Open access • Journal Article • DOI:10.2139/SSRN.3782027

Comparing Automated Content Analysis Methods To Distinguish Issue Communication by Political Parties on Twitter — Source link

Stiene Praet, Peter Van Aelst, Walter Daelemans, Tim Kreutz ...+3 more authors

Institutions: University of Antwerp

Published on: 08 Feb 2021 - Social Science Research Network

Topics: Content analysis

Comparing automated content analysis methods to distinguish issue communication by political parties on Twitter

Reference:
Praet Stiene, Van Aelst Peter, Daelemans Walter, Kreutz Tim, Peeters Jeroen, Walgrave Stefaan, Martens David.- Comparing automated content analysis methods to distinguish issue communication by political parties on Twitter
Computational communication research - ISSN 2665-9085 - 3:2(2021), p. 1-27
Full text (Publisher's DOI): https://doi.org/10.5117/CCR2021.2.004.PRAE
To cite this reference: https://hdl.handle.net/10067/1817990151162165141

Comparing automated content analysis methods to distinguish issue communication by political parties on Twitter

Stiene Praet[a], Peter Van Aelst[b], Walter Daelemans[c], Tim Kreutz[c], Jeroen Peeters[b], Stefaan Walgrave[b], and David Martens[a]

[a]Applied Data Mining Research Group, Department of Engineering Management, University of Antwerp

[b]Media, Movements and Politics Research Group, Department of Political Science, University of Antwerp

[c]Computational Linguistics and Psycholinguistics Research Group, Department of Linguistics, University of Antwerp

Abstract

Party competition in Western Europe is increasingly focused on "issue competition", which is the selective emphasis on issues by parties. The aim of this paper is to contribute methodologically to the increasing number of studies that deal with different aspects of parties' issue competition and communication. We systematically compare the value and shortcomings of three exploratory text representation approaches to study the issue communication of parties on Twitter. More specifically, we analyze which issues separate the online communication of one party from that of the other parties and how consistent party communication is. Our analysis was performed on two years of Twitter data from six Belgian political parties, comprising of over 56,000 political tweets. The results indicate that our exploratory approach is useful to study how political parties profile themselves on Twitter and which strategies are at play. Second, our method allows to analyze communication of individual politicians which contributes to classical literature on party unity and party discipline. A comparison of our three methods shows a clear trade-off between interpretability and discriminative power, where a combination of all three simultaneously provides the best insights.

1 Introduction

Issues and issue preferences form the raw matter of politics. The classic theory of democratic representation states that voters are expected to vote for parties that best represent the issues they deem important and that best represent their positional policy preferences on those issues (Thomassen & Schmitt, 1997). Therefore, parties try to steer the debate in the direction of the issue they have a strong profile or reputation on; this yields them an electoral advantage. Furthermore, the fragmentation of party landscapes across Europe in recent decades has increased the number of issues parties put forward. This explains why party competition in Western Europe has increasingly focused on the battle about which issues should dominate the party political agenda, i.e. "issue competition" (Green-Pedersen, 2007). The growing importance of issues in party politics, is also reflected by the rising attention for and proliferation of theories dealing with issue competition and communication (e.g. De Sio & Lachat, 2020).

Traditionally, research would examine party manifestos, campaign ads or press releases to study strategic issue communication choices (Tresch et al., 2017). However, nowadays social media represents an interesting alternative, as it is perhaps the most widely accessible form of party communication, with higher temporal adaptability and interaction potential (De Sio & Lachat, 2020). There is growing scholarly interest in parties' issue communication and strategies on social media (Vargo et al., 2014; Van Dalen et al., 2015; Van Ditmars et al., 2020). However, the high volatility of social media communication in combinations with relatively short and less formal text complicates automatic coding methods and party-level analysis. Therefore, the main aim of this study is to contribute to the rapid increase of studies that deal with different aspects of parties' issue communication on social media.

Especially Twitter is increasingly used by political parties and politicians to communicate with citizens, but even more so with opinion leaders and journalists (Jungherr, 2016; Vargo et al., 2014). We accept the press-release assumption of political parties on Twitter as suggested by De Sio & Lachat (2020) and extend this to individual politicians of a party. It states that, irrespective of the amount and type of followers a party's Twitter account might have, parties use Twitter as a way to communicate messages to the media and the public, like a press release, even in countries with low or elite-only Twitter penetration (Kreiss, 2016; Parmelee & Bichard, 2011).

In this study, we contribute to the issue competition literature by analyzing the issue communication of Flemish political parties on Twitter. More specifically, we are interested in how political parties differentiate themselves issue-wise from other parties in a multi-party system. We specifically focus on the emphasis they put on issues and not on their position towards issues. For instance, the theory of issue ownership states that parties can "own" issues if they are considered by the voters at large as the "best" party to deal with the issue (Petrocik, 1996; Walgrave et al., 2015). Hence, it is in a party's interest to make sure that the issues it owns are high on the priority list of voters. That is why parties tend to focus on their owned issues in their communication. Although several studies confirm that parties indeed focus on their issues, others show that parties "trespass" frequently and also address issues owned by their competitors (Damore, 2005). According to the recently developed issue yield theory, parties are more flexible and (ideologically) free to address issues that are not associated with the party as long as the party has a policy position on the issue that matches the party and if that position is also widely shared in the general electorate (De Sio & Lachat, 2020). While issue ownership and issue yield theory expect differences in the issue communication of parties, issue salience theories stress that parties and politicians address the issues that are high on the public and/or media agenda. By surfing the waves of issues that dominate the news, politicians can attract media attention for their political work (Van Santen et al., 2015; Wagner & Meyer, 2014).

Second, as we do not study the party as a single, united actor but rather study individual parliamentarians; we examine how consistent and coherent parties communicate about issues. Or, in other words, do politicians of the same party communicate about the same issues? Especially in election times a consistent issue strategy and clear, recognizable communication are valuable assets for persuading and retaining voters. Aligning online communication of all party representatives might be a beneficial strategy (Van Dalen et al., 2015). There also are reasons for politicians of the same party to address different issues. For instance, individual politicians may try to emphasize the issues they are specialized in to signal their expertise, and compete with politicians inside and outside their own party by emphasizing distinct issues (Peeters et al., 2019).

