In recent years, 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; when the classifier is unable to distinguish parties they are closer in policy preferences and vice versa. What has not yet been explored, however, is how this exercise works on multi-party systems, and whether pre-processing decisions affect the validity of this measure. In this paper, we explore how pre-processing decisions can influence these measures in a multi-party system by utilizing a richly annotated dataset of parliamentary speeches in the Norwegian Storting from the 1998-1999 to 2015-2016 session. We build several models, from a bare-bone classifier to classifiers taking more of the language complexity into account. Our findings suggest that using classifier performance as a measure of substantive interest should only be done with great care and consideration. In the Norwegian multi-party system, there is little reason to argue that misclassification exclusively shows traces of party position; the parties normally perceived to be closer together in the ideological space are not mistaken for each other more often than parties further apart.

Introduction

Measuring ideology has long had an important standing in political science because these measures help us describe political change and explain the behavior of political actors. 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 measures of interest based on speech. Most relevant here, text-based automatic party classification performance has recently been used to show trends of polarization between parties in a two-party system (Peterson and Spirling 2017). Here, polarization is defined as the distinctiveness of party labels in a given political system. The underlying assumption is that better classifier performance indicates more polarization and worse
classifier performance indicates lower polarization. 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 whether this measure can be applied to a multi-party setting.

In this paper, we set the logic of using party label classification performance as a measure of polarization to the test in the Norwegian multi-party setting, over different pre-processing feature sets. We find that the measure shows trends across periods do make sense in relation to conventional wisdom about polarization. However, our classifiers have a harder time estimating polarization of party labels between party dyads. Further, we test the sensitivity of the polarization measure with different pre-processing feature sets, showing that feeding the classifier with linguistic features and meta-data variables can affect the interpretation of polarization trends between parties.

Thus, our contribution has two important implications: First, we argue that the underlying assumption of polarization as varying differences in policy positions makes classifier performance a less useful measure of polarization in a multi-party setting. Second, our analyses suggest that we should be careful in weighting computational time, model simplicity, and reproducibility too heavily against predictive power. If we want to produce as precise measures as possible, the latter should not be disregarded in favor of the former, because giving more information to our models can affect the values of the measure. This could, further, lead us to make biased inferences, both in describing political change based on the measures, and in utilizing the measures for explaining political behavior. Last, we argue that this latter point can be generalized to other measures derived from text models. For example, extracting policy position measures from text would demand a highly precise model behind the measures, which again is affected by the information we feed to our model.

Our study builds on debates from the Norwegian parliament (Stortinget) in the period from 1998-2016. Our case selection serves two distinct purposes: First, Norway is a multi-party system with more than two parties consistently occupy a significant amount of seats. This puts the polarization measure to a harder test than in a two-party analysis where all misclassifications only can travel to one party. Second, we are able to test numerous pre-processing specifications with the richly annotated Talk of Norway (ToN) corpus on debates in the Storting (Lapponi and Søyland 2016).

We start with sketching out notable work on party classification in different political systems using text-as-data approaches. The major bulk of this literature is situated in majoritarian systems such as the US, Canada, and the UK, but quantitative party classification based on debates have also been conducted on the multi-party European Parliament. Further, we review some of the literature on measuring polarization, with emphasis on the speech-based approaches. Next, we describe the data (the ToN corpus), the linguistic annotations it contains, and the support vector machine classifier (SVM) used for our analyses, before outlining the different pre-processing sets used to build the SVM models. Finally, we analyze the output of the SVMs, and discuss the generalizability
and implications of our results.

**Classification as polarization?**

**Classifying parties**

The classification of parties based on data, such as roll call votes, has a long-standing tradition in political science. Here, we focus on the studies using speech in legislatures. Naturally, this literature is most developed on majoritarian electoral systems, such as the US (Yu, Kaufmann, and Diermeier 2008; Diermeier et al. 2011), Canada (Hirst, Riabinin, and Graham 2010), and UK (Peterson and Spirling 2017). But, Høyland et al. (2014) also predict party labels from speeches in the multi-party European Parliament. 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.

