Multi-party Classification of Parliamentary Debates: intra-party cohesion and Inter-Party Relations Measured in Text

Martin Søyland[*]

May 5, 2020

Abstract

Recently, quantitative studies have started to utilize at the natural language content in parliamentary debates as a source of data, arguing that party classification and misclassification can be used as measures of substantive interest. With polarization, for example, the intuition is that better performing classifiers indicate more polarization between parties, and worse performing classifiers indicate less polarization. In this paper, I utilize parliamentary speeches in the multi-party setting of the Norwegian legislature to show how pre- and post-processing choices for classification task can yield widely different results. First, I find that including linguistic features and meta-data not only matters for classifying parties correctly (intra-party cohesiveness) but also for which parties are more associated with each other (inter-party relations). Second, I show that using pure classification results can lead to differences in interpretation compared to using the underlying probabilities for the classifier. Consequently, I argue that pre-processing sensitivity should be regarded as an important part of validation when using classification for explaining political behavior and change. In sum, there is little reason to argue that classifiers exclusively show traces of party cohesion or position in the Norwegian multi-party context; party cohesion is sensitive to including contextual meta-data and parties normally perceived to be closer together in the ideological space are not systematically mistaken for each other more often than parties further apart.

[*] University of Oslo, Department of Poltical Science, e-mail: m.g.soyland@stv.uio.no

Measuring ideology has an important standing in political science because these measures help us describe political change and explain the behavior of political actors across political institutions. For this purpose, studies constructing party policy position measures often rely on the use of data such as voter surveys, expert coding, roll call votes, and manifestos. The recent increase in availability of large text corpora has spawned several studies generating such measures of interest based on parliamentary speech. Most relevant here, text-based automatic party classification performance has recently been used as substantive measures explaining parliamentary behavior or political trends (Peterson and Spirling 2017; Gentzkow et al. 2016; Goet 2019). For example, Peterson and Spirling (2017) and Goet (2019) use classification of speech as a measure of polarization. The underlying assumption is that better classifier performance indicates more polarization and worse classifier performance indicates lower polarization. Classification does offer a promising approach to such tasks as it arguably side-steps the problem of dimensionality (see Hix et al. (2007)). That is, the focus is on distinguishing parties from each other, not to give a numerical representation of party position in a spatial map. Further, Peterson and Spirling (2017) show that even computationally less demanding models of language can give a reasonable picture of longitudinal trends of political polarization between parties. What has not yet been explored, however, is to take a step back and investigate to what extent language preprocessing has an effect on the results from these efforts.

In this paper, I test the effect of preprocessing decisions – including language and metadata features – on the effectiveness and substantial sensibility of party label classification in the Norwegian multi-party setting. By distinguishing between intra-party cohesion (accuracy) and inter-party relations (inaccuracy), I show how informing the classifier with more language features and contextual meta data significantly alters the results both in terms of intra-party cohesion and inter-party relations. The Norwegian case selection serves a distinct purposes: Norway is a multi-party system with more than two parties consistently occupy a significant amount of seats in parliament. This puts the classification task to a harder test than in a two-party analysis where all misclassifications only can travel from one party to the other party – making accuracy (correct classification) and misclassification be opposite sides of the same coin. Because misclassifications can travel to multiple parties in the multi-party context, we can more easily investigate whether there is information about policy positions or closeness between parties in the classification experiments; something implicitly assumed when using classification as a substantial measure. Further, the paper builds on the richly annotated Talk of Norway (ToN) dataset of parliamentary speeches, covering debates from the Norwegian parliament (*Stortinget*) in the period from 1998-2016. These data enables testing both the effect of language features, such as lemmatization and parts-of-speech, and contextual data inclusion on the accuracy and substantial sensibility of classification of party labels.

My main finding is that the classifier is sensitive to preprocessing decisions. First, intra-cohesiveness differs between preprocessing feature sets, but only in that the overall accuracy is somewhat lower in less informed models; the relative differences between parties are similar across models. Second, inter-party relations are more sensitive to including contextual meta data, but not language features. Throughout the paper, I also show that utilizing the underlying probabilities of classifiers, instead of categorized predictions, can help paint a more nuanced picture of how successful the model is at capturing the concepts of intra-party cohesiveness and inter-party relations. To further highlight the practical consequences of my findings, I proceed by replicating a model from Søyland and Høyland (2020) looking at who gets to speak in parliament and include classification precision as an additional variable. The results confirm that preprocessing decisions can have large impacts on the effect of the measure and thus also subsequent inference made from these results.

Thus, my contribution has important implications: first, and most importantly, the analyses suggest that we should be careful in weighting computational time, model simplicity, and reproducibility too heavily against predictive power; including the context of debates in addition to language features can have consequences for subsequent inference. Second, using the densities of classification probabilities can reveal problems that is impossible to trace if we only analyze categorized predictions based on the highest probabilities. In sum, my results show that choices made in the preprocessing stage of an analysis could lead us to make biased inferences down the line.

Classifying parties

The goal of correctly classifying parties based on various data sources has a long standing tradition in political science. Here, I focus on the studies using speech in legislatures. Naturally, this literature is most developed on majoritarian electoral systems, such as the US (Yu et al. 2008; Diermeier et al. 2011), Canada (Hirst et al. 2010), and UK (Peterson and Spirling 2017; Goet 2019). Overall, the studies on the majoritarian system deem pretty high classification performance. This is not necessarily surprising because there are only two classes to predict in most of these experiments. In contrast, multi-party systems can have plenty more classes, and therefore also plenty more room for classifying the wrong party for a given speech. But, Høyland et al. (2014) show that predicting party labels from speeches in the multi-party European Parliament also can yield fairly high accuracy.

Yu et al. (2008) argue that training an ideology classifier is possible and fairly generalizable based on congressional speeches in the US. They build a classifier that reaches almost 90% predicting accuracy on the US Senate with training data on the House. In the opposite experiment – predicting House party affiliation based on Senate data – their best classifier falls somewhat short of the first with an accuracy of just over 65%. They also show that the results are somewhat time-dependent.

Further, building on Poole and Rosenthal's (1991) argument that a lot of the variation in voting behavior can be explained by a low-dimensional issue space, Diermeier et al. (2011) set out to explore what the contents of this dimension is. To achieve this, they utilize structured records from the US Senate, and apply standard preprocessing procedures. On the one hand, and similar to the arguments of Poole and Rosenthal (1991), they find that Senators do separate on economic issues – although in different feature sets. On the other hand, they also show that more value and moral ridden terms are used frequently, and that speeches are used to "appeal to partisan constituencies, as in the use of 'gay' versus 'homosexual"' (Diermeier et al. 2011, 51).

In their experiments on the Canadian parliament debates, Hirst et al. (2010) find that the driving features in party classification are those describing roles of opposition and government. They conduct experiments in three steps. First, they show that oral question periods are more polarized than regular debates based on the classifier having much higher accuracy on question periods. They attribute this to being driven by language of *attack* (government parties) and *defense* (opposition parties), rather than ideology. Second, Hirst et al. (2010, 737) show that, in order to more directly test the first findings, training the classifier on one parliamentary period and testing on another with inverse government/opposition roles results in a drastically poorer performance (in their case, a 40 percentage points drop in classifier accuracy), which is attributed to significant features "swapping sides". In other words, this also points in the direction of words typically being used by parties in attack and defense positions rather than to clarify their ideological stance. Last, Hirst et al. (2010) show, based on a dictionary of negative and positive words, that informing the classifier with sentiment does not seem to increase its accuracy in a significant way. In sum, Hirst et al. (2010, 740-741) emphasize that parties institutional position (cabinet vs. opposition) is more defining for parliamentary debates than their political position, and that this should be taken into account in research on such debates. This finding is particularly interesting in light of our study; if institutional position is more defining than parties' ideological position, the quest for using classification accuracy as a measure cohesion and party relations becomes problematic.

For the multi-party setting Høyland et al. (2014), by using a similar approach, classify party affiliation in the European Parliament based on speech data. While the results are generally less accurate, mostly because it is easier to get wrong classifications in a multi-party setting (in contrast to the two-party system of the US, where guessing the same party for all speeches would yield around 50% accuracy), they also demonstrate that some parties are harder to classify than others. For example, the Liberal (ELDR) party is argued to be a hard case because it shifted coalition allegiance between parties in the period under investigation, and consisted of an ideologically heterogeneous party group based on the MPs country of origin. Most interestingly, these experiments are performed to investigate a specific problem, namely whether freshmen MEPs from new member states joined parties for reasons other than ideological affinity. To do this, one has to assume that classifier performance is driven by the political/ideological characteristics of speeches, and observe that a drop in classifier performance on freshmen compared to incumbents can be attributed to differences in ideological cohesiveness. While their results do hint that this could indeed be the case, they note that the narrative of speeches in the European Parliament is for the most part driven by the topic of the debate itself, rather than party specific policies and ideology. This affects the performance of a party classifier negatively.

In sum, classification of parties from parliamentary speech has been used to make inference about a variety of substantive units of political interests. For consistency in arguments, I focus on the exercise of using classifier precision as a measure of substantive interest with regard to both intra- and inter-party relations. In the following section, I summarize some of the work on using parliamentary debates to make inference of intra-party cohesion and inter-party relations based on classification.

Intra-party cohesion and inter-party relations

One particular way of using party classification to illustrate substantial topics have been to use it as a measure of polarization i various studies (see Goet (2019), Peterson and Spirling (2017), and Gentzkow et al. (2016)). The idea is simply put that lower classification accuracy means that polarization is lower (parties are harder to distinguish from each other) and higher accuracy means polarization is higher (it is easy to distinguish parties). Because most of this literature also works with binary classification problems (two-party systems), distinguishing between correct classifications and incorrect classifications become opposite sides to the same coin; misclassifications always travel to the other party. In the multi-party setting this is quite different. Here, it is useful to distinguish between classifying correctly and where misclassifications travel; for example, a party X can be mistaken for party Z in one speech, but then mistaken for party Y in another. If we then believe that party X is closer to party Y in the ideological policy space, we should expect party X to be more often confused with party Y than party Z by the classifier. That is, if classification is a good illustration of the parties' ideological relations, classification can be utilize to not only say something about how cohesive a party is (intra-party cohesion), but also how parties relate to each other (inter-party relations).

In general, politicians' position on various political issues is one of the epitomes of modern democracies; journalists, historians, political scientists, and ordinary citizens have used the left-right dimension to distinguish between political actors and their stance since it was coined in the French parliament over 200 years ago (Rosas and Ferreira 2013, 3). Although it is beyond the scope of this paper to review the literature on (and measuring of) policy positions, it is important to emphasize that the concept of cohesion within parties and the spatial closeness between them relies on the concept of parties' policy position. Indeed, Peterson and Spirling (2017) argue that, for the UK, polarization is the "[...] difference between the *positions* of the two main parties that have held Prime Ministerial office in modern times".

More specific, both the concept of intra-party cohesion and inter-party relations spring from a long tradition of studying the American Congress, where individual roll call votes often have been used to show how divided parties are and how close legislators are to each other (Poole and Rosenthal 1984; Clinton et al. 2004; McCarty et al. 2006; Garand 2010). Intra-party cohesion studies based on roll-call data are, however, not exclusive to majoritarian systems. For example Hix et al. (2005) find that party cohesion in the European Parliament has increased over time.

Roll call votes have, nevertheless, shown to be challenging in parliamentary democracies (Owens 2003), where high party elite control over MP votes often lead to near perfect voting along party lines (Rosenthal and Voeten 2004); roll call votes in multi-party parliamentary democracies tell a story of party elite power, rather than inter- and intra-party policy differences. Consequently, the nearly perfect separation of parties in roll call votes would also indicate nearly complete cohesiveness within parties, which goes against common phenomena such as party switching and factions within parties. These problems also spill over to inter-party relations; if parties are perfectly separated, there is no way of knowing who is ideologically closer to whom. In sum, we might want to append roll-call analyses with other types of methods for determining how cohesive parties are and who they are more likely to cooperate with.