We propose an exploratory approach based on predictive modeling to find the most discriminative issues per party. The advantages of this exploratory approach are threefold. First, it allows researchers to move beyond an exclusive focus on frequency when analyzing issue communication. Rather than focusing on the most frequent issues per party (which could be similar for all parties), we argue it is more interesting to focus on the issues that differentiate one party from the others. Second, it does not require manual issue-coding of (a part of) the tweets, which is often labor-intensive and time-consuming. Third, an exploratory approach can contribute to existing theory by increasing our understanding of how parties try to profile themselves and which mechanisms and strategic choices drive issue communication. More specifically, per political party and based on the content of the tweet, a classification model is built to predict whether the author of the tweet belongs to the political party. We systematically compare three ways to represent the content of a tweet: (1) an expert-driven approach based on dictionaries, (2) a data-driven approach based on a bag of words method, and (3) another data-driven approach based on topic modeling. Before we turn to explain our data collection and discuss our results, we summarize established text classification methods in the field of politics and motivate our alternative approach.

2 Automated content analysis

Grimmer & Stewart (2013) argue that the understanding of language to know what political actors are saying and writing is central to the study of politics. Yet, the sheer volume of existing political texts does not allow for the manual reading and interpretation of all these documents. Automated content methods, however, can make the systematic analysis of large-scale text collections possible. For content analysis of political texts, typically two methods are considered: dictionary methods, based on the relative frequency of predefined key words in a document and supervised learning methods where the algorithm learns to classify documents into categories using a labeled training set (Grimmer & Stewart, 2013). Typically, when one is interested in party-level issue communication, one would classify texts into policy issues using one of both approaches, and aggregate results to learn the frequency of communication per issue at the party level. Next, we discuss how both methods can be used for the automated classification of policy issues in texts; after which we will explain why focusing on issue frequency might not be optimal to study issue communication by political parties.

To define issues, political scientists around the world often refer to the Comparative Agendas Project (CAP) codebook, consisting of 21 major issues (e.g. Environment, Macroeconomics), and more than 200 sub-issues[1]. Sevenans et al. (2014) manually compiled a Dutch dictionary of indicator words for each of the 21 CAP issues and showed it performs relatively well for issue classification. An important limitation of dictionary methods is that they depend on the quality of the predefined keywords and that dictionaries are of limited length, meaning that dictionaries are unable to capture all possible words related to a certain issue. When working with short texts such as tweets, the probability for dictionary words to appear in such a short text is low (Zirn et al., 2016). Moreover, with new words or terms being generated, a dictionary —mostly designed for formal text— soon becomes outdated (Wu et al., 2018). At the same time, extending dictionaries to improve coverage might come at the expense of lower precision.

To overcome the drawbacks of dictionaries, supervised learning has become a popular alternative. With supervised learning, the relevant features of the text and their weights are automatically estimated from a labeled data set (Barberá et al., 2019). Often-used methods for text classification are Logistic Regression, Support Vector Machines and Naive Bayes (Paul et al., 2017). Also recently, different variations of neural networks have been proposed for text classification (Lai et al., 2015). A notable challenge for the use of supervised learning, however, is that training a well-performing classifier requires a large training dataset coded by humans, where all policy issues of interest are well represented.

Annotating data is labor-intensive and several solutions have been proposed to reduce the coding work to a minimum; such as employing labeled data from a related task but different corpus, or using hashtags or well-defined keywords as annotations instead of human codings (Hasan et al., 2014; Gupta & Hewett, 2020). Next to that, semi-supervised learning (Van Engelen & Hoos, 2020) and transfer learning (Terechshenko et al., 2020) can be relevant to train a classifier when labeled data is scarce. The latter have been shown to outperform traditional classifiers with the same amount of (coded) training data (Terechshenko et al.,

[1]http://www.comparativeagendas.net/pages/master-codebook

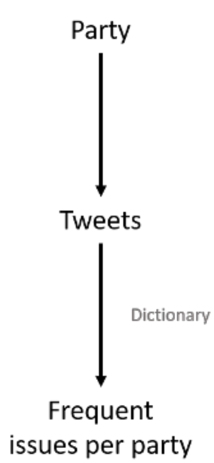
2020) but are increasingly complex and computationally demanding.

To sum up, achieving reliable document classification is hard, especially when considering a large number of classes. It requires compiling and/or updating dictionaries that are applicable to fast-evolving social media texts, or training a classifier on labeled data for which reaching sufficient accuracy is challenging to say the least. Moreover, even if we manage to classify documents to predefined categories reasonably well, the conclusions based on these results can be biased. The reason is that we try to optimize the classification of individual documents (e.g. tweets) in predefined categories (e.g. CAP issues), while the end goal is in fact to estimate the frequency or proportion of communication about a certain issue in a collection of documents (e.g. what percentage of tweets is about Macroeconomics). Unfortunately, even a well-performing individual classification model can be biased when the goal is to estimate category proportions. Suppose that all misclassifications happen in the same category, then the statistical bias in estimating the aggregate proportions could be very high (Hopkins & King, 2010). Methods exist to correct for this bias, or that give approximately unbiased estimates of category proportions directly, but they still require a sufficient set of labeled data (Hopkins & King, 2010).

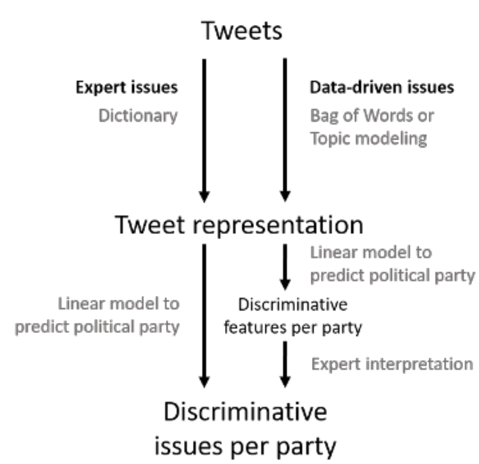
Finally, we argue that frequency of communication about a certain issue is in most cases not the object of interest. If all parties talk a lot about a certain issue, it is not inherent to a particular party's communication strategy. Therefore, it is more insightful to learn which policy issues are specific to one party but not to the others. In other words, how political parties differentiate themselves issue-wise from other parties. To illustrate this, have a look at the results of a frequency-based dictionary approach in Table 1. For half of the parties (left and center) the most-frequently discussed issues are almost completely identical. With a focus on frequency of communication we cannot differentiate between the issue strategies of these parties, as they seem similar at first sight.

