-DF and Embedding Vectorization

B Classifier Performance & Model Dependency

As part of the classification pipeline, I apply three different classification algorithms to the same set of coalition speech data. These algorithms differ in their mathematical foundations, in the way they learn from data, and in how they make predictions. The differentiation scores that emerge from this step of the analysis are, at their core, a quantification of prediction error. Higher scores indicate that the classifier found it more difficult to correctly assign speeches to the correct coalition party, which in turn is taken as evidence of greater coalition differentiation.

Because these scores are derived from prediction errors, it is important to ensure that they are not simply an artefact of a specific algorithm's tendencies or limitations. If a single classifier were used, there would always be the possibility that its unique biases or assumptions were shaping the results. However, when multiple classifiers, each operating according to different principles, produce highly similar errors on the same data, this suggests that the patterns in the scores are being driven by features inherent to the data itself rather than by idiosyncrasies of the model. In this case, the use of three distinct classifiers allows for such a cross-check. If the classifiers converge on similar patterns, one can have greater confidence that the observed variation reflects a genuine latent characteristic of the speeches, rather than an artefact of model dependency.

It is therefore important to establish whether the different classifiers are in fact measuring the same underlying variation. To that end, Figure A.2 presents the coalition differentiation scores for each of the three classifiers separately across all countries included in the sample. Each point in the figure represents the average monthly cabinet score for a given country, with results shown for all six countries under study.

In every country, the resulting scores for each classifier follow a very similar trajectory. This high degree of correspondence strongly suggests that the classifiers are capturing the same latent variable, which I interpret as coalition differentiation. The implication of this finding is twofold. First, it confirms that the differentiation measure is not being driven by the particular characteristics of any single classifier. Second, it indicates that the latent signal is sufficiently strong and consistent that it can be reliably detected by algorithms with very different underlying structures. This provides a strong basis for confidence in the robustness of the measure, and removes the concern that the results might be unduly influenced by model dependency.