One such appendment is given by Peterson and Spirling (2017) who use party classifier precision based on speech as a measure of polarization in their study on the UK. In short, they utilize British parliamentary speeches from 1935 to 2013 and supervised machine learning to predict the party labels of MPs. The intuition is straightforward: the better a classifier predicts the correct party label on average, the higher polarization there is at that time. They also show that the polarization trends of this measure is very similar to the same trends in data generated from other sources such as the Comparative Manifestos Project RILE measure. The results are also shown to be stable across different specifications of the classification algorithm. Importantly, Peterson and Spirling (2017, 7) argue that:

> [O]ur aim is not high predictive accuracy *per se* but rather predictive consistency: i.e. a maintained assumption is that variations in accuracy from one time period to another are indeed a result of substantive differences in speeches and not an artifact of data collection problems or the failure of the algorithm to identify the relevant features.

I also note the work of Gentzkow et al. (2016), who use a parametric approach for classification to show the evolution of polarization in the US from 1873 to 2009. They find that polarization in the US Congress had a watershed moment when the Republican party united around the *Contract with America* platform in the mid-nineties, and that polarization has increased to unseen heights after that.

Summed up, the main interest in the party classification literature is not to maximize classification precision, but rather the relative difference in classifier performance over time. By separating between intra-party cohesion and inter-party relations, I will put this approach to the test in the Norwegian multi-party context.

Data and methods

Data

For the analyses, I use the Talk of Norway (ToN) corpus of parliamentary speeches from the *Storting* in the period from 1998 to 2016 (Lapponi et al. 2018). The unprocessed data frame consist of 25373 speeches over 99 variables, including speaker characteristics, party attributes, institutional variables, the text of the speech, date, time, and more. In the analyses, I exclude speeches from the parliamentary President, speeches by deputy MPs, and speeches held by MPs from parties that are not represented through all parliamentary periods. This gives us a subset of 113741 speeches.

	Min	**Mean**	**Median**	**Max**
Speeches pr. MP	1.00	38.21	30.00	291.00
Words pr. speech	51.00	343.70	205.00	4788.00
Age	20.82	48.32	49.26	75.05
Gender (female)	0.00	0.35	0.00	1.00
Cabinet party	0.00	0.24	0.00	1.00
Opposition	0.00	0.73	1.00	1.00
Support	0.00	0.03	0.00	1.00
Nynorsk	0.00	0.13	0.00	1.00
Interpellation	0.00	0.08	0.00	1.00
Oral question	0.00	0.09	0.00	1.00
Ordinary question	0.00	0.13	0.00	1.00
Ordinary debate	0.00	0.69	1.00	1.00

Table 1: Descriptive stats for selected variables

Table 1 shows descriptive statistic for the speeches used in the analyses. There is great heterogeneity in the amount of speeches MPs hold during a parliamentary session with a minimum of 1 speech and a maximum of almost 300 speeches. As I cut speeches shorter than 51 words long, this is also the minimum speech length in the data. The maximum length speech, however, is almost 5000 words long – this speech lasted for over 30 minutes. The age distribution is close to normally distributed with a large amount of speeches among MPs between 40 and 60 years old. In line with other parliaments (see Bäck and Debus (2016)), there is an underrepresentation of female MPs in Norwegian parliamentary debates, accounting for only 35% of the speeches (while occupying 40% of the seats). MPs from the opposition parties are also naturally more prone to take the floor in the plenary, with 73% of the speeches, whereas support parties have very few speeches at 3% because only the Solberg I cabinet has relied on formal support from two parties. As for the two written languages – bokmål and nynorsk – MPs can themselves decide which language to be transcribed in. Most MPs opt for bokmål, however, with only 13% being transcribed in nynorsk. Finally, ordinary debates are the most common foundations for debates with almost 70% of the activity in the plenary, whereas all three question types make up for about 10% each.

I also note that, as shown in figure A-1 in the appendix, although the Labor Party is the largest class in our data, they also speak least per seat, and the smaller parties (SV, Sp, KrF, and V) speak less in absolute terms but more in relative terms.

Annotations

One of the major benefits with the ToN corpus, in terms of text-as-data, is that the whole corpus has been run through the automatic Oslo-Bergen tagger (OBT).[1] Here, all speeches of the corpus are split into individual speech files in a CoNLL-like formatted tab separated file.[2] In these files, the tokens of a given speech are ordered in rows from the first token of the speech to the last token of the speech (where empty lines indicate sentence boundaries). The columns of the annotated files include the lemma, part of speech, and morphosyntactic tags for each row (token). Further, the OBT collapse multi-word terms into one token, so that, for example, the multi-word phrase *i dag* (today) is one token, instead of two.[3]

Classifier

For each of the preprocessing feature sets described below, I train a stochastic gradient boost (SGB) classifier with class weights to learn a function that maps text vectors to

[1] See `http://www.tekstlab.uio.no/obt-ny/english/index.html` for more detailed information on the tagger.

[2] See `http://ufal.mff.cuni.cz/conll2009-st/task-description.html` for an example of the CoNLL format.

[3] For reading the tagged ToN speeches into R, an under development package can be found at `https://github.com/ltgoslo/talk-of-norway/tree/master/src/R/tonR`.

party labels. The SGB is a simplified version of gradient descent (GD), which iteratatively uses the sum of the squared residuals as loss function to optimize the classification of a training set. GD can, however, be very computationally time consuming, as it solves the classification problem for all data points in each iteration (Bottou 2010). SGB improves this by only using a subset of the data in each iteration, greatly increasing predictive performance and lowering computational time (see Greenwell et al. (2019) for an outline of the particular SGB implementation used here).

I use a 10-fold cross-validation method for each parliamentary session (1 year from October to October), where the data is split into 10 equally large samples, train the model on 9 of these training sets and predict party label on the remaining test set (development test set or dev-test set). I then switch the dev-test set out for one of the sets in the training set, and follow the same procedure until there is a party label prediction for all speeches in the data.

As a robustness check for the SGB classifier, I also estimate random forest and neural network classifiers for some selected feature sets. The results are shown in tables A-1 and A-2 in the appendix.

Preprocessing

Language is complex and hard to hold constant; some countries might have languages unrecognizable in others due to spelling or grammatical structure. Further, within language variations are also common even at close geographical locations or social groups through dialects. Then there are contextual differences through style; a party manifesto will have a very different style of language than a fictional novel. And, within styles the topics discussed will often guide what words are used. Even at the personal level, we tend to talk differently when all of the above factors are held constant. Consequently, political texts are inherently difficult to analyze. But, there is a large bag of preprocessing tools available for reducing the complexity of language in order to make these texts as comparable as possible. Reducing complexity does, of course, come at a cost; any alteration will make the text move away from its original form and meaning.

Importantly for the purpose of this paper, there are a myriad of decisions to take in the process of going from text to numbers and making the data ready for analyses. I will highlight the differences between the models in this section. As pointed out by Denny and Spirling (2018), each preprocessing decision is a binary choice and there are a vast amount of decisions. For example, Denny and Spirling (2018) focus on seven common preprocessing decisions, which amounts to a total of 128 (2^7) combinations of decisions. They also highlight that model results may vary substantially across these combinations.

In order to make the ananlyses as precise as possible, I hold a number of preprocessing decisions constant across all models. First, I use TF-IDF (Term Frequency – Inverse Document Frequency) as values instead of raw frequency counts of the features. TF-IDF has a relatable function to weighting in standard regression analyses, as it provides a way of weighting the frequency of an observation with its degree of ubiquitousness across documents in the data (Manning et al. 2009); the intuition behind TF-IDF is that the features appearing across many documents are going to have less disambiguating potential than those that do not. Second, I remove the 100 highest scoring IDF-tokens. This is a technique for removing stop-words: frequent words that seldom contribute to discrimination and potentially decrease computational time significantly. I opt for this solution because the stop word dictionaries available in Norwegian are somewhat limited. Third, I balance the classes in each fold by removing parties that did not hold a seat in the Storting over all periods in the period our data covers. Thus, the Green Party (*Miljøpartiet de Grønne*), Non-Partisan Deputies (*Tverrpolitisk Folkevalgte*), the Coastal Party (*Kystpartiet*), and independent MPs are not used in our models. Finally, I only keep speeches longer than 50 tokens and remove tokens that occur less than 21 times across all speeches in the session in order to be as sure as possible that speeches do contain positional statements and retain as much data as possible at the same time.[4] All these preprocessing steps are done for all the models shown in this paper. Next, I will highlight the 5 preprocessing decisions that are used in the analyses.

[4] Some speeches in the corpus can, for example, be a shout out from an MP: "I voted wrong" or "No!". Other short speeches can be statements on procedure.

Baseline

The first feature set draws loosely on Grimmer and Stewart (2013) for the most used preprocessing decisions in quantitative text analyses within political science. I label this specification as the *baseline*. It is important to note that this framework is not used in *all* applications of quantitative text analysis in political science, but rather an approximation of a preprocessing feature set that would be plausible in a political science application to my data.

First, I lowercase all tokens in order to not differentiate between same tokens in the start of a sentence and later in the sentence. Second, I remove numbers and punctuation. Third, I split the speeches into tokens by using the *tokenizers* package for R, which strips all whitespace and punctuation and returns a vector of tokens based on it (Mullen 2016). Fourth, I stem the tokens with the SnowballC stemmer for Norwegian (Bouchet Valat 2014). This is a procedure used for keeping only the stem of a token, converting the tokens in different grammatical forms the same token (for example "party" and "parties" are converted to "parti" by the English version of the SnowballC stemmer). The intuition is that a word in different forms is still the same word, and should denote similarity rather than difference.

Lemma

The remaining feature sets consist of token lemmas, retained from running the corpus through the OBT tagger. Using lemma is seen as a less crude method for normalizing the form of words than stemming (Manning et al. 2009, 32), which is used in the *baseline* feature set. Compared to stemming, which only keeps the stem of a token, lemmatization converts the same token in different forms to the root form (dictionary form) of the token. For example, irregular verbs are such as "did" and "done", would be converted to its root – "do". Or, irregular plural nouns such as "elves" would be converted to "elf" (instead of the stemmed version "elv"). The tagger is also trained to look at the context a token occurs in to pick the correct root for words that are written identically but has different meaning in different contexts (for example, Rose as a name is different from the flower).

Part of speech

Further, a set of models include part of speech (PoS), obtained from the OBT. PoS denotes the word group a token belongs to in terms of syntactic function. For example, the token "walking" is assigned to the category "verb", the token "weird" to the category "adjective", and so on. These tags can thus be "[...] considered to be a crude form of word sense disambiguation" (Pang and Lee 2008, 21). In other words, PoS can the classification model distinguish between identical tokens, occurring with different grammatical functions in relation to the context.

N-grams

Next, I supplement the token unigrams from the previous models with token and lemma bigrams and/or trigrams in another set of models. N-grams are n number of co-occurring words in a sequence of text. Token unigrams are, thus, single tokens, bigrams are pair of words, and trigrams are three words in sequence.[5] Importantly, the OBT tagger provides us with sentence boundaries, which is helps us construct n-grams that do not cross from one sentence to the next.

The main benefit of including n-grams in the model is that it does account for some level of word order where unigrams completely disregards the order words come in. In Norwegian, the word *gift*, for example, can mean both "married" and "poison". Thus, it is important to know the context of the word *gift* in order to understand the meaning of the word. From the ToN corpus, I can exemplify with the phrases *ulikhet er gift* ("inequality is poisonous") and *lykkelig som gift* ("happily married"), where *gift* would be the same token with unigrams, but very different with token trigrams.

Metadata

Last, I also feed a set of variables to a set of SGB models – similar to including controls in a regression analysis. The variables included are both at the speaker and debate level:

[5] The sentence "build a straw man argument" is a vector of five unigrams (each word for itself), four bigrams ("build a", "a straw", "straw man", and "man argument"), and three trigrams ("build a straw", "a straw man", and "straw man argument").

gender and county of provenance are the speaker level attributes, and type of debate (minutes, question hour, interpellations, and so on), keywords (for instance, “taxes”, “research”, “immigration” and so on), the name of the committee leading the debate, and finally the type of case (ordinary issues, budget, legislation) are the debate level attributes. Further, I include one feature set where I inform the classifier on whether the party was a part of the cabinet or in opposition. As will be discussed below, this feature set is only used as an illustration of how powerful a single contextual meta data variable can be when we classify parties.

The selected variables are by no means meant to be an exhaustive list of relevant covariates, but rather serve as an illustration for how meta data variables can contribute to increase model accuracy.