Therefore, we propose to focus on discriminative issues (issues which distinguish one party from the others). We classify individual tweets according to the 21 CAP topics, using a dictionary. Subsequently, we apply supervised learning to automatically label the political party that authored the tweet. When learning this task, the machine will learn which features (policy issues) are relevant to a specific actor's (party) communication (Gentzkow et al., 2016). As discussed, this approach has the downside that results will be biased by the performance of the dictionary. Hence, we propose a data-driven approach, that eliminates the need to classify individual tweets according to the 21 CAP issues upfront. Based on textual features, tweets are directly classified to the political parties, and the machine learns which textual features are relevant. Subsequently, human coders or domain experts can analyze the relevant features and label them with policy issues, which significantly reduces the amount of work compared to labeling the original texts. The disadvantage of this data-driven approach is that it will be harder to draw conclusions on issue competition, as also other aspects of communication are taken into account. On the other hand, the exploratory nature of this approach can also be an advantage, as it provides a more fine-grained look into party communication. Figure 1 provides a schematic representation of a frequency-based dictionary approach and the alternative methods we propose.

Figure 1. Schematic representation of a traditional dictionary approach (a) and the alternative methods we propose in this paper (b).



(a) Traditional dictionary approach



(b) Methods proposed in this paper

3 Data and methods

In this study, we propose and validate the use of an exploratory approach to learn about issue communication and emphasis in Flanders (Belgium)[2]. We have collected tweets

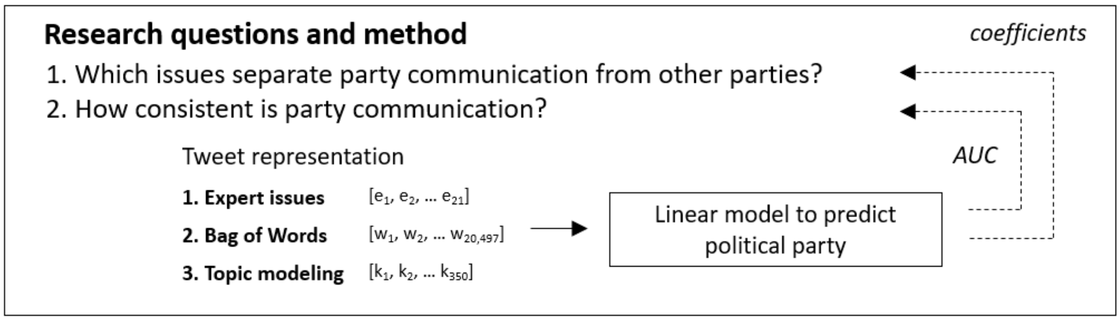
[2]Replication code can be found on Github: `https://github.com/SPraet/issue_communication`

Table 1
Most frequent CAP issues for Flemish parties on Twitter when applying a traditional dictionary approach.

Party	CAP issues
Groen	1. Transportation 2. Environment 3. Macroeconomics
Sp.a	1. Environment 2. Macroeconomics 3. Transportation
CD&V	1. Education 2. Transportation 3. Macroeconomics
Open VLD	1. International affairs 2. Education 3. Transportation
NVA	1. Immigration 2. Macroeconomics 3. International affairs
Vlaams Belang	1. Immigration 2. Law and crime 3. Government operations

from six Flemish political parties and their elected politicians. Per political party, we train a classification model that predicts whether the author of a tweet belongs to the political party or not, based on the representation—defined in three ways—of a tweet. The properties of the trained models are investigated to analyze issue communication per political party. First, the most discriminative features (with the highest coefficients in a linear model) show which issues distinguish parties' communication from one another (RQ1). In this study, we will focus on the top three most discriminative issues, but note that any other number can be chosen depending on the research desires. Second, the performance or discriminative power of the model per political party (measured by AUC, see Section 3.5) indicates how well the classification model can distinguish one party from the others. High discriminative power suggests that internal party communication is consistent and different from other parties (Gentzkow et al., 2016). Therefore, we will consider discriminative power per party as a proxy for internal consistency in party communication (RQ2). The research questions and method are summarized in Figure 2.

Figure 2. Overview of our exploratory approach to investigate issue communication by political parties on Twitter.



3.1 Data collection

For a time period of two years between October 2017 and October 2019, we collected more than 256,000 tweets from the official Twitter accounts of the six political parties represented in the (Flemish and federal) parliament: the Greens (Groen), Social Democrats (Sp.a), Christian-Democrats (CD&V), Liberals (Open Vld), Flemish Nationalists (NVA) and the Radical Right (Vlaams Belang, VB) and all their elected party representatives in the national or regional parliament including cabinet ministers and party leaders. First, we only select original tweets from these accounts, i.e. we do not include replies or retweets. Next, we separate the issue tweets, namely tweets that deal with a policy issue, from the tweets that deal with private life or refer to non-issue related aspects of politics such as messages to announce a campaign rally. We use a trained classifier[3] to select the issue tweets, which results in a final dataset of around 56,000 tweets by 227 individual politicians

[3] An external classifier (https://ccm.technology/) was trained on more than 37,000 labeled Facebook posts of Flemish politicians, to distinguish between issue-related tweets, private tweets and non-issue related (campaign) tweets. To test the performance of this classifier on our dataset, a random subset of 500 tweets was selected and manually labeled. The accuracy of the classifier on this test set was 84% and AUC was 92%. Removing private and non-issue related tweets results in a higher quality (less noise) dataset for our purpose. However, our approach is still applicable without this additional step and provides very similar results. The CAP issues per political party are largely the same and the predictive power is slightly lower (because of more noise) but this does not alter the conclusions.

and six political parties. The number of accounts and tweets per party can be found in Table 2.

Table 2
The number of accounts and tweets per party

Party	**Number of accounts**	**Number of tweets**
NVA (Flemish nationalists)	80	18,860
CD&V (Christian-democrats)	53	12,400
Open Vld (Liberals)	36	6,023
sp.a (Social-democrats)	31	6,545
Groen (Greens)	21	7,201
VB (Radical Right)	12	5,195

3.2 Preprocessing of tweets

Since the main interest of this research is to see how word usage in tweets might relate to political issues, we aim to reduce the event-specific information the tweets contain. Through intensive preprocessing we also want to reduce the noise that is common to social media texts (Han & Baldwin, 2011).