Yu, Kaufmann, and Diermeier (2008) argue that training an ideology classifier is possible and fairly generalizable based on applying a Support Vector Machine (SVM) and Naive Bayes approach 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 Senate, and apply standard pre-processing procedures, such as stemming, sparse term reduction, and removal of stopwords, but they also use parts of speech tagging for extracting different feature sets. 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, Riabinin, and Graham (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* and *defense*, rather than ideology, by showing the discriminating
words in their model. Second, Hirst, Riabinin, and Graham (2010, 737), in order to more
directly test the first findings, show that 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% 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, Riabinin, and Graham (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, Riabinin, and Graham
(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 of polarization between
parties 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.

The latter reflection raises a more general question on the evaluation of political party
classifiers – what is a reasonable upper bound for performance? A general approach in
Natural Language Processing (NLP) is to look at the performance of human annotators,
and consider factors such inter-annotator agreement: annotators might disagree when
annotating a piece of information, and low agreement scores might suggest that the
task in question is very complex or controversial, with classes lacking clearly separating
characteristics. To our knowledge, there is no such annotation study in the context of
parliamentary proceedings, making it hard to assess whether a system guessing the correct
party label for a speech 50% of the times (or 90, depending on the number of parties, the
majority class and the health condition of the democracy in question) is a good or a bad classifier. As will be discussed below, however, the absolute accuracy of a classifier is not necessarily believed to affect the measure of polarization.

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 our arguments, we focus on the exercise of using classifier precision as a measure of polarization, mainly leaning on the paper by Peterson and Spirling (2017). In the following section, we summarize some of the work on measuring polarization from different sources of data stemming from parliamentary debates.

**Polarization**

For simplicity, we define political polarization in similar terms as Peterson and Spirling (2017): the level of polarization is determined by how easy it is to distinguish parties from one another, which means that the easier this is at a given time, the more polarized the system is. This stems from a long tradition of studying polarization in the American Congress, where individual roll call votes often have been used to show how divided parties are (Poole and Rosenthal 1984; Clinton, Jackman, and Rivers 2004; McCarty, Poole, and Rosenthal 2006; Garand 2010).

Roll call votes have, however, shown to be less informative in parliamentary democracies, 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, given our polarization definition, the nearly perfect separation of parties in roll call votes would also indicate nearly complete polarization between parties, which further goes against the main view of some multi-party systems as parliamentary systems where parties often cooperate on producing policy (Lijphart 2012). Thus, roll calls can be problematic when investigating polarization in parliamentary democracies. A possible objection against using parliamentary speech as an alternative source is that, whereas votes actually matter directly for policy output, speech have little importance for producing policy. Although we find this objection plausible, we follow the logic of Gentzkow, Shapiro, and Taddy (2016, 23) that the use of external speech-writers is “an investment that only makes sense if language matters”.

Underlying for the concept of polarization is the large literature on party positions in a n-dimensional spaces: Following our definition of polarization, the farther apart parties are in a given policy space, the more polarized they are. 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 political polarization between parties relies on the concept of parties’ policy position. Indeed, Peterson and Spirling (2017) argue that polarization is the “[...] difference between the positions of the two main parties that have held Prime Ministerial office in modern times”.

Peterson and Spirling (2017) 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.

Thus, the main interest is not to maximize classification precision, but rather the relative difference in classifier performance over time. We also note the work of Gentzkow, Shapiro, and Taddy (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.

As outlined above, the studies using speech for measuring polarization has mainly operated within majoritarian party systems with two classes for classification. In the analyses below, we show that (1) the underlying assumption of using classification performance as a polarization measure – that it says something about positional differences between parties – seems to be violated in the Norwegian multi-party setting, and (2) pre-processing decisions matter for how polarization between parties is mapped.

**Data and methods**

**Data**

Our data builds on the Talk of Norway (ToN) corpus of parliamentary speeches from the Storting in the period from 1998 to 2016 (Lapponi and Søyland 2016). The unprocessed data frame consist of 250,373 speeches over 99 variables, including speaker characteristics,
party attributes, institutional variables, the text of the speech, date, time, and more. In our analyses, we exclude speeches from the parliamentary President and speeches held by MPs from parties that are not represented through all parliamentary periods. This gives us a subset of 152405 speeches.