Classifier performance

Feature sets

In this section, I show the overall performance differences between the classifiers built on the various combinations of preprocessing choices. This makes me able to test whether the classification performance is significantly affected by the different preprocessing decisions. As a first step, I use the highest probability class for each speech in each model as the predicted party for that model and compare this to the actual party of the MP holding the speech.

I opt for using F_1 scores over only accuracy in this section for comparing the performance of these models. F_1 scores rewards the model for not producing both false positives and false negatives, whereas accuracy only accounts for true positives.[6].

Table 2 shows the F_1 scores and accuracy for the 13 feature sets and the accuracy for the majority class. Noticeably, the baseline model performs worse than all other models in predicting the correct party, although it has over double the accuracy than predicting Labor Party – the largest class (majority class) in the data with 21.7% of the speeches

[6] The F_1 score is calculated as: $\mathrm{F}_1 = \frac{2\mathrm{PR}}{\mathrm{P}+\mathrm{R}}$, where P is precision and R is recall. Precision is defined as: $\mathrm{P} = \frac{\text{true positives}}{\text{true positives+false positives}}$, and recall as $\mathrm{R} = \frac{\text{true positives}}{\text{true positives+false negatives}}$.

– on all speeches. The differences are, however, not large between the baseline and the non-meta data feature sets. Further, including PoS and n-grams seems to not improve the model at all, and even make it gradually worse as more information is put into the model. Including meta data has a big impact on the performance of the classifier. Even the meta data only feature set (set 7) has better performance than the feature sets without meta data. Also here, including more language features than token lemmas does not improve the model.[7] Noticeably, feature set 12 has a very high accuracy compared to the other feature sets. This is because this model includes an indicator for whether the party had cabinet power or was in the opposition at the time of the speech. I will discuss the impact of this feature set later. For now, I focus on the baseline and 1 through 11 feature sets.

In short, the best classifier (set 9) does a good job of classifying the speeches in Stortinget, with an F_1 score of 0.560 and accuracy of 54.5%. For a seven class classification problem, all models do a pretty good job of assigning the correct party labels. But, even though the accuracy of the different feature sets are close, the discrepancies between the models in terms of F_1 score are significant enough for us to suspect that the incorrect classifications are distributed differently across the sets. If classifier performance is to be used as a measure of substantive interest, we need to capture the spatial distribution of parties both in classification and misclassification.

In sum, the initial impression is that feeding our classifier token lemmas and meta data significantly increase classification performance, but including further language features seems insignificant. In contrast to Lapponi and Søyland (2016), n-grams and PoS does not seem to improve the performance. This might be a consequence of a stricter sampling of the data into one year 10 folds, instead of the four year 10 folds used in Lapponi and Søyland (2016), meaning such language features might only improve performance if there is a lot of data in the train and test sets. In the next section, I build on the results above to test whether using classification as a measure of intra-party cohesion and inter-party relations is a viable strategy in the Norwegian case.

I do this in three steps: first, I study the intra-party cohesion by using both class

[7] Running the same feature sets with a random forest classifier largely give the same results, although the general prediction power is higher – peaking at an F_1 score of 0.782 for feature set 8.

Set	Lemma	PoS	Bigram	Trigram	Meta	F_1	Accuracy
Baseline						0.479	0.466
1	X					0.504	0.482
2	X	X				0.503	0.481
3	X		X			0.488	0.465
4	X	X	X			0.487	0.465
5	X		X	X		0.486	0.463
6	X	X	X	X		0.485	0.463
7					X	0.511	0.491
8	X				X	0.559	0.543
9	X	X			X	0.560	0.545
10	X	X	X		X	0.548	0.534
11	X	X	X	X	X	0.545	0.532
12	X				X	0.672	0.670
Majority							0.217

Table 2: List of feature sets accompanied by the macro F_1 score and accuracy.

predictions and the probability distributions these classifications. Second, I analyze the inter-party relations by looking at whether parties probabilities be spatially closer are misclassified to each other more often than those further apart. Here, I also use both class predictions and classification probabilities. Finally, I replicate a regression model from Søyland and Høyland (2020) looking at the determinants of who gets to speak in parliament and include the probability of the classifier for each MP as a latent measure of agreement with her party. Within these sections, I also discuss the effect of different feature set on the inference we can make about substantive questions.

Intra-party cohesion

Class predictions. In the first test of using party classification as a measure of intra-party cohesion, I use the class predictions: classifying each speech to the party with the highest probability from a model. First, I show how accurate speeches are predicted within parties. In order to insure that the difference of performance between our feature sets is not produced by chance, I use approximate randomization on the F_1 scores for parties (Noreen 1989). With this method, I throw out half the sample at random, calculate the F_1 scores, repeat this 1000 times, and extract the 2.5% quantile, mean, and 97.5% quantile

F_1 scores over all 1000 samples for each model.

Figure 1 shows the F_1 scores for all parties (points) across four of the feature sets presented above (baseline, 1, 6, and 8), with their confidence intervals from the approximate randomization. As illustrated above, the figure shows that the *lemma/meta* feature set outperforms the other models for all parties in terms of macro F_1 score. As for party specific F_1 scores, the *baseline* model scores lowest for most parties and the *lemma/meta* model the highest. Notably, the *baseline*, *lemma*, and *lemma/PoS/trigram* are very similar.

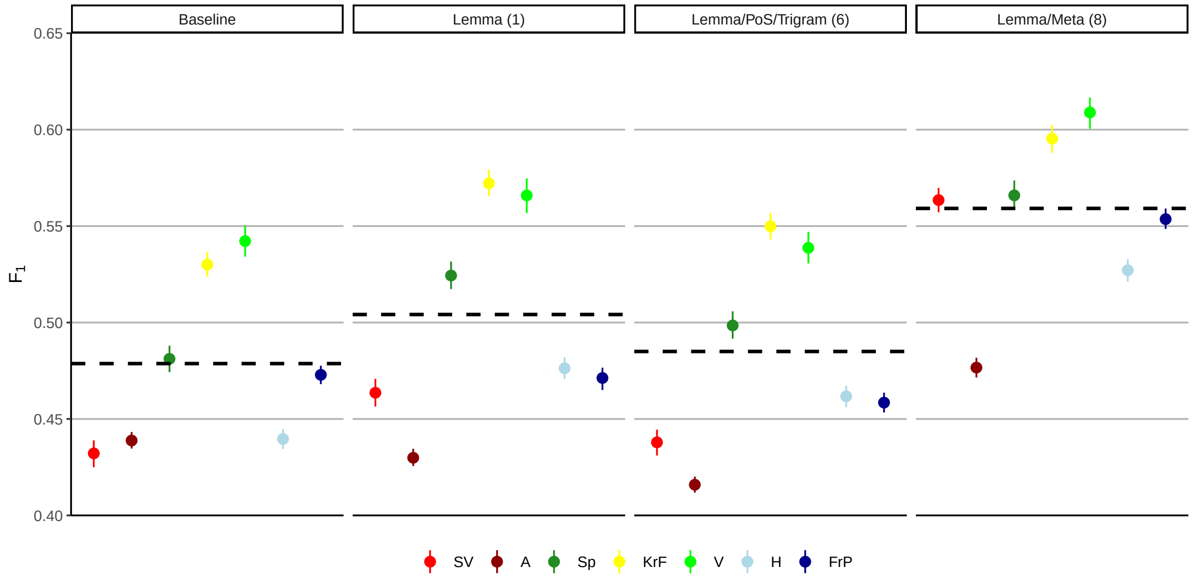


Figure 1: Baseline, lemma, lemma/PoS/trigram, and lemma/meta feature set F_1 scores and approximate randomized 95% confidence intervals for all parties. The horizontal darker lines through the plot shows the mean F_1 score for each feature set.

As for differences between parties, Figure 1 indicates that the classifiers do a better job of classifying the two parties with the least amount of speeches: the Liberal Party (V) and Christian People's Party (KrF)[8]. Further, the models generally have a harder time correctly classifying the Socialist Left Party (SV), Labor Party (A), and Conservative Party (H) in all feature sets, except the *lemma/meta* model, where the Labor Party has a far lower score than the others. The right-wing Progress Party (FrP) is below the macro F_1 score in all models and the Center Party (Sp) is slightly higher except in the *lemma/meta*

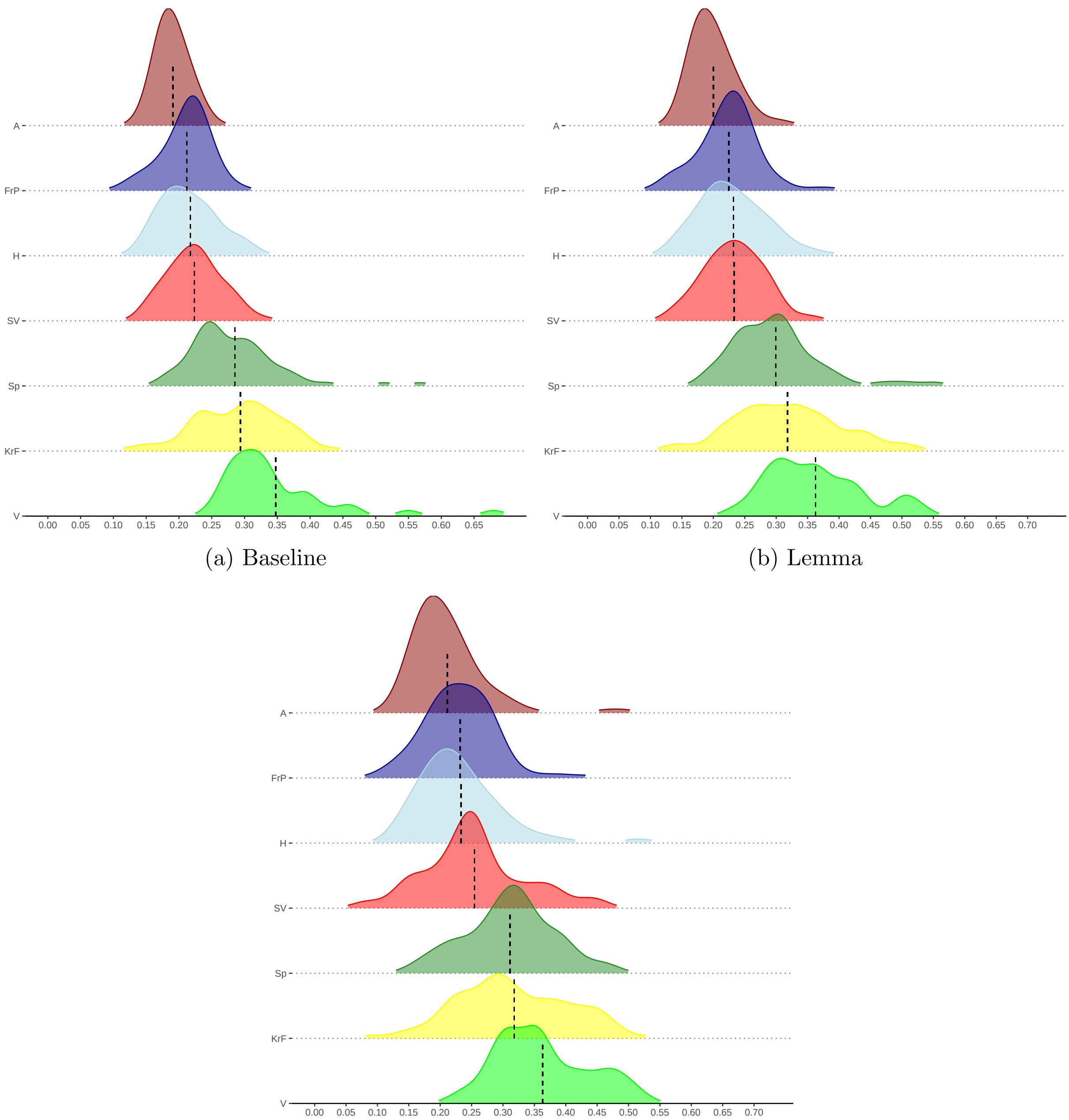
[8] This could be a result of the agressive class weights used in the modelling.

model. Importantly, however, the differences within models between parties are pretty stable in all feature sets; the variance between party F_1 scores is close across the four models[9]. This could suggest that the misclassifications from the different feature models consistently show intra-party cohesion. If figure 1 is a good representation of intra-party cohesion, we can conclude that the party occupying the center of Norwegian politics are more cohesive than the ones on the left and right side. This seems counter intuitive at face value; we would expect the more “extreme” parties to be more distinguishable than the parties occuping the center. However, it is hard to give a definitive answer as to which parties are more cohesive based on the results from figure 1 alone.