Tweets are first split into tokens and non-alphanumeric characters and stopwords[4] are removed. For Twitter specifically, this means that hashtags lose their '#-prefix and are handled as any other word. The use of user mentions, numbers and URLs in tweets is commonplace and might be informative for certain political issues; numbers playing an important role in financial news for example. However, we are not interested in the specific user, number or URL since it is unlikely that we can generalize from these. For that reason, these tokens are replaced with distinct placeholders.

Similarly, we argue that specific named entities (NE) in tweets are less informative to detect general policy issues. Using these words as features will cause our system to model specific events that occurred in the time-period of our data collection, rather than the more general policy issues that would be comparable to the expert dictionary. However, when it comes to named entities, the type of entity can still be informative for our purposes. Frequent mentioning of locations, for example, could be more indicative of issues like foreign affairs or defense, while frequent occurrence of organizations and

[4]We use the Dutch stopwords corpus from NLTK (`https://www.nltk.org/`).

products could relate to national economy. We use the Python library spaCy[5] for fine-grained tagging of named entities. We distinguish several types of named-entities such as locations, persons, organizations, products and events,[6] and replace them with their respective placeholders.[7] Lastly, we reduce word variation by lemmatizing the remaining tokens.[8] We are only interested in the lemma form of words because we aim to model their relatedness to political issues, regardless of their inflectional form.

3.3 Tweet representation

Before the actual modeling can start, the preprocessed tweets are transformed to a numerical representation. This will be done in three different ways, ranging from expert-driven to data-driven.

3.3.1 Expert issues In the first method, we will use the Dutch CAP dictionary compiled by Sevenans et al. (2014) to transform every tweet in our collection to 21 CAP issues. More specifically, every tweet is transformed to a binary vector of length 21, where each value represents the presence of a CAP issue in the tweet (1 if the issue is present in the tweet and 0 if not). Multiple issues can be present in one tweet. Consequently, predictive models are built on this representation to predict to which of the six parties the tweet belongs.

To evaluate the performance of the CAP dictionary, a random subset of 9,280 tweets was manually coded for the 21 CAP issues.[9] We first separate political tweets from non-political tweets[10] and then apply the CAP dictionary to code issues. We experimentally found that the CAP dictionary provides the best results when assigning an issue to a text as soon as one relevant dictionary word appears in the text, in which case the accuracy[11]

[5] `https://spacy.io/`

[6] For a complete list of entity types, see `https://spacy.io/api/annotation#named-entities`

[7] To assess how named entities influence our results, we have also repeated the same experiments (as will be explained in the following sections) for the data with named entities included. These results indicate that it is indeed the case that we model very specific short-term events as well as names of party representatives etc. Though the results are -as expected- better in terms of classification performance (AUC), they provide little insight in the general political issues of party communication

[8] We used the pattern.nl module developed by CLiPS: `https://github.com/clips/pattern`

[9] The tweets were coded by two coders who agreed in 44% of the cases on all labels. A more detailed overview of intercoder reliability per issue can be found in Table A1.

[10] Again, we apply the external classifier described before. The number of political tweets is 4954, or 54% of the evaluation set.

[11] Since this is a multi-label problem, *accuracy* refers to the percentage of tweets for which all labels were

Table 3
Overview of the 21 CAP issues (Sevenans et al., 2014).

Code	Issue
t100	Macroeconomics
t200	Human rights
t300	Health
t400	Agriculture
t500	Labor and employment
t600	Education
t700	Environment
t800	Energy
t900	Immigration
t1000	Transportation
t1200	Law and crime
t1300	Social welfare
t1400	Community development and housing
t1500	Banking, finance and domestic commerce
t1600	Defense
t1700	Space, science, technology and communications
t1800	Foreign trade
t1900	International affairs and foreign aid
t2000	Government operations
t2100	Public lands and water management
t2300	Culture and arts

of the CAP dictionary is 35%, recall is 20% and precision is 63%.[12] The low recall of the dictionary resulted in many zero input features (only 24% of the tweets could be assigned at least one issue, see Appendix A). Since the performance of the CAP dictionary on our tweets is low[13], we introduce two data-driven approaches below.

3.3.2 Bag of Words A first data-driven representation is a basic Bag of Words (BoW) approach, where each unique word corresponds to an input feature for the classification model[14]. This is still among the most commonly utilized methods in text classification (Barberá et al., 2019; Dun et al., 2020). Words are transformed into a

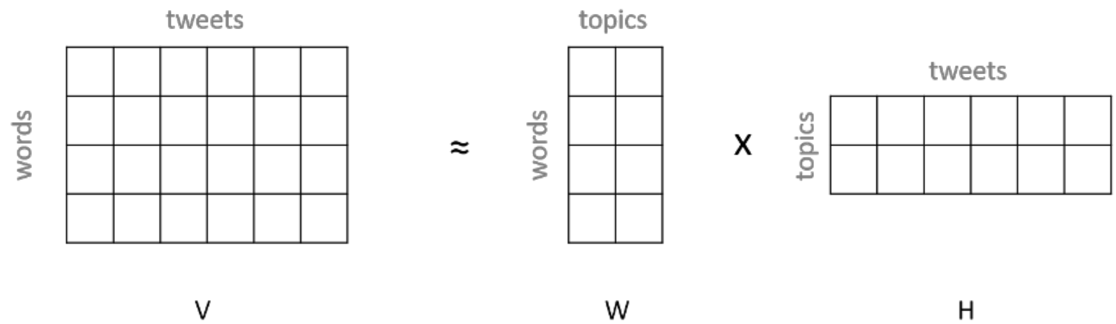
classified correctly.

[12]A more detailed evaluation per issue can be found in Appendix A

[13]We tried to improve the performance of the dictionary by extending it with word embeddings (see Appendix B). Although this results in higher recall; precision and accuracy are much lower.

[14]Including n-grams did not improve performance of the models, nor interpretation of the results. In fact, n-grams hardly were included in the most predictive features, and when they did it was in combination with a named entity, e.g. "ORG URL" or "says MENTION".

Figure 3. Illustration of non-negative matrix factorization of a matrix V consisting of m words in n tweets into two non-negative matrices W and H of the original n words by k topics and those same k topics by the m original tweets (Kuang et al., 2017).



numerical matrix using term frequency-inverse document frequency (tf-idf). The tf-idf matrix is used as input to predict to which of the six parties the tweet belongs. Afterwards, the most discriminative words will be manually interpreted in terms of the 21 CAP issues (see Section 3.5).