<table>
<thead>
<tr>
<th>Party</th>
<th># tokens</th>
<th># sentences</th>
<th>tokens/sentences</th>
<th>% nynorsk</th>
</tr>
</thead>
<tbody>
<tr>
<td>SV</td>
<td>7064169</td>
<td>315815</td>
<td>22.37</td>
<td>18.34</td>
</tr>
<tr>
<td>A</td>
<td>15596225</td>
<td>731024</td>
<td>21.33</td>
<td>9.43</td>
</tr>
<tr>
<td>Sp</td>
<td>5724137</td>
<td>265312</td>
<td>21.58</td>
<td>34.02</td>
</tr>
<tr>
<td>KrF</td>
<td>6498372</td>
<td>309307</td>
<td>21.01</td>
<td>21.05</td>
</tr>
<tr>
<td>V</td>
<td>3703190</td>
<td>164168</td>
<td>22.56</td>
<td>8.84</td>
</tr>
<tr>
<td>H</td>
<td>11203083</td>
<td>502470</td>
<td>22.30</td>
<td>2.60</td>
</tr>
<tr>
<td>FrP</td>
<td>9363542</td>
<td>429172</td>
<td>21.82</td>
<td>5.87</td>
</tr>
<tr>
<td>Total</td>
<td>59152718</td>
<td>2717268</td>
<td>21.85</td>
<td>14.31</td>
</tr>
</tbody>
</table>

Table 1 shows the number of tokens, sentences, and the percentage of speeches in nynorsk over all parties before the pre-processing. In total, the corpus subset consist of over 59 million tokens, where the Labor Party (A) has the most tokens by a fair margin, in front of the Conservatives (H). However, tokens per sentence is very similar across parties; there does not seem to be systematic differences in the relative amount of speech between parties. As there are two official languages in Norway, we have also included an indicator in the main ToN frame for whether the speech was held in Bokmål or Nynorsk. Table 1 shows that the Center Party (Sp) has the highest percentage of nynorsk speeches, with well over double the average.

Figures 6, 7, and 8 in the appendix also show number of speeches, proportion of speeches, and speeches per seat across parliamentary periods and parties, respectively. Here we note that, although the Labor Party is the largest class in our sample, 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). Here, all speeches of the corpus are split into individual speech files in a CoNLL-like formatted tab separated

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

**Classifier**

For each of the pre-processing feature sets described below, we train a Support Vector Machine (SVM) classifier to learn a function that maps instance vectors to party labels. The SVM is a vector-space-based supervised machine-learning method that optimizes a classification function by finding a decision boundary between classes with maximal distance from any point in the training data, save for those that the algorithm deems to be “outliers”. We use the Linear SVM implementation available through the Python package Scikit Learn (Pedregosa et al. 2011). We tune the classifier across an exponential range of C-parameter values, and let the classifier assign class weights (higher or lower C parameter values) as a function of the size of the training sample set for each class.

We use a 10-fold cross-validation method, where we split the data into 10 equally large samples, train the model on 9 of these sets (training set) and predict party label on the remaining set (development test set or dev-test set). We then switch the dev-test set out for one of the sets in the training set, and follow the same procedure. This is done for all 10 folds, so that we have party label prediction for the full dataset.

**Pre-processing**

Importantly, there are a myriad of decisions to take in the process of going from text to numbers and making the data ready for analyses. We will highlight the differences between our models in this section. As pointed out by Denny and Spirling (2017), each pre-processing decision is a binary choice and there are a vast amount of decisions. For example, Denny and Spirling (2017) focus on seven common pre-processing 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 our argument as precise as possible, we hold a number of pre-

---


3 For reading the tagged ToN speeches into R, an under development package can be found at https://github.com/martigso/tonR.

4 The C parameter governs the trade-off between training error and margin size between classes.
processing decisions constant across all models. First, we 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; the intuition behind is that the features appearing across many documents are going to have less disambiguating potential than those that do not. Second, we 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. We opt for this solution because the stop word dictionaries available in Norwegian are somewhat limited. Third, we balance the classes in our folds 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, we remove speeches with less than 100 tokens. All these pre-processing steps are done for the five models we outline below.