Probability distributions. One feature seldom used in classification analyses is the underlying probabilities of the class predictions. That is, for each class prediction the model predicts the class scoring the highest probability over all classes – for the seven class problem at hand, if the classifier had no idea about which class was higher in probability for a speech, all parties would be assigned a probability of 14.3% ($\frac{100}{7}$) for that speech. In order to get a deeper grasp on intra-party cohesion, I extract these probabilities for the true party of each speech in the data. In other words, since we know the correct party of each speech, I extract the probability of that party in the classification model output. The results for the three selected feature sets is shown in figure 2.

The figure shows a more nuanced picture than the absolute classification output used above. The parties are ranked similar across models, which indicates that preprocessing does not have a large effect on the prospect of using classification as a measure of party cohesion. Also, the smaller parties (V and KrF) still have the highest average probability of being correctly classified, as shown by the vertical dashed line for each party. However, the distributions for these parties are flatter than the other parties; that is, the small parties have a greater variety in how difficult their speeches are to classify. This does make sense, as parties occupying the center of politics often have potential cooperation partners on both sides, depending of the issue at hand. In sum, the results for

[9] The variance for the F_1 scores: 0.00199 (Baseline), 0.00296 (Lemma), 0.00257 (Lemma/PoS/Trigram), and 0.00194 (Lemma/Meta)



(a) Baseline

(b) Lemma

(c) Lemma/Meta

Figure 2: Probability distribution from SGB classifiers for classifying the correct party over four feature sets. The vertical line shows the mean probability within each party.

using classification as a measure of intra-party cohesion seems to be reasonable. There is one exception though; the model including the meta data variable for a party's role in parliament (cabinet/opposition) does have an effect of the ranking at the lower levels of cohesion, shown in figure A-2 in the appendix.

Inter-party relations

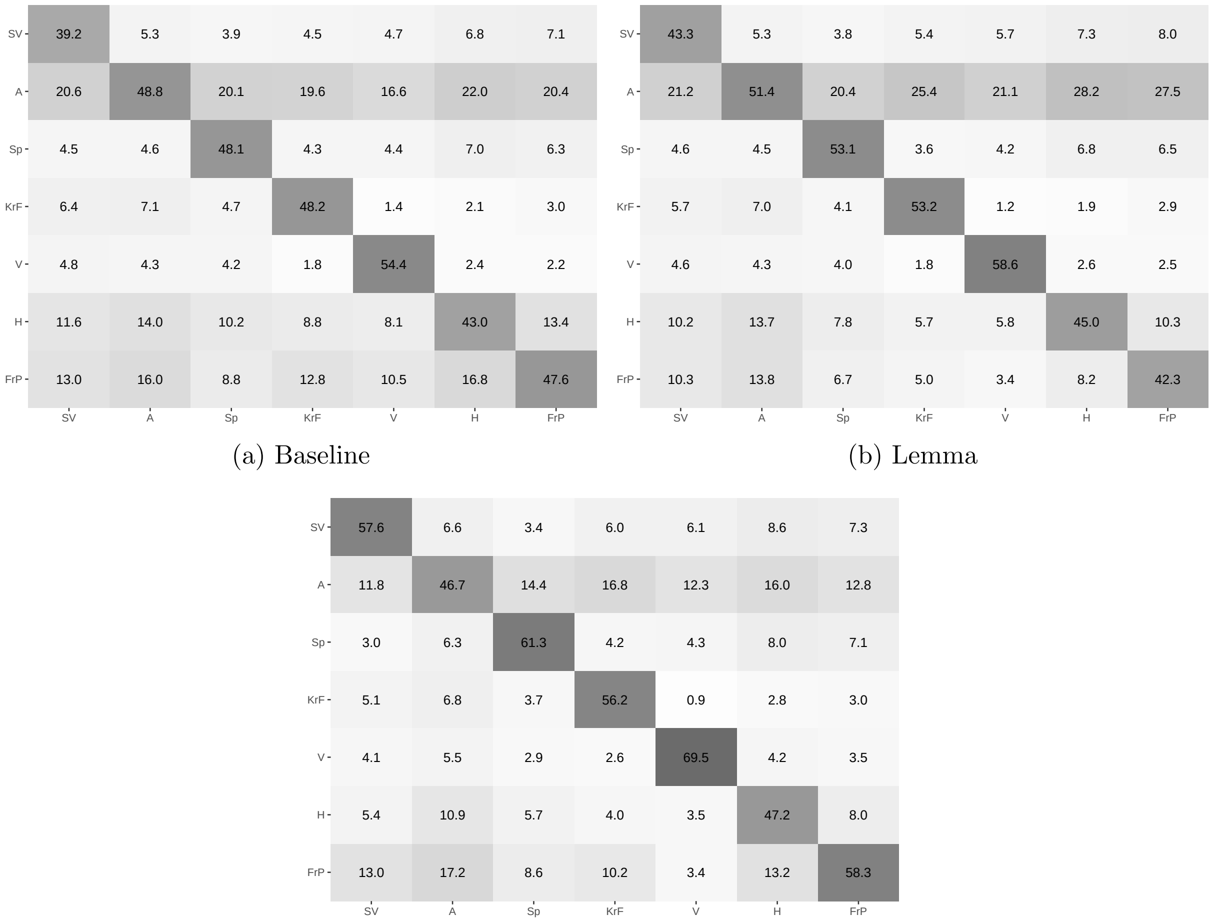
Absolute classification Similar to the analyses above, I use both absolute classification and the underlying probabilities to explore whether the classifiers captures inter-party relations.

First, figure 3 shows the percentage of correctly classified speeches on the diagonal, with the true party on the y-axis and the prediction on the x-axis. In other words, the lower left tile in each panel shows how many percent – 13% – of Socialist Left Party (SV) speeches that were classified as the Progress Party (FrP) and the upper right tile how many percent – 7.1% – of FrP speeches were classified as SV. Parties are aligned according to the their perceived position in the left-right policy space – with the Socialist Left Party (SV) being the furthest left and the Progress Party (FrP) furthest right. If the classifiers contain information on close inter-party relations, we would expect the squares closer to the top-down diagonal to be more shaded than those further from the diagonal.

Panel 3a, 3b, and 3c show little signs of such patterns.[10] Indeed, all the three panels have higher probabilities in the lower left corner, indicating strong relations between the two leftmost parties hand the two rightmost parties. Further, the Labor Party (A) seems to concede a fair amount of classifications to all parties. In sum, the misclassifications do not seem to give a good picture of inter-party relations here. Parties perceived to be closer are not more commonly mistaken for each other than parties further away in the ideological space. But, the analyses so far have not accounted for issue dimensionality; although the Progress Party might generally be perceived as being the rightmost party in Norwegian politics, the Christian Democrats may safely be assumed to be the most conservative on issues regarding religion.

Probability distribution. As for probability distributions, figure A-4 in the appendix shows the probability distribution for the Labor Party (A, 2005-2009) to highlight that the spatial sorting is identical between the feature sets. Here, I zoom in on dimension

[10] Although including party role, shown in figure A-3 in the appendix does improve spatial closeness somewhat.



	SV	A	Sp	KrF	V	H	FrP
SV	39.2	5.3	3.9	4.5	4.7	6.8	7.1
A	20.6	48.8	20.1	19.6	16.6	22.0	20.4
Sp	4.5	4.6	48.1	4.3	4.4	7.0	6.3
KrF	6.4	7.1	4.7	48.2	1.4	2.1	3.0
V	4.8	4.3	4.2	1.8	54.4	2.4	2.2
H	11.6	14.0	10.2	8.8	8.1	43.0	13.4
FrP	13.0	16.0	8.8	12.8	10.5	16.8	47.6

(a) Baseline

	SV	A	Sp	KrF	V	H	FrP
SV	43.3	5.3	3.8	5.4	5.7	7.3	8.0
A	21.2	51.4	20.4	25.4	21.1	28.2	27.5
Sp	4.6	4.5	53.1	3.6	4.2	6.8	6.5
KrF	5.7	7.0	4.1	53.2	1.2	1.9	2.9
V	4.6	4.3	4.0	1.8	58.6	2.6	2.5
H	10.2	13.7	7.8	5.7	5.8	45.0	10.3
FrP	10.3	13.8	6.7	5.0	3.4	8.2	42.3

(b) Lemma

	SV	A	Sp	KrF	V	H	FrP
SV	57.6	6.6	3.4	6.0	6.1	8.6	7.3
A	11.8	46.7	14.4	16.8	12.3	16.0	12.8
Sp	3.0	6.3	61.3	4.2	4.3	8.0	7.1
KrF	5.1	6.8	3.7	56.2	0.9	2.8	3.0
V	4.1	5.5	2.9	2.6	69.5	4.2	3.5
H	5.4	10.9	5.7	4.0	3.5	47.2	8.0
FrP	13.0	17.2	8.6	10.2	3.4	13.2	58.3

(c) Lemma/Meta

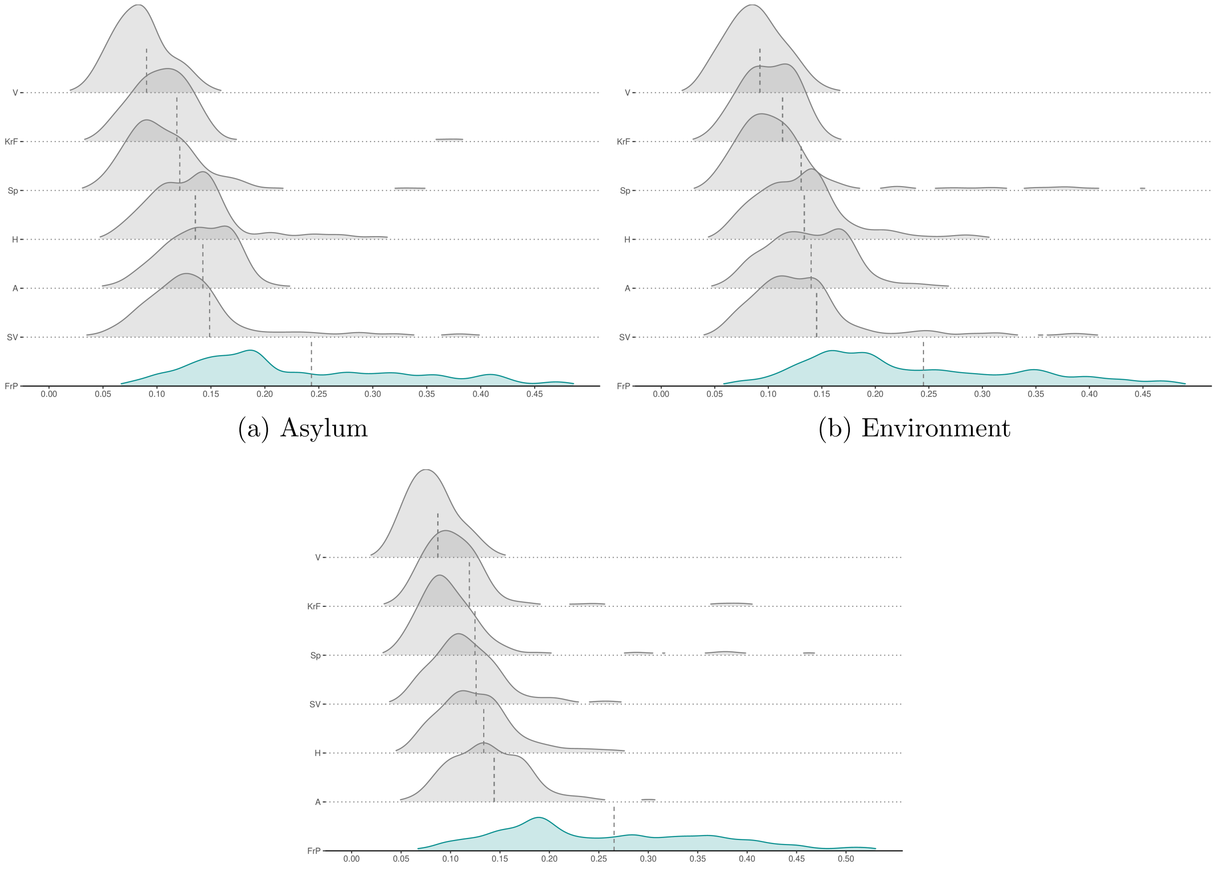
Figure 3: Accuracy of predicting correct party in percent with true party on the y-axis and predicted party on the x-axis.

specific inter-party relations.