3.3.3 Topic modeling Alternatively, feature construction can be done using topic modeling techniques. The idea is to extract latent topics from the collection of tweets, where each topic is a multinomial distribution over words, and to represent each tweet as a mixture of these topics (Chang et al., 2009). Albeit useful to discover hidden topic structures in the data, topic detection techniques do not always improve final classification performance, especially when working with short texts (Conover et al., 2011). We will apply Non-negative Matrix Factorization (NMF)[15] to automatically extract topics from the political tweets. NMF is applied in multiple domains to decompose a non-negative matrix into two non-negative matrices. In the context of topic modeling, the term-document matrix is represented by two matrices, one containing the topics and one containing the coefficients to approximate the original matrix as close as possible (O'callaghan et al., 2015). This is visually represented in Figure 3.

The NMF topics are learned from the collection of political tweets,[16] and the original

[15] We have also experimented with another technique: Latent Dirichlet Allocation (LDA) (Blei & Lafferty, 2006). In our setting however, the predictive models based on the topics produced with NMF achieve higher discriminative power than with LDA, which is why we will report the results using NMF topics

[16] Additionally, we tried to build the NMF topics on a larger background collection, including tweets from all Flemish media channels and political journalists. It did not lead to more interpretable or more accurate

tweets are represented by k topics. Next, classification models are built on this representation. We optimize the number of topics (k) based on the performance of the subsequent supervised task: classification to one of the six parties. This way, the number of topics is set to 350, which is considerably higher than the 21 expert issues. Our data-driven topics are thus much more specific than the expert issues. Again, these data-generated topics will be manually interpreted in terms of the 21 CAP issues (see Section 3.5).

3.4 Classification models

Per political party, a classification model is built to predict whether the author of the tweet belongs to the political party or not, based on the representation of the tweet (see Figure 2). From these models, we want to analyze the most discriminative features for each of the six parties. For this reason we choose to work with Logistic Regression with l2 regularization[17], since the coefficients of this model are straightforward to interpret. Moreover, the discriminative power of this model showed higher or similar to the other classifiers in our benchmark[18] for the three different tweet representations. The coefficients and discriminative power of the trained models are investigated to draw conclusions on issue communication per political party.

3.5 Evaluation

We will systematically compare the three tweet representations defined in Section 3.3 in function of two evaluation criteria: discriminative power, or the ability to discriminate between political parties, and interpretability. First, to report the *discriminative power* of each model the last 20% of the tweets in our dataset are used as a separate out-of-time holdout set. We use the Area Under the ROC Curve (AUC) to measure how well the

results than topic detection on the political tweets only.

[17] More specifically, we use the scikit-learn implementation for logistic regression (Pedregosa et al., 2011). The model parameters are optimized (for AUC) using 5-fold out-of-time cross validation: the training data is split in 5 folds, where first the 5th fold is used as a validation set while the previous folds are used for training, then the 4th fold is used for validation and the previous folds for training, etc. The regularization parameter (C) is optimized in the interval [0.001, 0.01, 0.1, 1, 10]. For the topic modeling representation, we first optimize the number of topics k, which ranges from 0 to 400 with a stepsize of 50 and then we optimize the regularization parameter C for the optimal k.

[18] Other classifiers in our benchmark include (Multilayer) Perceptron, Lasso Regression, Linear Regression, Support Vector Machine, Naive Bayes, Decion Tree and Random Forrest

trained models can classify the political parties based on the tweet representations. AUC is a frequently used metric in data science to measure the performance of a classification model, independent of the frequency of the classes. It can be interpreted as the probability that the model ranks a random positive example higher than a random negative example (Flach et al., 2011). A perfect model would achieve an AUC of 100%, while an AUC of 50% indicates a random model (Provost & Fawcett, 2013). We calculate the weighted average AUC for the six classification models (one for each political party) to evaluate the discriminative power of our three different methods.[19]

Second, we define *interpretability* as the extent to which the most discriminative features correspond with the 21 CAP issues. When using the expert issue representation, the three most discriminative features are CAP issues and therefore by definition 100% interpretable. For the BoW and topic modeling representations we ask two independent domain experts to manually label the most discriminative features of the classification models with CAP issues (see example in Appendix C). Usually, topics extracted by a topic model are interpreted by humans by looking at the top-weighted words per topic (Chang et al., 2009). We will look at the top 15 words[20] to assign a CAP issue to an NMF topic. Similarly, for the BoW we will assume that 15 words represent one CAP issue. Since we want to report the three most discriminative issues (see Section 3), we will show 45 words. We repeat the same experiment with different domain experts and a different set of most discriminative features from a model trained on a random subsample of the data. The average percentage agreement of the two experts is used as a measure for interpretability (referred to as INT).

4 Results

In the following sections we provide our results regarding the two questions we introduced earlier: (1) which issues separate the communication of parties from each other and (2) how consistent is party communication? The first question is answered by looking at the top three most discriminative issues per party. Additionally, we explore to what extent this issue communication is in line with existing theory on issue competition. The

[19]Note that the weighted average AUC is used to compare the discriminative power of our three methods, while the AUC per political party is used to investigate consistency of party communication (see Figure 2).

[20]Usually between 6 to 30 words are considered, so other options are possible as well.

discriminative power of the model per political party provides us with an answer to the second question. A high discriminative power indicates that communication is coherent and consistent across individual politicians of the same party, while being distinct from other parties. Before we answer these questions, we will start with an evaluation of our three tweet representations.

4.1 Comparison of tweet representations

The classification models are built on tweet representations defined in three different ways: expert issues, BoW and topic modeling (NMF). When comparing these three approaches, a trade-off between classification performance of the classifiers and interpretability of the features becomes apparent. With the BoW representation the classification models are best able to distinguish between parties, while the expert issues offer the most direct interpretation of policy issues (Figure 4). The topic modeling representation seems to balance both criteria.

The models based on expert issues have an average AUC of 59% meaning they are only slightly better at discriminating between parties than random. One explanation is the limited performance of the CAP dictionary when converting tweets to the expert issues (see Section 3.3.1). Additionally, even with a perfectly accurate dictionary, valuable information (e.g. specific word usage) is lost when reducing the tweets to 21 issues, and we cannot discriminate between different sub-themes within the same issue. On the other hand, results are 100% interpretable as the issues are constructed top-down from the CAP dictionary itself.