Baseline

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

First, we lowercase all tokens in order to not differentiate between same tokens in the start of a sentence and later in the sentence. Second, we remove numbers and punctuation. Third, we split the speeches into tokens by using the tokenizers package for R, which strips all whitespace and punctuations and returns a vector of tokens based on it (Mullen 2016). Fourth, we 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 the same word in different forms is still the same words, and should denote similarity rather than difference.

Lemma

The second feature set 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, Raghavan, and Schütze (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 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, depending on the context, a “party” can denote both politics and celebration).

**Part of speech**

Our third feature set also includes the 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).

Table 2 shows the descriptive statistics of some selected PoS tags in the ToN corpus over the speeches used in our analyses. Some of the PoS tags that are present in the corpus have been removed here; punctuation (commas, quotes, etc), for example, are not included in any of the feature sets used in the analyses. Unsurprising, the main word categories of the PoS tags are adjectives, prepositions, nouns, and verbs.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>N</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td># adjectives</td>
<td>152,405</td>
<td>37.911</td>
<td>41.193</td>
<td>0</td>
<td>1,406</td>
</tr>
<tr>
<td># adverbs</td>
<td>152,405</td>
<td>18.552</td>
<td>16.565</td>
<td>0</td>
<td>458</td>
</tr>
<tr>
<td># conjunction clause boundries</td>
<td>152,405</td>
<td>21.738</td>
<td>21.160</td>
<td>1</td>
<td>595</td>
</tr>
<tr>
<td># commas</td>
<td>152,405</td>
<td>10.616</td>
<td>10.485</td>
<td>0</td>
<td>343</td>
</tr>
<tr>
<td># conjunctions</td>
<td>152,405</td>
<td>13.922</td>
<td>14.918</td>
<td>0</td>
<td>450</td>
</tr>
<tr>
<td># determiners</td>
<td>152,405</td>
<td>26.158</td>
<td>26.010</td>
<td>0</td>
<td>751</td>
</tr>
<tr>
<td># infinitive marker</td>
<td>152,405</td>
<td>6.864</td>
<td>7.139</td>
<td>0</td>
<td>192</td>
</tr>
<tr>
<td># interjections</td>
<td>152,405</td>
<td>0.238</td>
<td>0.700</td>
<td>0</td>
<td>31</td>
</tr>
<tr>
<td># prepositions</td>
<td>152,405</td>
<td>48.841</td>
<td>47.914</td>
<td>4</td>
<td>1,396</td>
</tr>
<tr>
<td># pronouns</td>
<td>152,405</td>
<td>30.932</td>
<td>26.426</td>
<td>0</td>
<td>745</td>
</tr>
<tr>
<td># subjunctions</td>
<td>152,405</td>
<td>19.130</td>
<td>16.114</td>
<td>0</td>
<td>376</td>
</tr>
<tr>
<td># nouns</td>
<td>152,405</td>
<td>80.802</td>
<td>84.790</td>
<td>8</td>
<td>2,665</td>
</tr>
<tr>
<td># verbs</td>
<td>152,405</td>
<td>69.208</td>
<td>61.605</td>
<td>6</td>
<td>1,617</td>
</tr>
</tbody>
</table>

**Table 2** Descriptive statistics of the speeches in the corpus, with sentences, tokens, and some selected PoS tags.
N-grams

As for the fourth feature set, we supplement the token unigrams from the previous models with token and lemma bigrams and trigrams. 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 \textit{gift}, for example, can mean both “married” and “poison”. Thus, it is important to know the context of the word \textit{gift} in order to understand the meaning of the word. From the ToN corpus, we can exemplify with the phrases \textit{ulikhet er gift} (“inequality is poisonous”) and \textit{lykkelig som gift} (“happily married”), where \textit{gift} would be the same token with unigrams, but very different with token trigrams.

Metadata

In our last model, we also feed a set of variables to the SVMs – similar to including controls in a regression analysis. The variables included are both at the speaker and debate level: gender and county of provenance belonging 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 (general issue, budget, law) are the debate level attributes.

This is 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

In this section, we show and discuss the overall performance differences between our six models. This makes us able to test whether the classification performance is significantly affected by our different pre-processing decisions. We opt for using \( F_1 \) scores in comparing the performance of our models, because it rewards the model for not producing both false

\(^5\)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”).
Figure 1. F_{1} scores on the full dataset. Points represent the F_{1} score for the parties on the x-axis and the dashed lines the macro F_{1} scores for all models. The models are differentiated according to the color of the points and lines.