Figure 4 shows all speeches from the Progress Party (FrP) over three selected policy dimensions. The dimensions are filtered through searching for keywords in the keyword lists provided by *Stortinget*.[11] Being the far right party, FrP is expected to be furthest away from the more immigration friendly parties (SV, V, and KrF) in panel 4a. For the latter two, this seems to be correct, but SV is actually the closest party to FrP in this panel. This could be a consequence of the issue being more salient for these two parties; they talk more about it, and are thus more likely to use similar words on those subjects. For panel 4b on environmental debates, the same expectation and result appears. FrP is closest to the party it should be the furthest away from. Finally, panel 4c considers a

[11] Three simple regex searches were done for filtering dimensions: 1) asylum – ”[Aa]syl”, 2) environment – ”[Mm]iljøv—[Nn]atur”, and 3) road toll – ”[Bb]ompeng”.

contentious topic in Norwegian politics: road tolls. Here, FrP is strictly against using such tolls, whereas the Liberal Party (V) and Labor Party (A) traditionally have been more positive. The Liberal Party does emerge to be the furthest party from the Progress Party on this issue, but the Labor Party is its closest neighbor. It is also worth noting that even though the Liberal Party are furthest away in all these debates, this does not necessarily mean that we are capturing policy position differences; the Liberal Party is also the party furthest away from the Progress Party if we look at all speeches in the corpus combined. In sum, there is no reason to conclude that there is positionality describing inter-party relations in the classifiers used in this paper.



(a) Asylum

(b) Environment

(c) Road tolls

Figure 4: Probability distribution for the Progress Party (FrP) on three issues from SGB classifiers. True classification is colored in cyan, other classes colored in gray. Parties are ranked, so that the lower a party is on the y-axis, the higher the average probability is that that party is classified as the Progress Party.

As shown in the appendix, figure A-5 for the classifier including party role (cabinet/opposition), shows a more plausible picture, placing the two coalition partners of the Labor Party – the Socialist Left Party (SV) and Center Party (Sp) – as the closest neighbors in addition to the model placing the two far right parties, very far away. I am, nevertheless, hesitant to use this model as it might help the classifier too much when providing uniform party information.

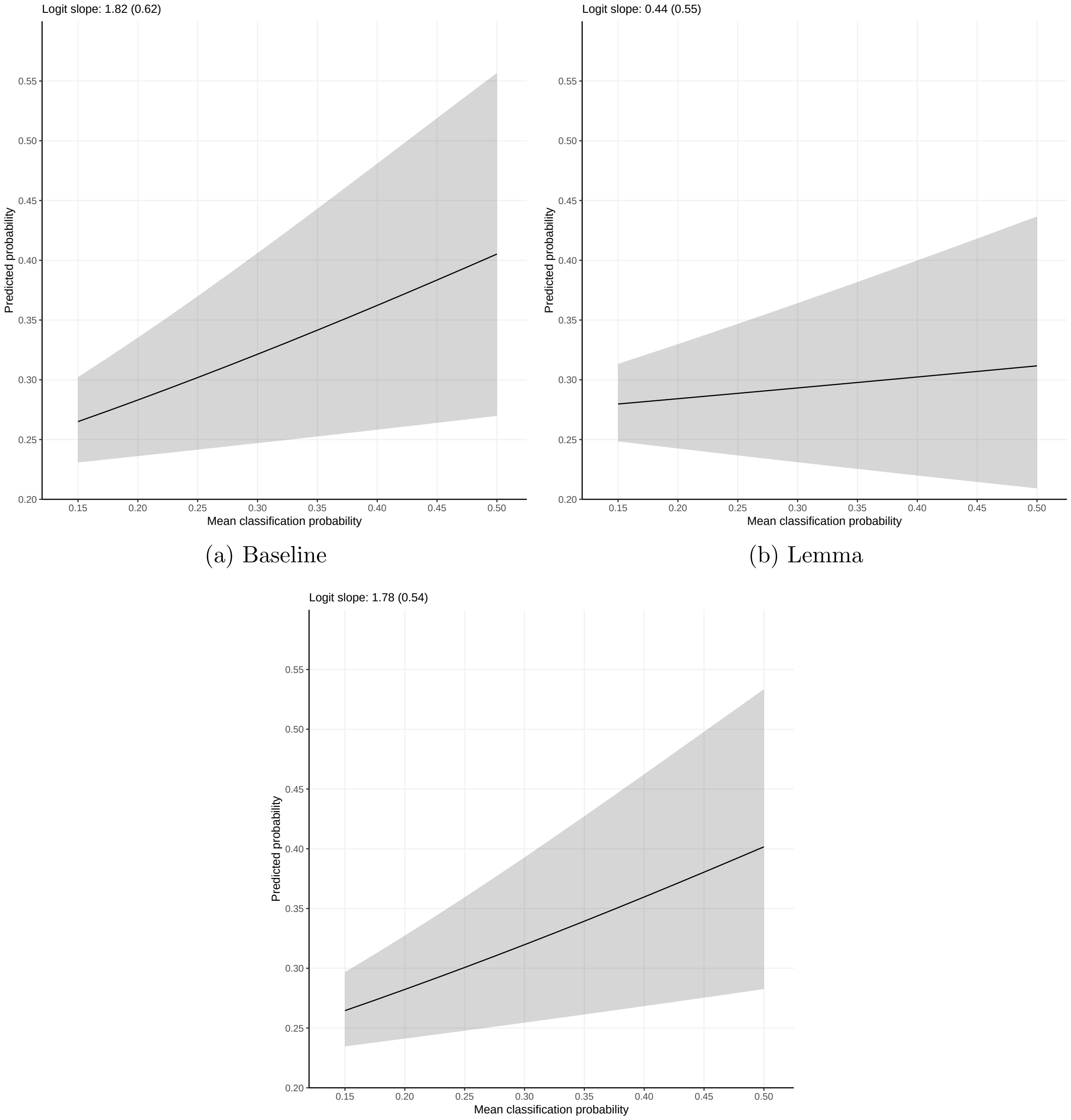
Classification as a measure in practice

Finally, in order to investigate the practical impact of using different preprocessing feature sets, I replicate the regression model from Søyland and Høyland (2020) (see table A-3 in the appendix). This analysis explores the determinants for which MPs are selected to speak in parliament. I append this with including classification as a measure of interest, aiming at measuring intra-party cohesion. I use the predicted probabilities of correctly classifying each speech over all feature sets, then averaging the score for individual MPs within each parliamentary period (of four years). In that way, the measure is supposed to represent an MPs closeness to her own party (intra-party cohesion).

Figure 5 shows the effect of this variable on speaker selection – holding all other variables constant – across the a) Baseline, b) Lemma, and c) Lemma/Meta. Although the estimates are positive in all three specifications, the slope of the effect varies substantially. For the baseline model, an MP with 25% probability of being correctly classified (x-axis) on average has a likelihood of about 30% to be selected for holding a speech in the plenary (y-axis). If we increase the classification probability to 45%, the same MP has over 36% chance of being picked – a jump of approximately 6%. This is a substantial effect. However, the two remaining models have less steep slopes; in panel 5b, the same jump would increase the likelihood with only about 2% and the effect is not significant. And, panel 5c gives an expected increase of about 5% for the jump from 25% to 40% cohesion.[12]

In sum, using classification in applications handling political processes should be handled with care. Even small pre-processing tuning can substantially affect the subsequent

[12] See fig A-6 and A-7 in the appendix for the same estimation over all feature sets. Note also, that the Lemma/Party role specification, shown in table A-3, has an even steeper curve.



(a) Baseline

(b) Lemma

(c) Lemma/Meta

Figure 5: Replication of Søyland and Høyland (2020) including cohesiveness measures as an independent variable in the full model. The effect is recalculated for four selected preprocessing feature sets. The logit slope and standard error is shown in the top left of each panel.

inference we draw from ananlyses.

Discussion

Producing substantive measures on policy positions is an important task in the field of comparative politics. Such measures are used not only to describe political change in democracies, but also used for explaining behavior within these systems. We thus depend on them being as precise as possible; if we want to investigate the effect of, for example, party policy position on some dependent variable of interest, the measure of policy position must reflect the actual position of the parties we study in order for our inference to be valid.

In this paper, I have explored how using classification of party labels based on parliamentary debates works for this task. By utilizing a unique dataset on speeches in the Norwegian parliament, my analyses of classification based on different preprocessing feature sets show relatively stable results on the macro level of each model. However, looking at the results as measures of interest with regard to both intra-party cohesion and inter-party relations shows that preprocessing decisions can have a large impact on subsequent inference. Further, I have shown that investigating the underlying probabilities of the classification prediction can help paint a clearer picture of how successful the model is at capturing concepts, rather than looking at only the highest probability categorical classification predictions.

In strict terms, if classification of party labels in analyses of text are to be used as a substantive measure, we have to assume that the our classification inhibit some sort of spatialness. As shown in this paper, this is not necessarily the case. At least, parts of misclassification is driven by omitted variable bias, in form of both controlling for relevant institutional attributes, MP specific variables, and the linguistic features we feed our text models. Further, the results only show traces of positionality, where more complex models seem to better explain inter-party relations. I thus argue, that using classification models as meaningful measure in these two ways is an optimistic approach. Current approaches are unable to differentiate between what is omitted variable bias and what is the subject of interest; it is still unclear what the explained and unexplained variance actually captures in these models. And, policy positions are inherently unobservable.

I also argue that these findings could go beyond the Norwegian case in that similar patterns should be expected to occur in other multi-party parliamentary systems. Whether the results are generalizable to majoritarian systems is, however, unclear. One possible avenue for testing this could be to predict party labels on smaller parties or independents. A hard test could, for example, be to see whether Liberal Democrats in the British parliament are more prone to be misclassified as Tories or Labor Party speeches on policy dimensions they are perceived to be closer to one or the other. A softer test could be to predict party labels of far left or right independents in the US. In any case, my analyses show that classifications are not exclusively driven by party positions, and I urge researchers to take this into consideration when utilizing such measures for substantive tests.

References

Bäck, H. and M. Debus (2016). Political Parties, Parliaments and Legislative Speechmaking. Palgrave.

Bottou, L. (2010, August). Large-scale machine learning with stochastic gradient descent. In Y. Lechevallier and G. Saporta (Eds.), Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010), Paris, France, pp. 177–187. Springer.

Bouchet Valat, M. (2014). SnowballC: Snowball stemmers based on the C libstemmer UTF-8 library. R package version 0.5.1.

Clinton, J., S. Jackman, and D. Rivers (2004, May). The Statistical Analysis of Roll Call Data. American Political Science Review 98(02), 355–370.

Denny, M. J. and A. Spirling (2018). Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. Political Analysis 26(2), 168–189.

Diermeier, D., J.-F. Godbout, B. Yu, and S. Kaufmann (2011, May). Language and Ideology in Congress. British Journal of Political Science 42(01), 31–55.

Garand, J. C. (2010). Income Inequality, Party Polarization, and Roll-Call Voting in the U.S. Senate. The Journal of Politics 72(4), 1109–1128.

Gentzkow, M., J. M. Shapiro, and M. Taddy (2016). Measuring Polarization in High-Dimensional Data: Method and Application to Congressional Speech. Technical report, National Bureau of Economic Research.

Goet, N. D. (2019). Measuring polarization with text analysis: Evidence from the uk house of commons, 1811–2015. *Political Analysis* *Early view*.

Greenwell, B., B. Boehmke, J. Cunningham, and G. Developers (2019). *gbm: Generalized Boosted Regression Models*. R package version 2.1.5.

Grimmer, J. and B. M. Stewart (2013, Jan). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis* *21*(3), 267–297.