With an average weighted AUC of 79%, the models based on BoW perform best at distinguishing between parties. The 45 most discriminative words are matched to the three most corresponding CAP issues (See Appendix C or one example in Table 4). This task is hard for domain experts since the most discriminative words are not necessarily thematically related, and therefore the average weighted interpretability is only 48%.

The discriminative power of the models based on the topic modeling representation (AUC = 68%) is higher than with the expert issues but lower than BoW. Per party we look at the three most discriminative NMF topics (each represented by 15 words) and manually

Figure 4. A comparison of our three methods on both evaluation criteria shows a clear trade-off between interpretability and discriminative power.

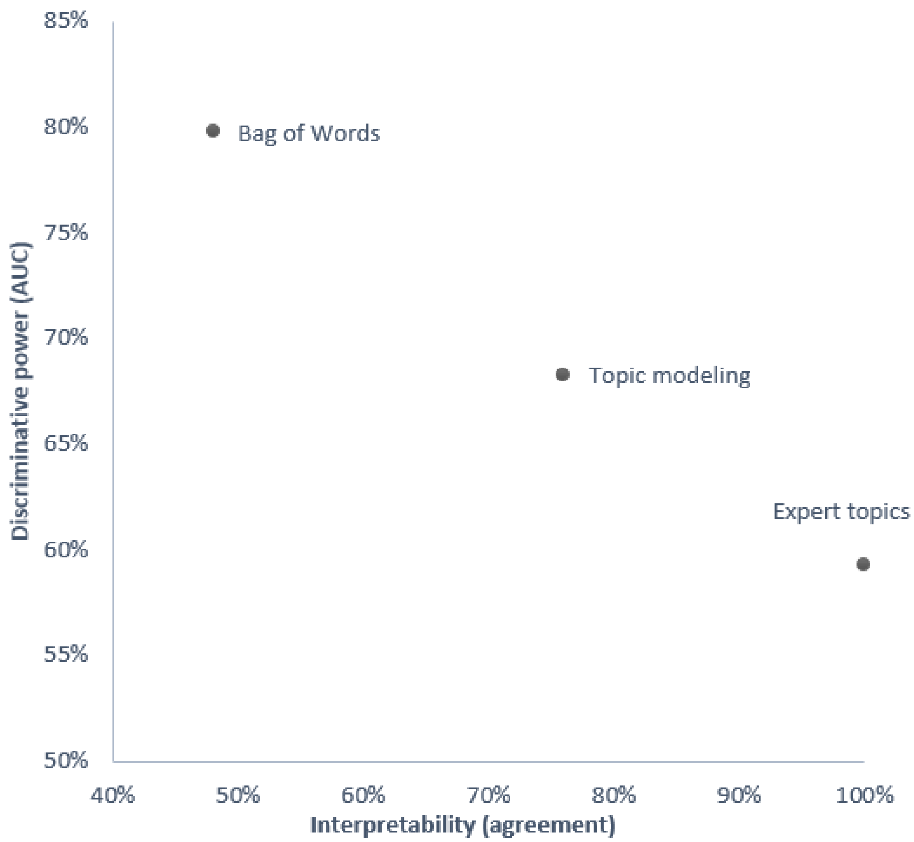


Table 4

The most discriminative features for the extreme right party (Vlaams Belang) when using the BoW approach, and the three most related CAP issues.

Party	Most discriminative features	CAP issues
VB	immigration, tomvangriek, islamization, vlaparl, immigration pact, mass immigration, islam, alien, immigration stop, immigrant, mosque, cordon, mosque, community, population, illegal, immigration policy, asylum seeker, multicultural, border, flanders ours again, concerning, URL, real, scum, immigrant, cause, country, people, people, terrorist, stop immigration, liberty, independence, our people first, protect our people, muslim, headscarf, so-called, government, even, elite, pact, madness	1. Immigration 2. Government operations 3. /

assign the most corresponding CAP issue (See Appendix C one example in Table 5). The expert interpretability is 84%, which indicates that domain experts mostly agree on which CAP issue corresponds to the NMF topic. This approach seems to find the best balance between discriminative power and interpretability.

Table 5
The most discriminative features for the extreme right party (Vlaams Belang) when using the topic modeling representation, and their corresponding CAP issues.

Party	Most discriminative features	CAP issues
VB	1. URL, action, and, due, youngsters, again, worry, ready, drawing, petition, life, share, right, thanks to, helping	1. Human rights
	2. country, border, safe, criminal, population, origin, illegal, deportation, alien, greatest, when, migrant, deport, hard, nationality	2. Immigration
	3. our, community, protect, security, proposals, economy, society, values, welfare, and, earn, pride, norm, farmer, resolution	3. Social welfare

4.2 Which issues separate party communication from other parties?

For every party, the most discriminative issues are shown in Table 6. For the more extreme parties on both sides of the political spectrum, the three methods give consistent results. For the greens (Groen), that started as a one-issue party, the issue focus on the Environment is still irrefutable, while radical right politicians (Vlaams Belang) have a clear focus on Immigration. These results are in line with issue ownership theory[21], stating that focusing on a few policy issues on which they have built a reputation is an effective strategy for parties to garner more votes. Another party that has a clear issue focus, at least partly in line with the issue ownership theory is, according to the different methods, is the NVA. Although the Flemish nationalists were traditionally not strongly focused on Immigration, in recent years they tried to "steal" the issue from the extreme-right party Vlaams Belang, which is also reflected in their communication on Twitter.

For the three traditional parties who are more situated in the center the issue focus is slightly more diffuse. The social-democrats of the Sp.a are linked to one of their core issues (Social welfare), but more often to an issue of a competitor (Environment, the core

[21]For issue ownership in Flanders, we rely on the study of Peeters et al. (2019) who asked Flemish respondents which party they instinctively though about when hearing a certain issue. We consider an issue owned by the party if the percentage of respondents that linked a certain party with the issue is higher than 20%.

issue of the Green party). The Christen-democrats (CD&V) most often communicate on Education, an issue that is traditionally linked to the many catholic schools in the country and for which the cabinet minister is a leading figure of their party. The (economic) liberals (Open Vld) seem to communicate least consistent on the issues they own (Macroeconomics), although several issues have an economic dimension (e.g. foreign trade, banking).