As for differences between parties, Figure 1 indicates that the baseline model do a worse job of classifying the party with the least amount of speeches: the Liberal Party (V). Further, the model has a harder time correctly classifying the Christian People’s positives and false negatives.\(^6\)

\(^6\)The F_{1} score is calculated as: F_{1} = \frac{2PR}{P+R}, where P is precision and R is recall. Precision is defined as: P = \frac{true positives}{true positives+false positives}, and recall as R = \frac{true positives}{true positives+false negatives}.
Party (KrF), whereas it performs better on the right-wing Progress Party (FrP). More or less, the same pattern emerge from the other models. Importantly, however, the intra-model differences between parties do decrease substantially in all models compared to the baseline; the variance between party $F_1$ scores is six times higher in the baseline model than in the meta model. This in itself does suggest that, at least, parts of the misclassifications we do are not due to differences between the ideologies of parties, but rather omitted variable bias; giving more information to our classifier reduces the variance in wrong predictions between parties. Of course, some of the variance explained in the higher level models could include ideology. This will be discussed further below.

We are also interested in whether the difference of performance between our feature sets is not produced by chance. A common way to do this is by bootstrapping our $F_1$ score estimates (Jurafsky and Martin 2016, 15). Figure 2 shows the bootstrapped macro $F_1$ scores for all models. With this method, we 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. The points in Figure 2 show the mean point estimate score over all simulations, whereas the horizontal dashed lines show the lower and upper quantiles (confidence intervals). Note that the point estimate is very close to the corresponding $F_1$ score on the complete sample, marked by the dashed lines in Figure 1.

What first sticks out with Figure 2 is, again, the significant performance increase of the higher level models compared to the baseline model. There is, however, only a slight
increase in model performance when we include part-of-speech tags, compared to the token lemma model. Including both n-grams and metadata do, nevertheless, increase model precision with a fairly large margin.

In sum, the initial impression is (1) that feeding our classifier with more information does significantly increase its performance, although some feature sets increase accuracy more than others, and (2) that this precision increase truncates the variation in classification between parties. In the next section we delve deeper into the main issue at hand: using misclassification as a measure of polarization. We do this in two steps: first, we study the longitudinal trends of misclassification between parliamentary periods. Second, we analyze where the misclassifications go by mapping false positives and true negatives from our models. Following Peterson and Spirling (2017), the main intuition for using classifier performance as a measure of polarization does not hinge on how good the classification model is, but rather the relative difference in performance over time. We thus restrict the following analyses to our meta and baseline specifications.

Polarization

The multi-party setting of Norway gives us a unique opportunity to explore where the misclassifications go; in analyses of only two parties, the misclassifications can only go to the other party. For the remainder of this section, low polarization will be treated as a synonym of low classification performance (more misclassifications), and high polarization as high classification performance (fewer misclassifications).

Longitudinal trends

Figure 3 shows bootstrapped $F_1$ scores (same method as described above, but now for the subsamples in each panel) for the baseline and meta models over the parliamentary periods in our sample, with the points and corresponding confidence intervals showing party scores (but, the confidence intervals are often smaller than the area covered by the point, and is therefore not visible), and the panel-wide gray dashed lines showing the macro bootstrapped period specific score with upper and lower confidence intervals.

For the general trend of both models, we see that polarization is, on the one hand, at its lowest under the minority government periods of Bondevik I (1997-2001), Stoltenberg I (1997-2001), Bondevik II (2001-2005), and Solberg I (2013-2017). On the other hand, polarization is at its highest under the majority governments – Stoltenberg II and III. All the periods, except for the 2005-2009 and 2009-2013 also see a significant (confidence intervals not crossing with neighboring periods) change in polarization. These results do not seem unreasonable as minority governments in Norway often construct majorities with different parties on different issues, and majorities do not rely on the opposition for producing policy.
Consequently, we would expect the opposition to be more interested in distancing themselves from the government parties during majorities than in minorities, where they rely on the opposition for producing policies. We would also expect misclassification to be higher during coalitions because cooperating parties should communicate more similar, but our sample only has one short-lived single-party cabinet (Stoltenberg I) in the 1997-2001 period, making comparisons difficult.