Hirst, G., Y. Riabinin, and J. Graham (2010). Party status as a confound in the automatic classification of political speech by ideology. In *Proceedings of the 10th International Conference on Statistical Analysis of Textual Data (JADT 2010)*, pp. 731–742.

Hix, S., A. Noury, and G. Roland (2005). Power to the Parties: Cohesion and Competition in the European Parliament, 1979–2001. *British Journal of Political Science* *35*(2), 209–234.

Hix, S., A. G. Noury, and G. Roland (2007). *Democratic Politics in the European Parliament*. Cambridge University Press.

Høyland, B., J.-F. Godbout, Godbout, E. Lapponi, and E. Velldal (2014). Predicting Party Affiliations from European Parliament Debates. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. Association for Computational Linguistics.

Lapponi, E. and M. G. Søyland (2016). Talk of Norway. Hello World.

Lapponi, E., M. G. Søyland, E. Velldal, and S. Oepen (2018). The Talk of Norway: a richly annotated corpus of the Norwegian parliament, 1998–2016. *Language Resources and Evaluation*. Published online.

Manning, C. D., P. Raghavan, and H. Schütze (2009). Introduction to information retrieval. Online version.

McCarty, N., K. T. Poole, and H. Rosenthal (2006). Polarized America: The Dance of Ideology and Unequal Riches. The MIT Press.

Mullen, L. (2016). tokenizers: A Consistent Interface to Tokenize Natural Language Text. R package version 0.1.4.

Noreen, E. W. (1989). Computer-intensive methods for testing hypotheses. Wiley.

Owens, J. E. (2003). Part 1: Cohesion. The Journal of Legislative Studies 9(4), 12–40.

Pang, B. and L. Lee (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1-2), 1–135.

Peterson, A. and A. Spirling (2017). Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems.

Poole, K. T. and H. Rosenthal (1984). The Polarization of American Politics. The Journal of Politics 46(4), 1061–1079.

Poole, K. T. and H. Rosenthal (1991, Feb). Patterns of Congressional Voting. American Journal of Political Science 35(1), 228.

Rosas, J. C. and A. R. Ferreira (2013). Left and Right: Critical Junctures. In J. C. Rosas and A. R. Ferreira (Eds.), Left and RIght: The Great Dichotomy Revisited, Chapter 1, pp. 2–21. Cambridge Scholars Publishing.

Rosenthal, H. and E. Voeten (2004, Jul). Analyzing Roll Calls with Perfect Spatial Voting: France 1946-1958. American Journal of Political Science 48(3), 620–632.

Søyland, M. and B. Høyland (2020). Parliamentary debates in the norwegian parliament. In J. M. A. Fernandes, H. Bäck, and M. Debus (Eds.), The Politics of Legislative Debate, Chapter TBA, pp. TBA. Oxford University Press.

Yu, B., S. Kaufmann, and D. Diermeier (2008, Jul). Classifying Party Affiliation from Political Speech. Journal of Information Technology & Politics 5(1), 33–48.

Appendix

Set	**Lemma**	**PoS**	**Bigram**	**Trigram**	**Meta**	F_1	**Accuracy**
Baseline						0.540	0.521
1	X					0.585	0.567
2	X	X				0.580	0.562
4	X	X	X			0.557	0.542
5	X		X	X		0.570	0.554
6	X	X	X	X		0.567	0.552
7					X	0.458	0.311
8	X				X	0.783	0.776
11	X	X	X	X	X	0.774	0.768
Majority							0.217

Table A-1: List of feature sets accompanied by the macro F_1 score and accuracy for random forest estimation.

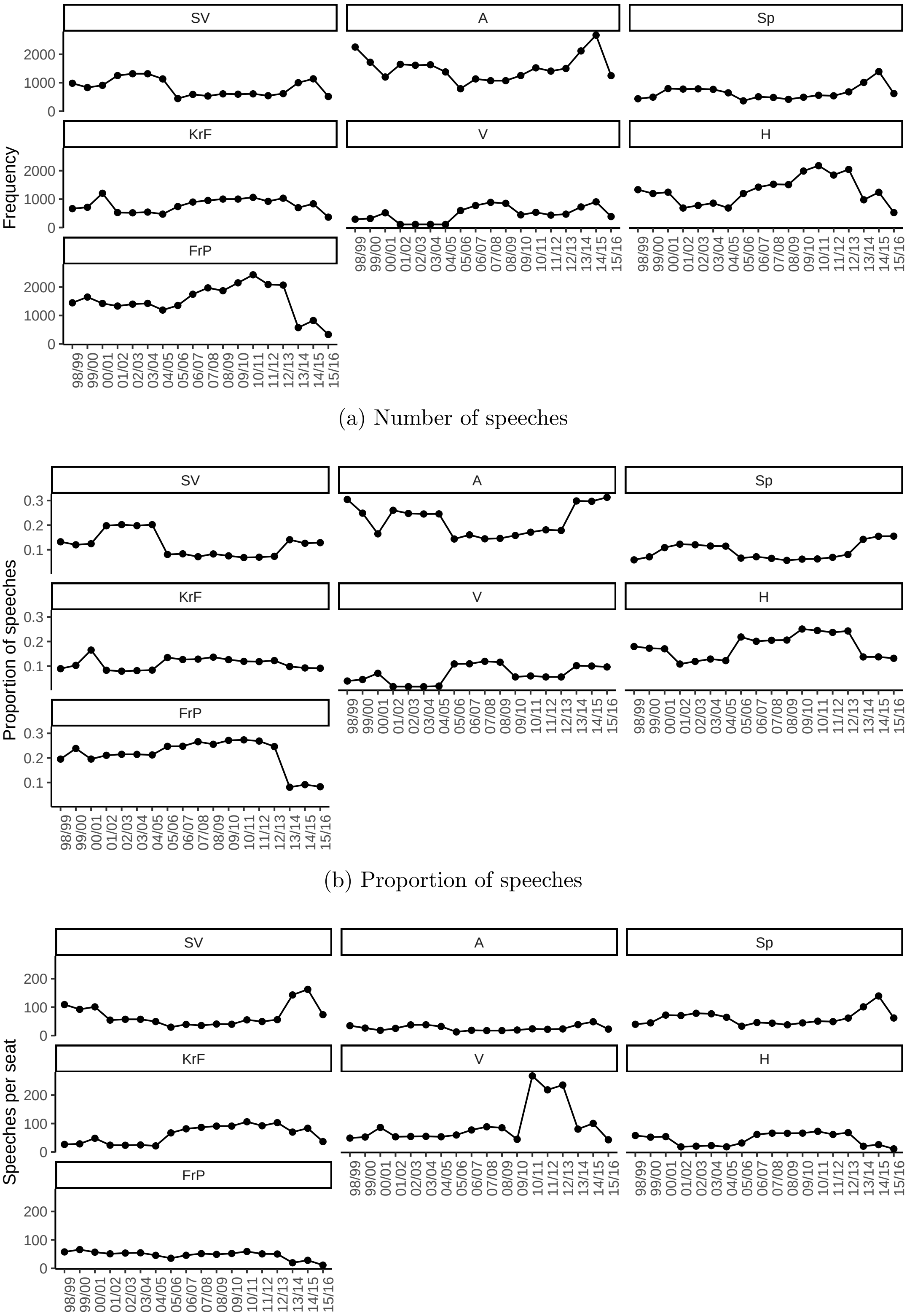
Set	**Lemma**	**PoS**	**Bigram**	**Trigram**	**Meta**	F_1	**Accuracy**
1	X					0.599	0.607
6	X	X	X	X		0.560	0.605
Majority							0.217

Table A-2: List of feature sets accompanied by the macro F_1 score and accuracy for neural network estimation.

Table A-3

	Dependent variable:				
	MP holding speech				
	Original	Baseline	Lemma	Lemma/Meta	Lemma/Party role
	(1)	(2)	(3)	(4)	(5)
Cohesion		1.818*	0.438*	1.782*	1.915*
		(−0.113)	(−0.126)	(−0.179)	(−0.178)
Deputy committee chair	0.569*	0.538*	0.560*	0.515*	0.513*
	(−0.003)	(−0.003)	(−0.003)	(−0.002)	(−0.002)
Committee chair	0.609*	0.613*	0.612*	0.623*	0.622*
	(−0.011)	(−0.012)	(−0.012)	(−0.018)	(−0.019)
Not committee member	−3.874*	−3.879*	−3.874*	−3.880*	−3.882*
	(−0.0003)	(−0.0004)	(0.0001)	(−0.003)	(−0.003)
Previous speeches (log)	0.286*	0.284*	0.286*	0.279*	0.278*
	(0.003)	(0.003)	(0.003)	(0.0005)	(0.001)
Parliamentary leader	−0.107*	−0.111*	−0.107*	−0.109*	−0.110*
	(−0.019)	(−0.015)	(−0.015)	(−0.032)	(−0.031)
Party leader	0.127*	0.129*	0.128*	0.131*	0.131*
	(−0.011)	(−0.011)	(−0.010)	(−0.041)	(−0.041)
List placement	0.013*	0.013*	0.013*	0.010*	0.010*
	(−0.002)	(−0.002)	(−0.001)	(−0.003)	(−0.003)
Age	−0.019	−0.019	−0.019	0.011	0.015
	(−0.015)	(−0.015)	(−0.015)	(−0.017)	(−0.017)
Age squared	0.0002	0.0002	0.0002	−0.0001	−0.0002
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Experience	−0.387*	−0.381*	−0.385*	−0.377*	−0.376*
	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)
Experience squared	0.065*	0.066*	0.065*	0.066*	0.066*
	(−0.002)	(−0.002)	(−0.002)	(−0.002)	(−0.002)
N MPs	−0.050*	−0.048*	−0.050*	−0.048*	−0.048*
	(−0.001)	(−0.001)	(−0.001)	(−0.001)	(−0.001)
Opposition party	−0.274*	−0.263*	−0.277*	−0.294*	−0.295*
	(−0.007)	(−0.008)	(−0.007)	(−0.003)	(−0.003)
Support party	−0.447*	−0.448*	−0.445*	−0.447*	−0.443*
	(−0.016)	(−0.016)	(−0.023)	(0.002)	(0.001)
Right wing party	−0.528*	−0.535*	−0.526*	−0.459*	−0.453*
	(−0.011)	(−0.015)	(−0.016)	(−0.010)	(−0.010)
Conservatives	−0.640*	−0.663*	−0.645*	−0.645*	−0.643*
	(−0.026)	(−0.025)	(−0.026)	(−0.017)	(−0.017)
Christian Democrats	−0.419*	−0.564*	−0.459*	−0.581*	−0.594*
	(−0.039)	(−0.030)	(−0.026)	(−0.003)	(−0.002)
Center	−0.177*	−0.276*	−0.195*	−0.295*	−0.301*
	(−0.042)	(−0.034)	(−0.038)	(−0.007)	(−0.006)
Socialist Left	−0.393*	−0.411*	−0.393*	−0.459*	−0.447*
	(−0.055)	(−0.055)	(−0.059)	(−0.037)	(−0.035)
Liberals	0.713*	0.532*	0.667*	0.497*	0.484*
	(−0.042)	(−0.030)	(−0.023)	(−0.044)	(−0.045)
Male	0.016*	0.025*	0.021*	0.054*	0.057*
	(0.001)	(0.001)	(0.00003)	(−0.001)	(−0.001)
2001-2005	−0.592*	−0.611*	−0.598*	−0.614*	−0.616*
	(−0.019)	(−0.018)	(−0.017)	(−0.011)	(−0.011)
2005-2009	−0.566*	−0.574*	−0.570*	−0.556*	−0.557*
	(−0.018)	(−0.016)	(−0.018)	(−0.012)	(−0.012)
2009-2013	−0.553*	−0.584*	−0.566*	−0.579*	−0.583*
	(−0.018)	(−0.017)	(−0.015)	(−0.017)	(−0.017)
2013-2017	−0.503*	−0.498*	−0.507*	−0.494*	−0.497*
	(−0.020)	(−0.022)	(−0.021)	(−0.021)	(−0.020)
Right wing party*Male	−0.078*	−0.066*	−0.076*	−0.168*	−0.176*
	(−0.006)	(−0.002)	(−0.003)	(0.007)	(0.007)
Conservatives*Male	0.050*	0.071*	0.056*	0.043*	0.042*
	(0.004)	(0.003)	(0.003)	(−0.0002)	(−0.0004)
Christian Democrats*Male	0.052*	0.112*	0.070*	0.101*	0.105*
	(0.011)	(0.005)	(0.003)	(−0.002)	(−0.002)
Center*Male	0.033*	0.101*	0.045*	0.098*	0.103*
	(−0.014)	(−0.024)	(−0.021)	(−0.021)	(−0.021)
Socialist Left*Male	0.019	0.062*	0.031*	0.085*	0.091*
	(0.010)	(0.009)	(0.012)	(0.010)	(0.006)
Liberals*Male	−0.582*	−0.551*	−0.571*	−0.544*	−0.544*
	(0.036)	(0.032)	(0.034)	(0.069)	(0.069)
Intercept	1.451*	1.022*	1.343*	0.352	0.241
	(0.412)	(0.434)	(0.456)	(0.486)	(0.486)
Observations	1,942,816	1,942,816	1,942,816	1,942,816	1,942,816

Note: *p<0.05; **p<[0.**]; ***p<[0.***]
* p¡0.05



(a) Number of speeches

(b) Proportion of speeches

(c) Speeches per seat

Figure A-1: Descriptive statistics on nuber of speeches by party and parliamentary period. The Liberal Party (V) is excluded from 2001-2005 because they only occupied two seats.