In sum, many parties' communication on Twitter is in line with the theory of issue ownership. For all parties, we find at least one issue that can be considered as an "owned" issue (see issues in bold in Table 6). However, most parties also seem to "trespass" their owned issues, in line with other issue competition theories. For example, the issue International Affairs is not owned by the liberal party Open Vld but they do have a minister for development cooperation in the federal government, which might be the reason for this specific issue focus. The reason opposition parties go beyond their owned issues is that they communicate about issues in reaction to what the government does. For example, the issue Defense is not owned by the socialist party Sp.a but in the period of data collection they heavily criticized the government decision to buy fighter planes. Finally, issue salience theory suggests that parties also respond to policy issues that are high on the public agenda (Van Santen et al., 2015; Wagner & Meyer, 2014). During the period of analysis these issues were Environment and Immigration. While concerns about the environment, and climate change in particular, were increasingly picked up by parties other than the Greens, the theme of immigration remained almost exclusively in the hands of the (radical) right. The data-driven methods allow to investigate sub-issues within issues, although this was not the focus of our study. For example, with respect to the salient issue of Environment, the Greens talk about a general climate policy, while the social-democrats and liberal party merely mention deposits on cans and small bottles, the Christen-democrats refer to their own important theme, namely quality of life, and finally, the Flemish nationalists discuss the efficiency of nuclear power plants driven by their approach of "eco-realism".

Table 6

The CAP issues Flemish party representatives communicate about on Twitter.

Party	Expert issues	Bag of Words	Topic modeling
Groen	**1. Environment** 2. Transportation 3. Agriculture	**1. Environment** 2. / 3. /	**1. Environment** 2. / 3. /
Sp.a	1. Defense 2. Environment 3. Health	**1. Social welfare** 2. Environment 3. Macroeconomics	1. Environment 2. Government operations **3. Social welfare**
CD&V	**1. Education** 2. Foreign trade **3. Social welfare**	**1. Social welfare** 2. Transportation **3. Education**	1. Environment 2. / **3. Education**
Open VLD	1. Foreign trade 2. Banking and finance 3. Agriculture	1. International affairs **2. Macroeconomics** 3. Banking and finance	1. International affairs 2. Environment 3. Immigration
NVA	1. Public lands and water **2. Immigration** 3. Science and technology	**1. Immigration** 2. Government operations **3. Law and crime**	**1. Immigration** 2. Energy **3. Immigration**
Vlaams Belang	**1. Immigration** 2. Government operations 3. Human rights	**1. Immigration** 2. Government operations 3. /	1. Human rights **2. Immigration** 3. Social welfare

Note: Issues printed in bold are owned by the party (Peeters et al., 2019). If none of the CAP issues matches with the set of words this is indicated with /.

4.3 How consistent is party communication?

To assess how consistent parties communicate we explore the discriminative power of the models per party (see Table 7). We assume that high AUC indicates consistent communication by the politicians of the considered party. For our three methods, the radical right party Vlaams Belang, is most consistent in their communication. This is partially due to the fact that this party pursues a clear positioning and association with one policy issue (Immigration). In addition, the lower number of party representatives is of course another explanation for more coherent communication. In that sense, it is remarkable that the N-VA, by far the biggest party with 80 representatives, scores not much lower in terms of consistency. This might be partly due to the high internal party discipline that characterizes Belgian parties (Depauw & Martin, 2009), and the N-VA in particular (Van Erkel et al., 2014). For all parties, AUC is higher for the data-driven methods than for the expert issues. This could indicate that party communication is more complex and not reducible to predefined issues. Indeed, with topic modeling we discover other characteristics of party communication rather than the policy issues they talk about. For example, one of the NMF topics for the liberal party (Open Vld) consists of English words (all other topics are in Dutch) and was apparently discriminative for Open Vld as it is the only party that occasionally tweets in English. Next to that, we often see party campaign slogans or hashtags among the most discriminative words, which can of course not be directly related to a policy issue.

Table 7

Classification performance and interpretability of the expert issues, Bag of Words and topic modeling representation.

	Expert issues		Bag of Words		Topic modeling	
	AUC	INT	AUC	INT	AUC	INT
Groen	60%	100%	82%	33%	71%	100%
sp.a	63%	100%	76%	50%	63%	67%
CD&V	57%	100%	81%	67%	70%	50%
Open Vld	61%	100%	79%	33%	71%	100%
NVA	56%	100%	76%	50%	66%	83%
VB	68%	100%	87%	33%	72%	67%
Weighted average	59%	100%	79%	48%	68%	76%

5 Conclusion and future research

Using three different tweet representations, we looked at which policy issues separate political parties on Twitter. Overall, our methods are remarkably good in distinguishing parties based on their (issue) communication. According to our results, especially the more extreme parties communicate clearly about the issues they "own". This finding is in line with issue ownership theory which suggests that political parties compete by raising attention for those policy issue that are positively associated with their party. On the other hand, several parties, mainly those in government, seem to trespass and also communicate about other issues, in line with other issue competition theories, such as issue salience or individual issue specialization and ministerial competences. The results indicate that our exploratory approach is useful to study how political parties distinguish themselves on Twitter and which strategies are at play. In addition, from the examination of the most discriminative words it becomes clear that a large part of communication on Twitter is event-driven, with parties talking about and reacting to current events that are limited in time. A more detailed temporal analysis could shed light on to what extend parties try and are successful to link these events to their owned issues.

By looking at the discriminative power of our models per political party we can draw conclusions about the consistency of communication by party representatives. This is highest for the more extreme (and also smaller) parties. Twitter is a much more personal communication channel than manifestos or press releases and individual politicians are free to tweet what they want (Peeters et al., 2019). Yet, for some political parties a classification model performs rather well in identifying their tweets based on the text only. As suggested by Gentzkow et al. (2016) the ease with which a machine learning model can infer a politician's party from their (written) language could be a measure for partisanship. A common language can be a key factor in creating group identity and party cohesion, but it can also increase inter-party hostility. An interesting direction for future research might be to look into how aligned all party representatives are in their communication, and to investigate communication strategy and its link to party composition (number, popularity, seniority, etc.) to explain the differences. This could be a useful contribution to the classic literature on party unity and party discipline that so far has not included the communication of individual politicians in their work (e.g. Depauw & Martin, 2009; Andeweg & Thomassen, 2011).