Although the baseline and meta models show similar trends with respect to the period macro $F_1$ scores, we note that the scores, as expected, are much lower across all periods in the baseline specification. The main relative differences between the models is in the party specific $F_1$ scores: the baseline model again gives much higher intra-model variation in scores between the parties than the meta model. Further, the intra-model party scores relative to the period macro scores are not consistent across the two models. For example, the Liberal Party (V) scores significantly lower than the macro $F_1$ score for the 2009-2013 period.

**Figure 3.** Baseline and meta model $F_1$ scores for all parties over parliamentary periods. The horizontal darker lines through the plot shows the mean $F_1$ score for the relevant period, with the lighter horizontal lines showing the lower and upper confidence intervals for this measure.
period with the baseline specification, whereas they score very close to the macro score in the same period for the meta specification. The picture is pretty similar for the other periods as well; with the models producing somewhat different variation between parties – similarly to what was shown in the previous section. The Progress Party (FrP), however, scores consistently higher than the macro $F_1$ scores for all periods and models except the 2013-2017 period, where they participate in cabinet for the first time.

In sum, the general picture, with the period specific macro $F_1$ scores, suggest that the polarization measure is plausible; the relative changes in polarization across parliamentary periods make sense according to conventional wisdom about how majorities and minorities cooperate with the opposition. Here, both the baseline and meta models show the same trends. What is different between the two models, however, is the party-wise $F_1$ scores and the intra-model variation between them. In the next section, we explore what parties are mistaken for which parties, under different policy dimensions and institutional settings. Our intuition, with the assumption of spatial model in mind, is that parties closer to one another are more prone to receiving misclassifications from one another.

**Party dyads**

In this section, we investigate how misclassifications travel between parties by using Sankey diagrams from the R package `riverplot` (Weiner 2017). Our presentation of these diagrams are exclusively dedicated to false positives and false negatives (disregarding true positives). We restrict this analysis to the meta specification in order to retain consistency. We do, however, also provide twin diagrams with the baseline specification in the appendix.

The diagrams show the true party label on the left side, with bands going to the party the true label was misclassified as. The bandwidth indicates the proportion of all classifications from one party that goes to another party – the label on the far left shows the percentage of speeches within each party that has been misclassified. Both the left and right side of the figures are order according to the policy dimension under investigation, with the most right-sided party on the top and the most left-sided party on the bottom. Thus, if misclassifications are to indicate polarization, we should expect the top party to concede more classifications to the second party than the third, more to the third than the fourth, and so on. Figure 10 in the appendix show Sankey diagrams for the full dataset, with the parties ordered by the traditional socio-economic left-right dimension. Figure 4 shows the misclassifications of party labels in the meta model under debates on (a) immigration, (b) the Norwegian church, and (c) European Union/EFTA.

In (a) immigration debates, the Labor Party (A) receives most false positives and the Liberal (V) and Center Party (Sp) the least, indicating that smaller parties get fewer false positives than larger parties. Further, the model shows mixed signs of misclassifying parties that are perceived to have closer ideological preferences in the immigration debate. Typically, Norwegian debates have produced the largest differences in opinion between the Progress Party (FrP) and the others, with FrP being the most negative and the Socialist
(a) Immigration debates (N = 353/1446)

(b) Norwegian Church debates (N = 356/1278)

(c) European Union/EFTA debates (N = 896/3301)

Figure 4. Sankey diagram of misclassification in (a) immigration, (b) Norwegian Church, and (c) European Union/EFTA debates. The left sides show true party label with the misclassification count and total amount of speeches (false negatives / N) to the left of the party label. The right side shows false party labels. The band between the left and right shows magnitude of true labels misclassified as the corresponding false label.
left Party (SV) being the most positive to immigration (Aardal 2011, 99-101). This is also reflected in the amount of false negatives in FrP (15.3%), which is far lower than the other parties. We do, nevertheless, see that the second largest set of misclassifications from SV go to FrP – the two parties with the greatest distance between them on this issue. Reversely, SV only receives a small chunk of FrP’s speeches. It is also interesting that FrP receives the second largest amount of false positives, although this does not lend itself to conventional wisdom about the policy position on immigration in Norway either. Both this and the band between SV and FrP could indicate that the story here is more about policy salience (how important immigration is for the parties), rather than policy positions. As for the difference between models, Figure 9 in the appendix shows (a) immigration debates with the baseline specification; this specification does not seem to show neither more or less patterns of positionality than with the meta specification.