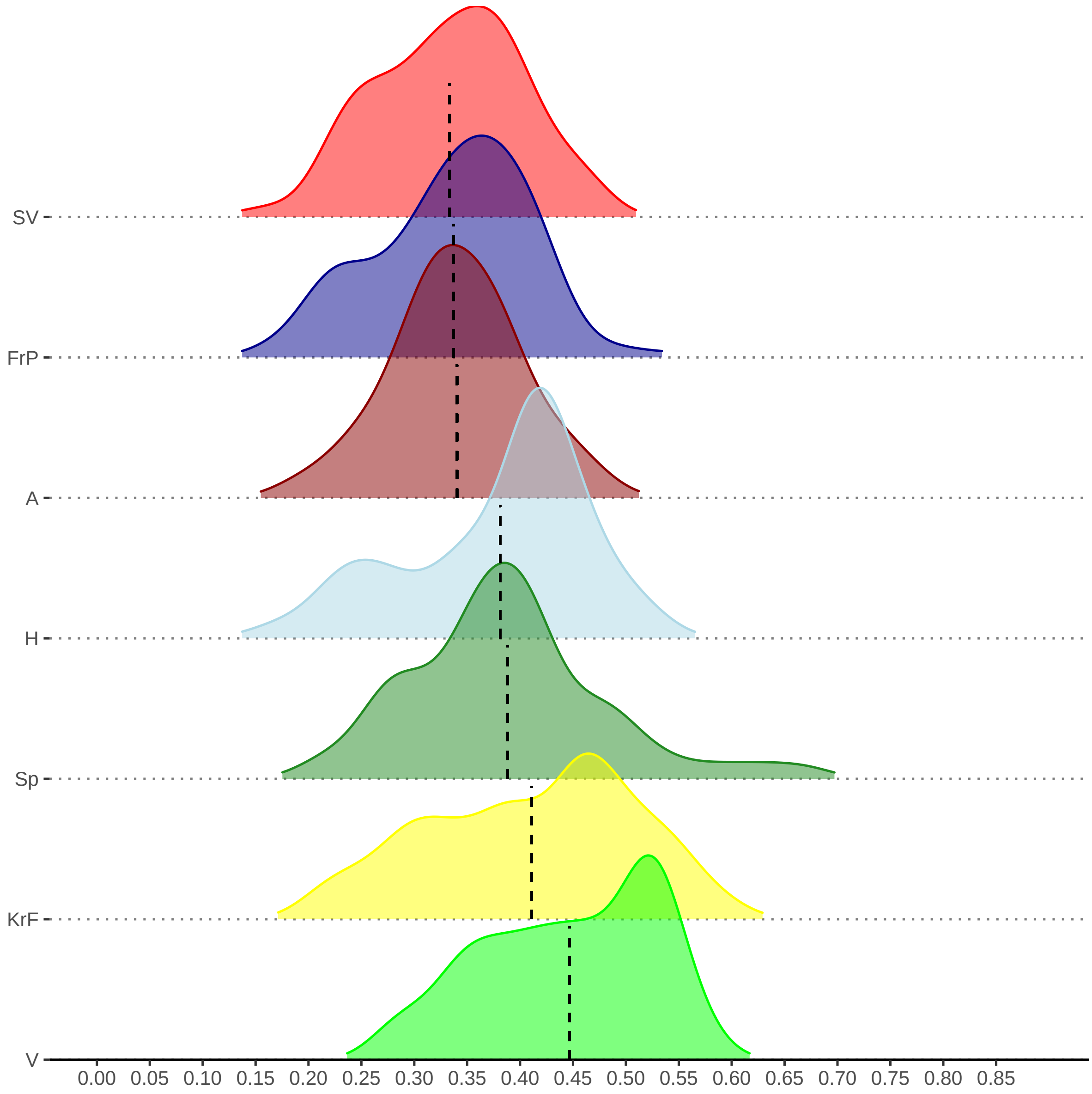
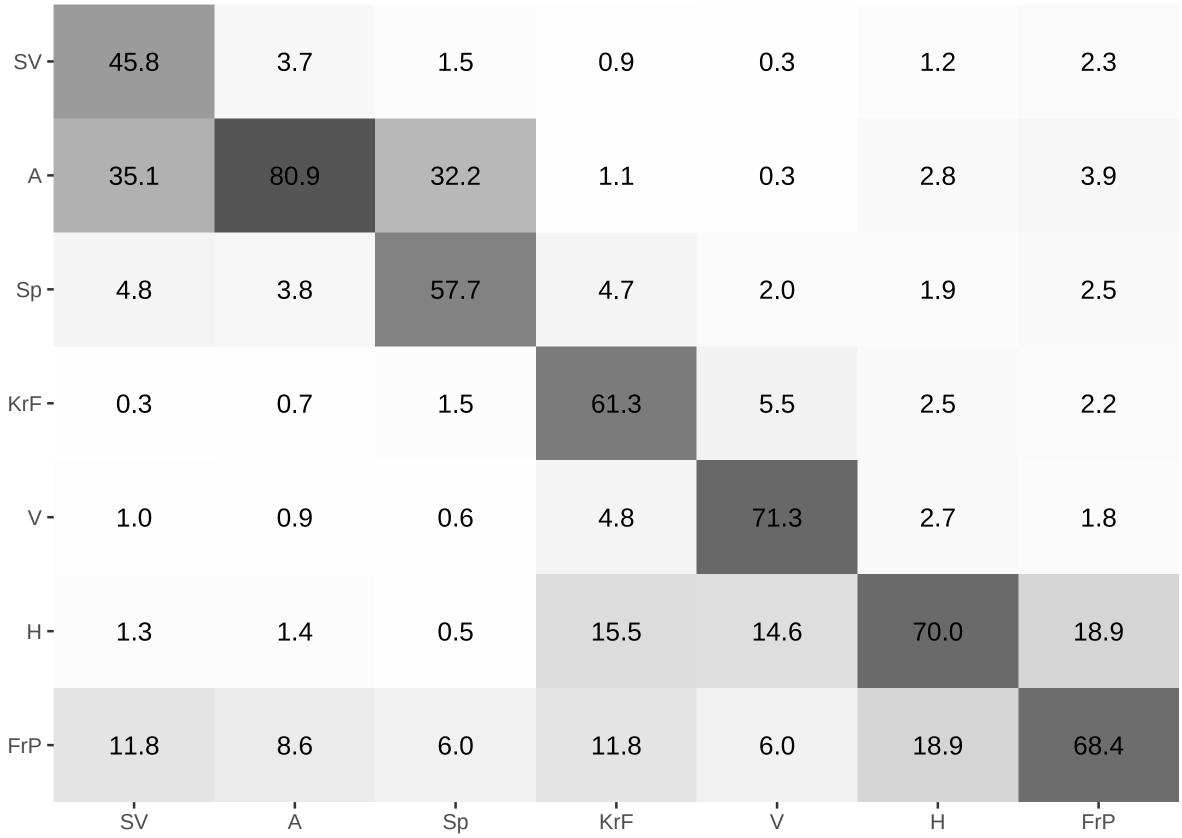
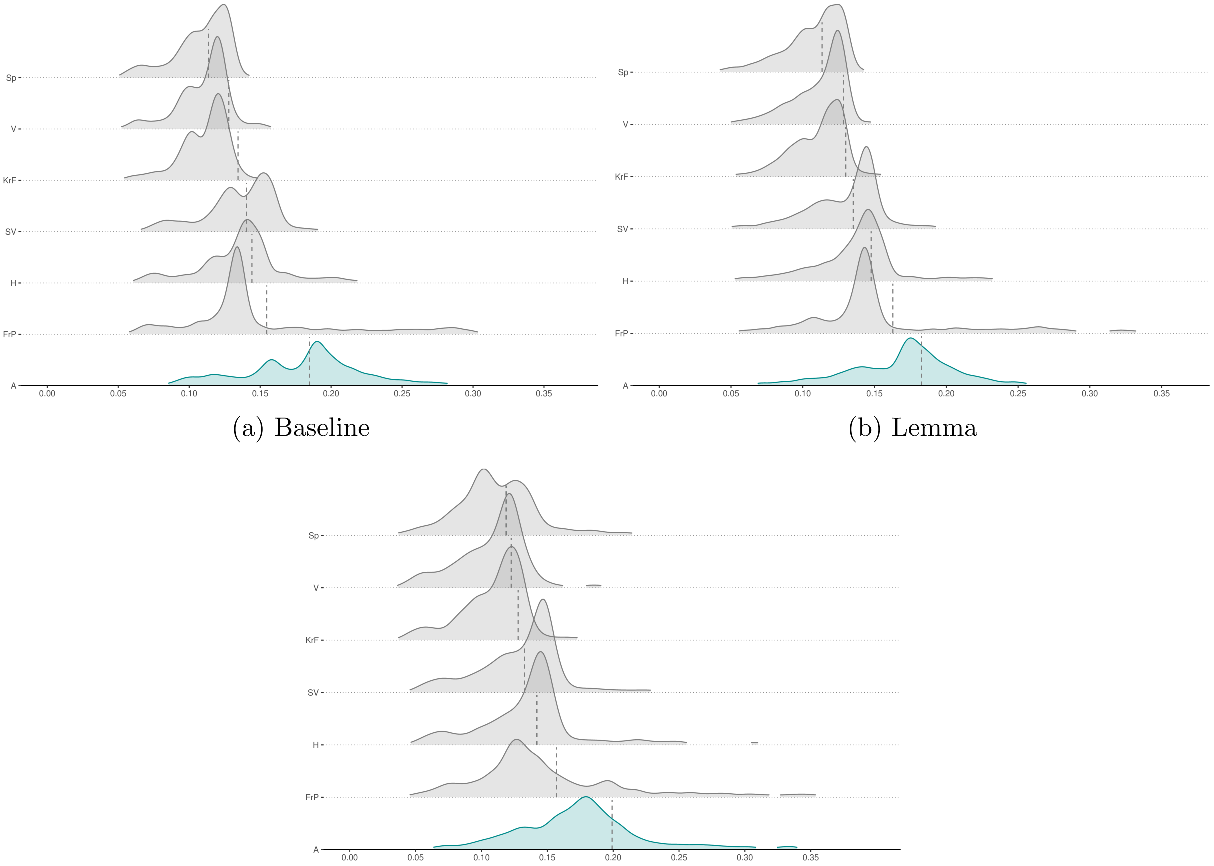


Figure A-2: Probability distribution from SGB classifier for classifying the correct party in the Lemma/Meta/Party role feature set. The vertical line shows the mean probability within each party.