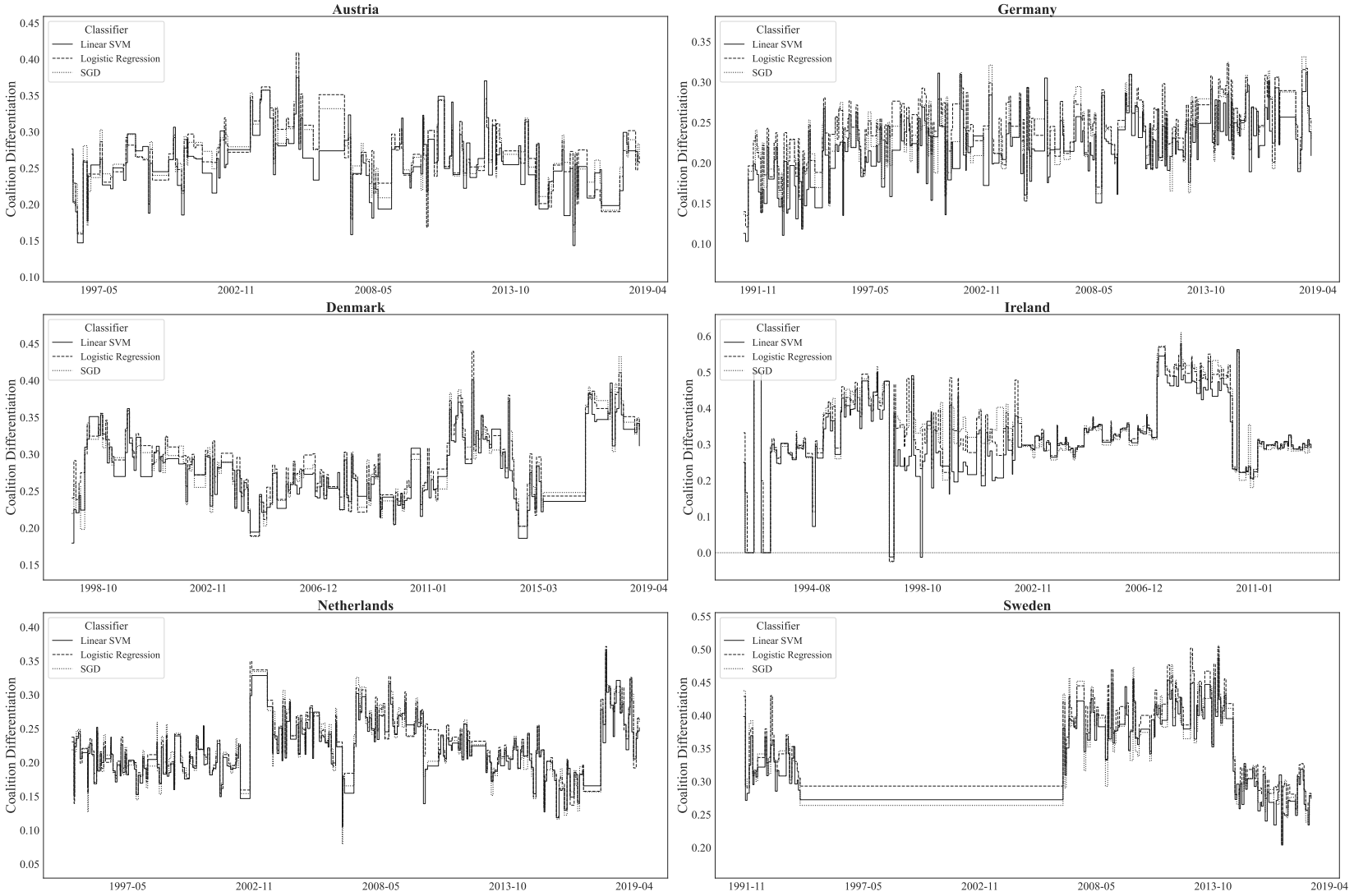

Figure A.2. Monthly Coalition Differentiation per Classifier for the Full Sample

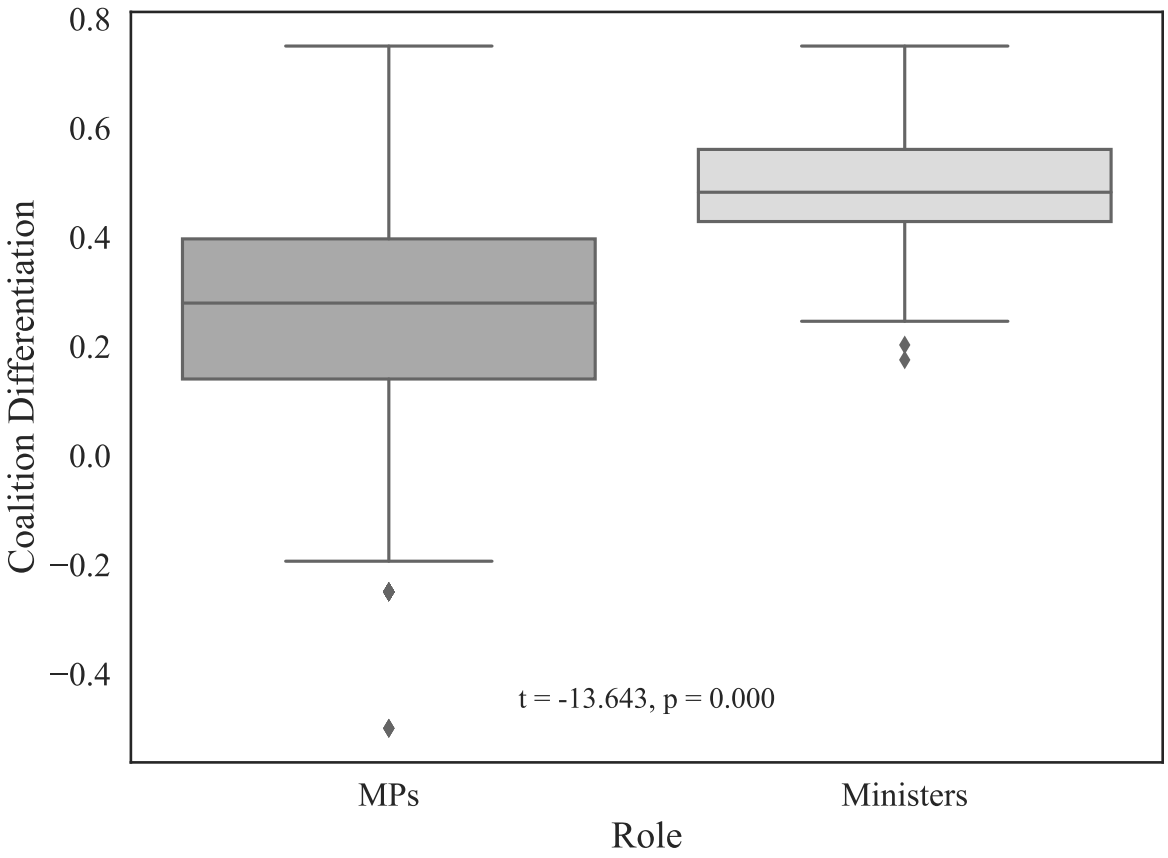

Figure A.3. Individual Speaker Coalition Differentiation for Ministers and MPs from Four Swedish Coalitions

C Coalition Parties as Unitary Actors

To test whether internal party division affects the differentiation scores, I consider scores of MPs and ministers separately. Metadata on roles of parliamentary speakers is only available for Sweden, for a total of four. For this sample, I calculate differentiation scores for each speaker, per cabinet, in order to investigate the difference in aggregated behaviour between parliament and cabinet. If MPs are indeed less bound by consensus than their colleagues in cabinet, one would expect MPs to have, on average, a higher score than ministers do. Yet when comparing these two groups in Figure A.3, the opposite is true. These boxplots show that ministers score higher on average, and that the spread of the minister scores is significantly smaller when compared to those of the MPs. Subjected to a t-test, these differences are statistically significant with a t-score of -13.643 and p-value of 0.000. This suggests cabinet ministers tend to differentiate more, and do so in a consistent fashion.

Nevertheless, this higher score for ministers may simply be driven by their frequency of speech. More precisely, if ministers speak more often during their incumbency, the classifier