Panel (b) in Figure 4 shows misclassifications under debates on the Norwegian church. Naturally, the Christian People’s Party (KrF) has been the least secular party on this issue, whereas the Socialist Left Party (SV) the most secular, the Center Party (Sp) and Progress Party (FrP) closest to KrF, and the Labor and Liberal Party leaning slightly over the the secular side (105-106). The picture in church debates is very similar to that of immigration: the least secular KrF is mistaken for the Labor Party more often than the other parties, V is still the party receiving the least amount of misclassifications although KrF is the furthest apart from the other parties on this dimension, and the larger parties generally receive more false positives than smaller parties. There are, however, traces of positionality in the misclassifications here as well. For example, excluding the Liberal Party (V), the false negatives from the Socialist Left Party (SV) follow the expected pattern by losing more speeches to the Labor Party (A) than the Conservatives (H), more to H than the Progress Party (FrP), and so on. The same pattern can be seen, to different degrees, for all parties except KrF. The same patterns do not seem to emerge with the baseline specification, shown by Figure 9 in the appendix. Here, KrF gets the largest set of false positives, which indicates that this model picks up policy salience; when parties talk about the Norwegian Church, they are more likely to be classified as the most salient party (KrF), rather than the party closest in position.

Finally, panel (c) of Figure 4 displays misclassifications from debates on the European Union and the European Free Trade Association (EFTA) in which Norway is a part of. The discussion over EU membership and the agreement between EFTA countries and the EU on access to its free marked has been a long and polarized debate in Norwegian politics. Traditionally, the Center Party (Sp) has been an opponent of both, whereas the Labor (A), Conservative (H), and Liberal (V) parties have been in favor of either one or both (Aardal and Bergh 2015, 271). Some of the parties are more fluctuant on this dimension than the two discussed above, both in terms of factions within parties having divergent positions (the Labor Party (A) is an example of this), and change their official stance (for example, the Progress Party). Hence, this policy area is slightly harder to interpret. On the one hand, Figure 4 shows that the two main misclassification veins go between the
Labor Party (A) and the Conservative Party (H), who have generally been among the most positive parties to European integration. Also, the Center Party (Sp) is mistaken for FrP more often than both the Christian Democrats (KrF) and Socialist Left Party (SV). On the other hand, Sp is confused more with A than any other party, and V exchange few misclassification with H and A. Thus, as with the two other panels, there seem to be traces of positionality here, but not exclusively. The baseline specification (panel (c) of Figure 9 in the appendix) does not seem to catch policy positions any better. Indeed, the bands A and H are relatively much smaller here.

In sum, the story of larger parties receiving more false positives, discussed above, is still the main story after looking at specific policy areas. We have shown that there are also traces of policy positions here, but they are mixed with variance that does not make sense and possibly some policy salience features. Consequently, this puts a dent in the prospect of using classification as a measure of polarization; the assumption of positional closeness is not clearly met in our models. At the very least, our best performing meta model does seem to produce errors that more often can be traced to positionality than the baseline model.