	SV	A	Sp	KrF	V	H	FrP
SV	45.8	3.7	1.5	0.9	0.3	1.2	2.3
A	35.1	80.9	32.2	1.1	0.3	2.8	3.9
Sp	4.8	3.8	57.7	4.7	2.0	1.9	2.5
KrF	0.3	0.7	1.5	61.3	5.5	2.5	2.2
V	1.0	0.9	0.6	4.8	71.3	2.7	1.8
H	1.3	1.4	0.5	15.5	14.6	70.0	18.9
FrP	11.8	8.6	6.0	11.8	6.0	18.9	68.4

Figure A-3: Accuracy of predicting correct party in percent with true party on the y-axis and predicted party on the x-axis.



(a) Baseline

(b) Lemma

(c) Lemma/Meta

Figure A-4: Probability distribution for the Labor Party (A, 2005-2009) from SGB classifiers. True classification is colored in cyan, other classes colored in gray. Parties are ranked, so that the lower a party is on the y-axis, the higher the average probability is that that party is classified as Labor Party.

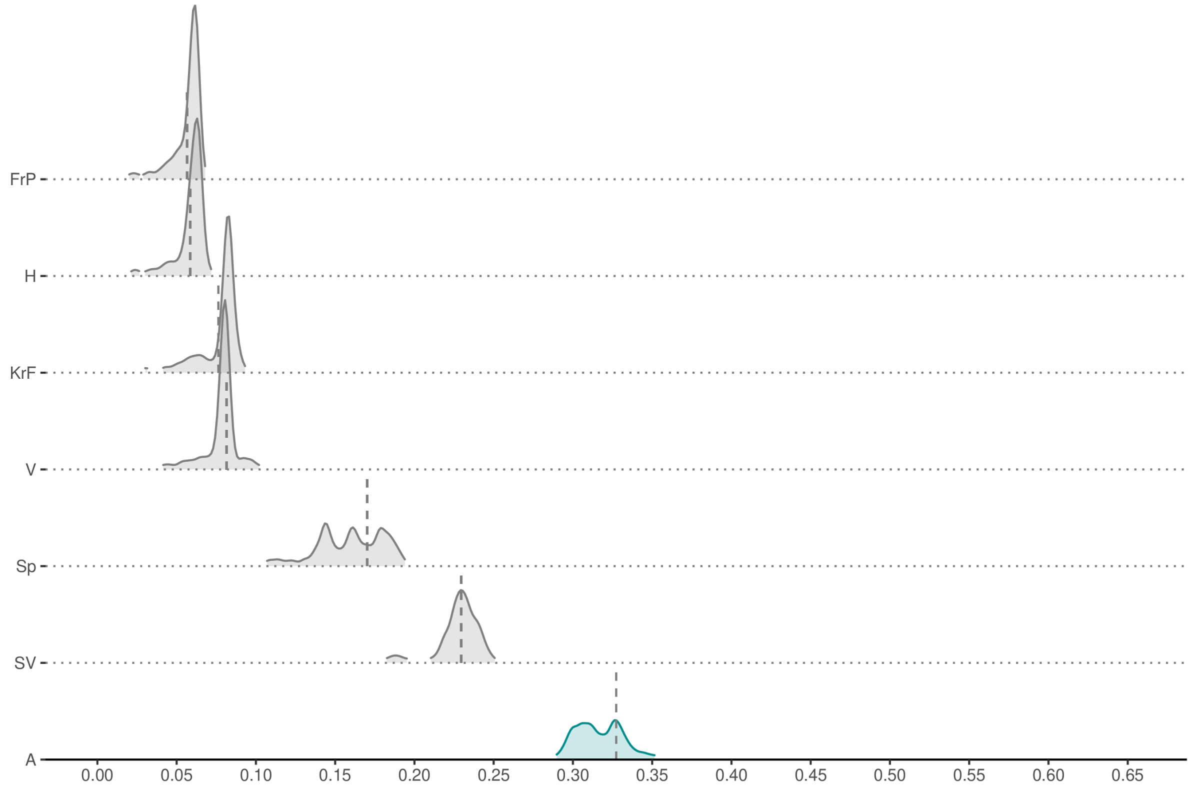
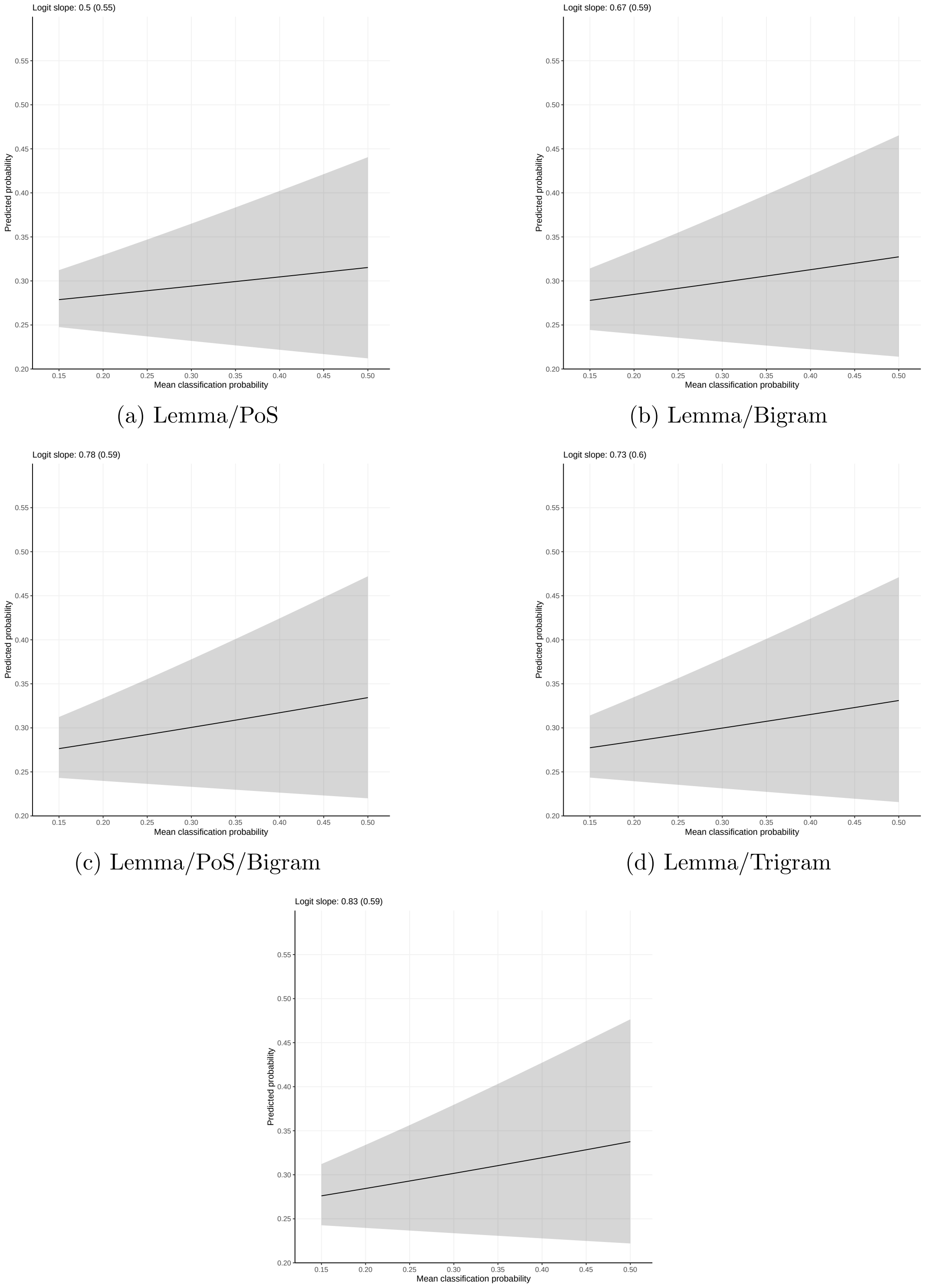


Figure A-5: Accuracy of predicting correct party in percent with true party on the y-axis and predicted party on the x-axis.



(a) Lemma/PoS

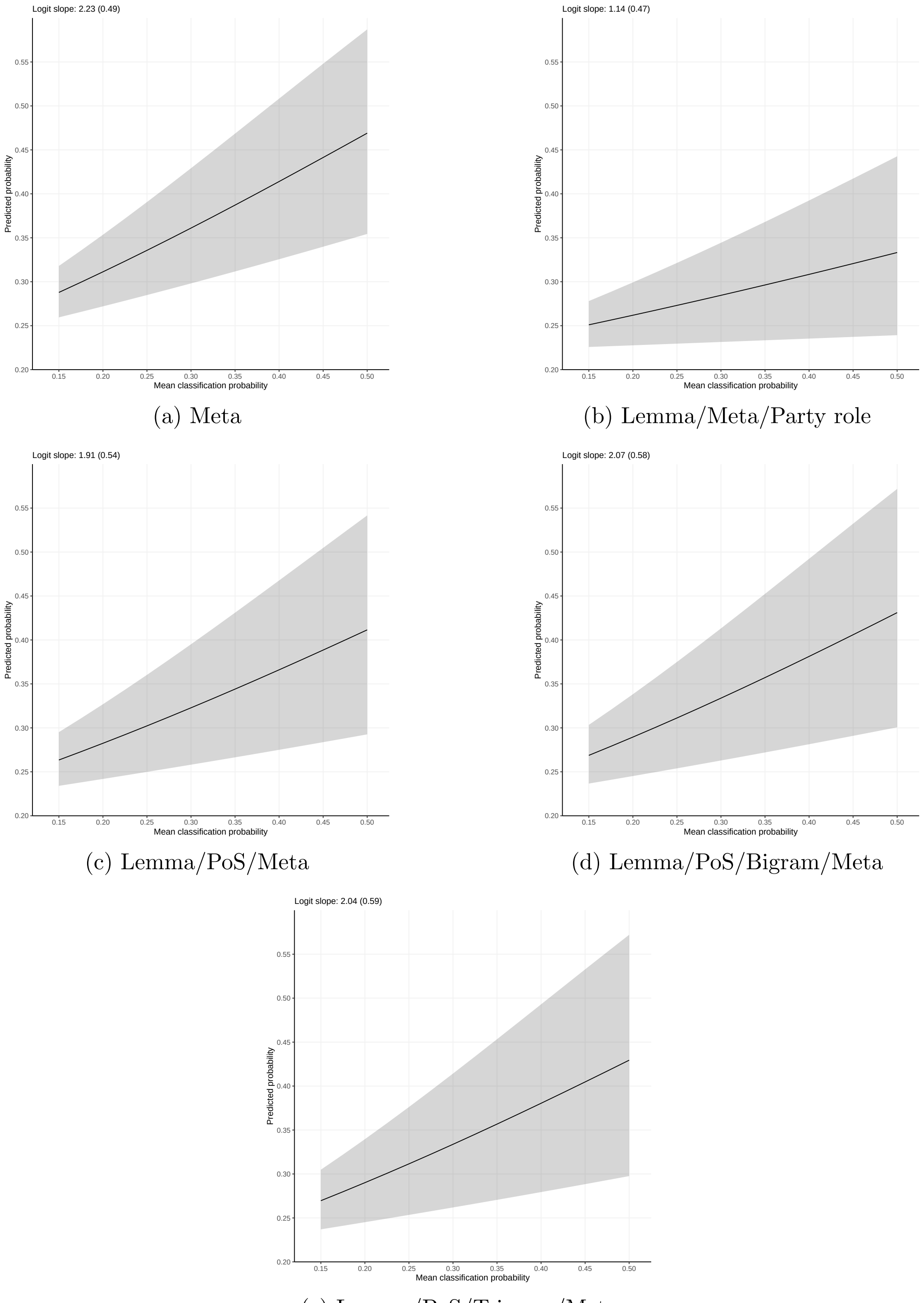
(b) Lemma/Bigram

(c) Lemma/PoS/Bigram

(d) Lemma/Trigram

(e) Lemma/PoS/Bigram

Figure A-6: Replication of Søyland and Høyland (2020) including remaining cohesiveness measures (with no meta data) as an independent variable in the full model.



(a) Meta

(b) Lemma/Meta/Party role

(c) Lemma/PoS/Meta

(d) Lemma/PoS/Bigram/Meta

(e) Lemma/PoS/Trigram/Meta

Figure A-7: Replication of Søyland and Høyland (2020) including remaining cohesiveness measures (with meta data) as an independent variable in the full model.