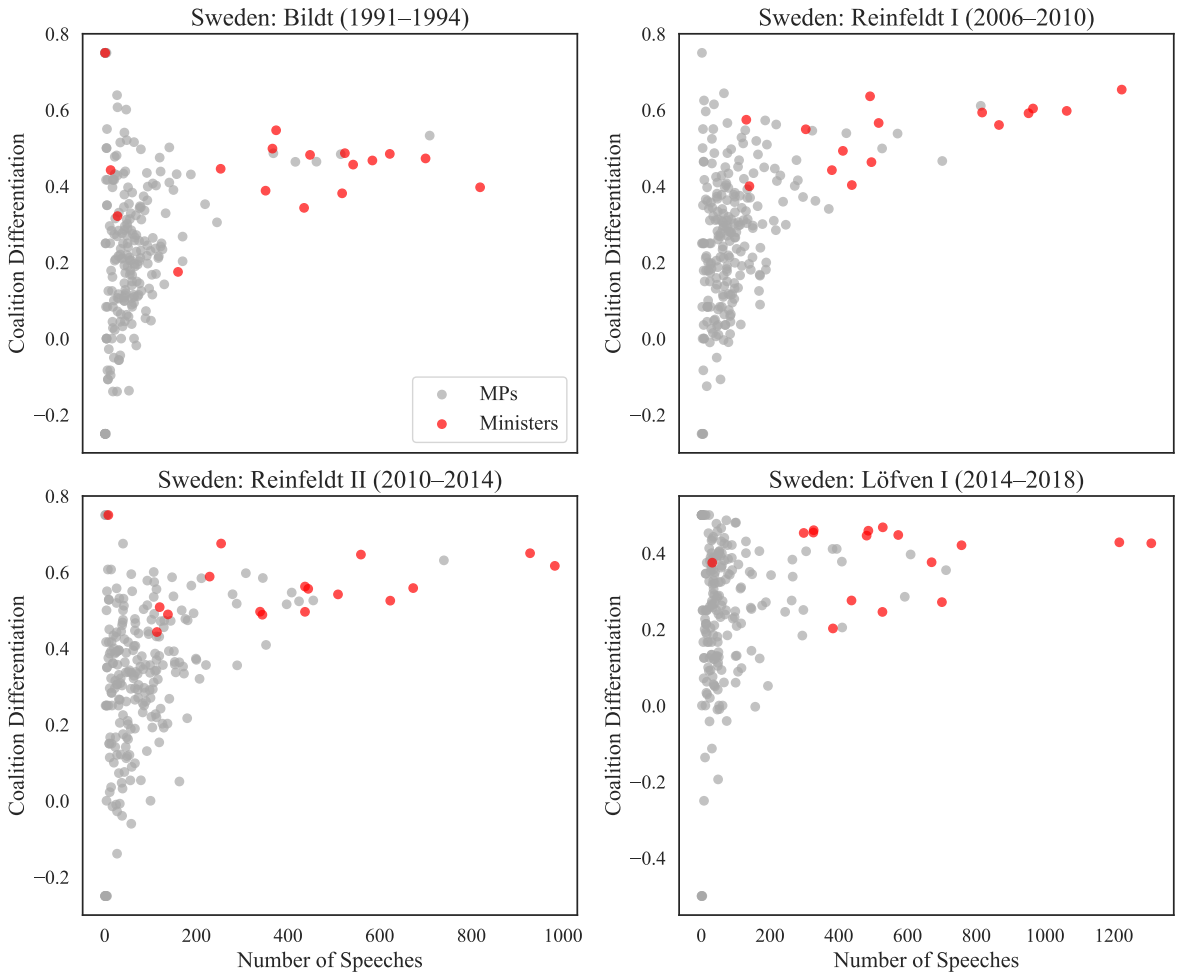

Figure A.4. Individual Speaker Coalition Differentiation for Ministers and MPs From Four Swedish Coalitions Compared Against Frequency of Speaking

will be better attuned to their language use and therefore classify them more accurately. To investigate this possibility, I plot these same scores across the number of speeches each speaker performs for the duration of the cabinet in Figure A.4. This shows that, indeed, cabinet members speak more often. Nevertheless, even those ministers who do not, still classify as high. This indicates that there is consistency across ministers of the same party in their language use, further enforcing the small spread seen in the boxplots in Figure A.3.

Does the observed difference in differentiation scores among MPs challenge the assumption that parties act as unitary actors? I argue that it does not necessarily. This is because the majority of the data used in the pipeline is produced by frequent speakers, who are typically members of the party elite. Consequently, the classifier is primarily trained on the language patterns of these elite speakers when identifying party labels. These frequent speakers exhibit a high degree of linguistic consistency, regardless of whether they hold min-

isterial roles or not, which results in relatively high differentiation scores. In contrast, less frequent speakers—who are more likely to display divergence in their speech—contribute far fewer observations. As a result, their influence on the overall classification is minimal and introduces only limited noise. Therefore, the differentiation scores predominantly reflect coherent party-level language use, supporting rather than violating the unitary actor assumption.

It remains possible, however, that divisions in language use emerge amongst the frequent speakers of a party. More precisely, this would occur if there were a complete separation in differentiation behaviour within the leadership of a party. In the case of a divided party leadership, the set of speeches the classifier is trained upon is much more heterogeneous for that respective party. Therefore, the classifier's understanding of what a party identity is, becomes more diffuse. For instance, consider a latent dimension of language use on which one can place a frequency distribution of speeches per party, as plotted in Figure A.5, plot A.[4] Here the mean value, indicated by a dashed vertical line, indicates where a party identity expressed in parliamentary speech lies on the dimension in that point in time, and the standard deviation indicates how diffuse this identity is within the party itself. The intention of the measure is to capture change in means as coalition parties choose to unite or differentiate in language, and indicated in plot B where party B's mean changes from six to four, thereby moving closer to party A. Here the shaded area where the two distributions overlap represents the linguistic similarities driving misclassification and lowering classification performance scores.

However, if frequent speakers become divided, and thereby the data for the party becomes more heterogeneous, the spread of the distribution is more affected than the mean. This is depicted in plot C of Figure A.5. In such situations, any increase in the surface area where the two distributions overlap is not due to a central party decision to differentiate, and instead more of a reflection of an internal party conflict on the extent of differentiation.