A possible different story behind the misclassification positions could be that they contain information about the placement between potential coalition parties. For example, the Center Party (Sp) “switched sides” from being center-right to center-left during the period we cover; Sp were part of all the center-right coalitions from the end of World War II to the fall of the Bondevik I cabinet in 2000. However, they went into the 2005 election as an official member of the center-left “red-green” coalition, which ultimately saw them forming a majority cabinet together with the Labor Party (A) and Socialist Left Party (SV). We should thus expect Sp to be more similar to the left parties after 2005 than before. Figure 5 shows misclassification from the meta model between parties during the (a) Bondevik I cabinet, a coalition between KrF, V, and Sp, and the (b) second “red-green” period (2009-2013) under Prime Minister Jens Stoltenberg. At face value, the bands from Sp to A seem to be larger under the Bondevik I cabinet. However, Sp’s accuracy for this period is much lower (48%) compared to the Stoltenberk III period (21%). The bands from Sp to both SV and A in Figure 5 are, in relative terms, larger during the Stoltenberg III period, whereas the bands are pretty thick between Sp, KrF and V under the Bondevik I cabinet and reduced to marginal errors for the Stoltenberg III period. Further, it is encouraging to see that a larger share of misclassifications from SV go to A when they cooperate in the Stoltenberg III coalition than under the Bondevik I cabinet, where both were opposition parties. A similar pattern is seen if we focus on the right side (false positives): the share of misclassifications Sp receive from A and SV are somewhat larger when they are coalition partners (Stoltenberg III) than when they are in cabinet with KrF and V (Bondevik I). Last, as discussed above, the model is much more precise under the majority Stoltenberg III cabinet (23% misclassifications) than the minority Bondevik I cabinet (34% misclassifications); the political horse trading of minority periods are also reflected in this graph.
Figure 5. Sankey diagram of misclassification from debates under (a) the Bondevik I and (b) Stoltenberg III cabinets based on the meta model. The left sides show true party label, the right side shows false party label, and the band between the magnitude of true labels misclassified as the corresponding false label.
In sum, the party dyad analyses do show traces of positionality – especially with
regards to coalition partners – in the misclassifications. However, we have also shown
that the larger a party is, the more misclassifications (false positives) it gets, and that only
parts of the unexplained variance in our models are produced by positionality. We are
thus hesitant to conclude that misclassification serves as a good measure of polarization
on the party level for our case. Nevertheless, we find that the more complex model (meta)
overcome some of the shortcomings found in our most basic model (baseline), which is
evidence that better models are also a better point of departure for further research and
experimentation in this setting.

Discussion

Producing substantive measures on policy positions is an important task in the field of
cumparative politics. Such measures are used not only to describe political change in
democracies, but also used for explaining behavior in 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, we have explored how one technique for measuring polarization works
in a multi-party setting. By utilizing a unique dataset on speeches in the Norwegian
parliament, our analyses of classification based on different pre-processing feature sets
show sensible results on the macro level between parliamentary periods, but varying
results at the party level. This does suggest that the pre-processing decisions we make can
affect how we perceive polarization between parties. Further, by analyzing what parties
are mistaken for each other, we show that party misclassifications mainly travel in the
direction of the larger parties, even on specific policy dimensions.

In strict terms, if polarization is realized through misclassification of party labels in
analyses of text, we have to assume that the residuals in our classification models are
indicators of two parties sharing preferences on a given issue. 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, these residuals only
show traces of positionality; in some configurations, the more complex models seem to
have explained both parts of policy positions, policy salience, and coalition patterns. We
thus argue, that using model residuals as meaningful measures about political systems is
an optimistic approach. Current approaches are unable to differentiate between what is
omitted variable bias and what is the subject of interest; residuals are, after all, unexplained
variance, and policy positions (as a measure) is inherently unobservable.

We also argue that our findings could go beyond the Norwegian case, in that we would
expect to see similar patterns 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, our analyses show that the unexplained variance in classifications are not exclusively driven by party positions, and we urge researchers to take this into consideration when utilizing such measures for substantive tests.

References


References


Jurafsky, Daniel, and James H. Martin. 2016. “Naive Bayes and Sentiment Classification.” Chap. 6 in Speech and Language Processing, edited by 3. Online draft.


Figure 6. Number of speeches over parliamentary periods and parties.

Figure 7. Proportion of speeches within parliamentary periods over parties.
Figure 8. Number of speeches per seat within parliamentary periods over parties.
Figure 9. Sankey diagram of misclassification in (a) immigration, (b) Norwegian Church, and (c) European Union/EFTA debates. The left sides show true party label with the misclassification count and total amount of speeches (false negatives / N) to the left of the party label. The right side shows false party labels. The band between the left and right shows magnitude of true labels misclassified as the corresponding false label.
Figure 10. Sankey diagram of misclassifications for the full sample with the (a) baseline and (b) meta feature sets. The left sides show true party label, the right side shows false party label, and the band between the magnitude of true labels misclassified as the corresponding false label.