In general, such division amongst frequent speakers is very rare in the sample this measure is currently applied to. The sample of countries is selected as they are characterised by party unity in parliament. The few cabinets in my sample where this may be an issue are those coalitions which include a populist party without governing experience, such as the FPÖ in the Austrian Schüssel I government and the LPF in the Dutch Balkenende I government. In these outliers, party discipline is less prevalent, specifically with regards to the extent to which the party ought to govern with consensus or continue their more extreme position from the electoral campaign. Therefore, in such cases, interpreting the differentiation scores should be done with caution, as internal party divisions may influence the results,

[4] I do not expect the spread to follow a normal distribution, this is merely for the sake of simplicity.

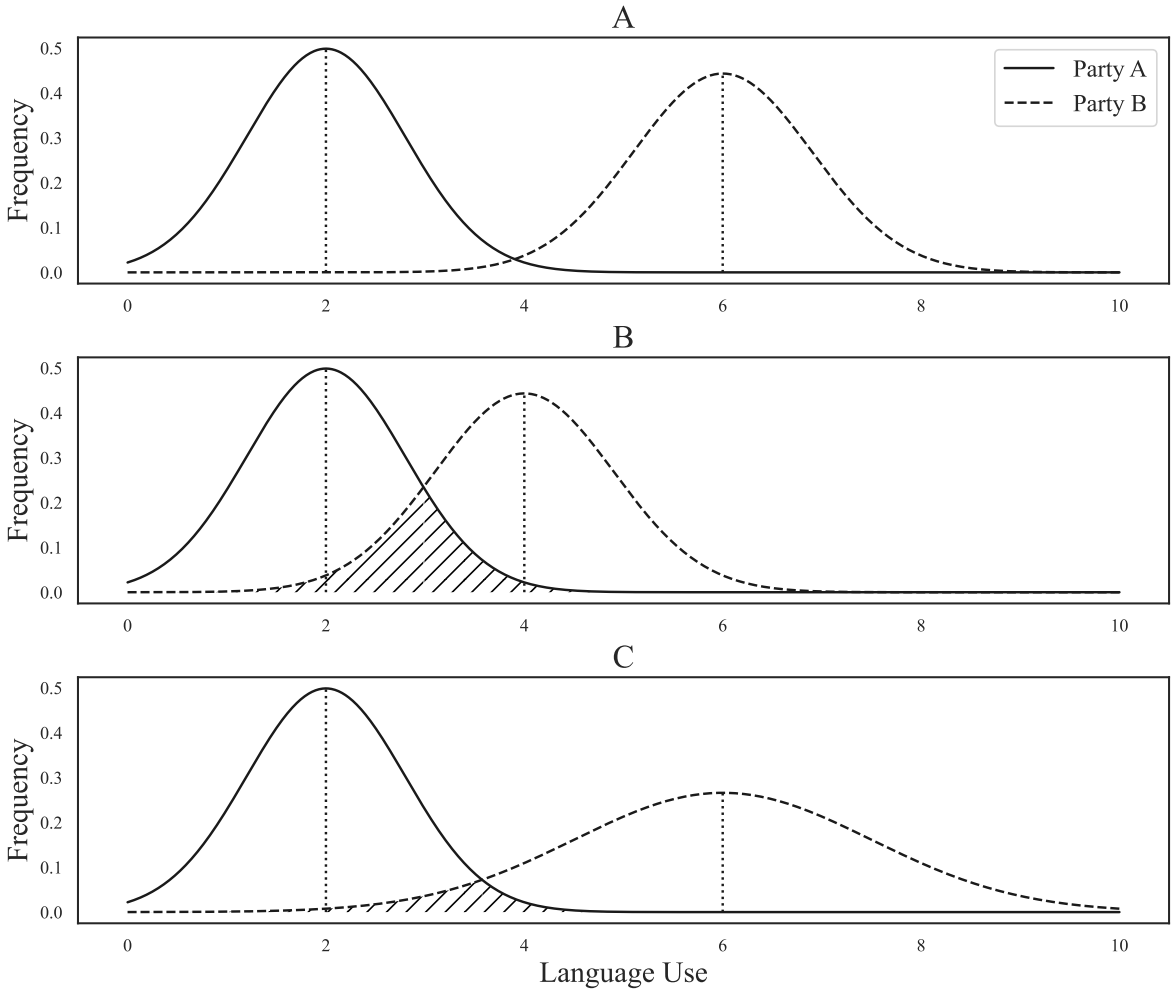

Figure A.5. Linguistic Distributions for a Hypothetical Cabinet