Lastly, with respect to our methodology, we think there's value in focusing on the distinctive character rather than just the frequency of communication. Classification models can distinguish one party from all other parties based on its communication, but they could also be applied to discriminate between two parties of interest (e.g. what is the difference in communication strategy of two nationalist parties NVA and Vlaams Belang). The expert- and data-driven approaches each have their advantages and disadvantages but by applying them simultaneously, different and complementary insights can be gained. The expert issues are insightful at the general issue level, but, next to being a result of low dictionary performance, the low AUC suggests that a lot of information is lost by trying to reduce political communication on Twitter to predefined issues. The low AUC could also suggests that political parties do not particularly differentiate themselves from their competitors in terms of issues but more in terms of specific content, as suggested by the higher AUC of the data-driven approaches. The data-driven approaches offer much more fine-grained insights at the event and even stylistic level of communication, at the expense of interpretability at the issue level. Moreover, the data-driven approaches allow to analyze sub-themes within issues. Although this was not the main focus of our study, our methods could help to study issues at a more fine-grained level. Additionally, the results could even help to improve issue dictionaries by bringing forward synonyms or other related terms. For example, the herbicide "glyphosate" was topic for debate during the time period of analysis. The term is not included in the current CAP dictionary, but is clearly related to the issue "Environment".

The methodology we propose is applicable to other (social media) text data and research questions as well. The expert-driven approach would benefit from improvements in document classification techniques. Recent advances in data-enhanced dictionaries, deep learning, transfer learning and semi-supervised learning offer exciting avenues for political text classification while at the same time introducing a lot of additional complexity and requiring ever more computing power. Adapting text classification to the volatility of social media remains a delicate exercise. Therefore, a promising method to study issue communication on social media is to start from a data-driven approach and use domain knowledge to interpret and understand the results.

6 References

Andeweg, R. B., & Thomassen, J. (2011). Pathways to party unity: Sanctions, loyalty, homogeneity and division of labour in the dutch parliament. *Party Politics*, *17*(5), 655–672.

Barberá, P., Boydstun, A. E., Linn, S., McMahon, R., & Nagler, J. (2019). Automated text classification of news articles: A practical guide. *Political Analysis*, 1–24.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on machine learning* (pp. 113–120).

Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Neural information processing systems* (Vol. 22, pp. 288–296).

Conover, M., Ratkiewicz, J., Francisco, M. R., Gonçalves, B., Menczer, F., & Flammini, A. (2011). Political polarization on twitter. *ICWSM*, *133*, 89–96.

Damore, D. F. (2005). Issue convergence in presidential campaigns. *Political Behavior*, *27*(1), 71–97.

Depauw, S., & Martin, S. (2009). Legislative party discipline and cohesion in comparative perspective. *Intra-party politics and coalition governments*, *103120*, 103–120.

De Sio, L., & Lachat, R. (2020). Making sense of party strategy innovation: challenge to ideology and conflict-mobilisation as dimensions of party competition. *West European Politics*, *43*(3), 688–719.

Dun, L., Soroka, S., & Wlezien, C. (2020). Dictionaries, supervised learning, and media coverage of public policy. *Political Communication*, 1–19.

Flach, P. A., Hernández-Orallo, J., & Ramirez, C. F. (2011). A coherent interpretation of auc as a measure of aggregated classification performance. In *Icml.*

Gentzkow, M., Shapiro, J., Taddy, M., et al. (2016). Measuring polarization in high-dimensional data: Method and application to congressional speech.

Green-Pedersen, C. (2007). The growing importance of issue competition: The changing nature of party competition in western europe. *Political studies*, *55*(3), 607–628.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, *21*(3), 267–297.

Gupta, V., & Hewett, R. (2020). Real-time tweet analytics using hybrid hashtags on twitter big data streams. *Information*, *11*(7), 341.

Han, B., & Baldwin, T. (2011). Lexical normalisation of short text messages: Makn sens a #twitter. In *Proceedings of the 49th annual meeting of the association for computational* *linguistics: Human language technologies - volume 1* (pp. 368–378). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from `http://dl.acm.org/citation.cfm?id=2002472.2002520`

Hasan, M., Agu, E., & Rundensteiner, E. (2014). Using hashtags as labels for supervised learning of emotions in twitter messages. In *Acm sigkdd workshop on health informatics, new york, usa.*

Hopkins, D. J., & King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, *54*(1), 229–247.

Jungherr, A. (2016). Twitter use in election campaigns: A systematic literature review. *Journal of information technology & politics*, *13*(1), 72–91.

Kreiss, D. (2016). Seizing the moment: The presidential campaignsâ use of twitter during the 2012 electoral cycle. *New media & society*, *18*(8), 1473–1490.

Kreutz, T., & Daelemans, W. (2018). Enriching general sentiment lexicons for domain-specific polarity classification. *Manuscript in preparation.*

Kuang, D., Brantingham, P. J., & Bertozzi, A. L. (2017). Crime topic modeling. *Crime Science*, *6*(1), 1–20.

Lai, S., Xu, L., Liu, K., & Zhao, J. (2015). Recurrent convolutional neural networks for text classification. In *Aaai* (Vol. 333, pp. 2267–2273).

O'callaghan, D., Greene, D., Carthy, J., & Cunningham, P. (2015). An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, *42*(13), 5645–5657.

Parmelee, J. H., & Bichard, S. L. (2011). *Politics and the twitter revolution: How tweets influence the relationship between political leaders and the public.* Lexington Books.

Paul, D., Li, F., Teja, M. K., Yu, X., & Frost, R. (2017). Compass: Spatio temporal sentiment analysis of us election what twitter says! In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining* (pp. 1585–1594).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, *12*, 2825–2830.

Peeters, J., Van Aelst, P., & Praet, S. (2019). Party ownership or individual specialization? a comparison of politicians' individual issue attention across three different agendas. *Party Politics*, 1354068819881639.

Petrocik, J. R. (1996). Issue ownership in presidential elections, with a 1980 case study. *American journal of political science*, 825–850.

Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking.* " O'Reilly Media, Inc.".

Sevenans, J., Albaugh, Q., Shahaf, T., Soroka, S., & Walgrave, S. (2014). The automated coding of policy agendas: A dictionary based approach (v. 2.0.). In *7th annual comparative agendas project (cap) conference, konstanz, june* (pp. 12–14).

Terechshenko, Z., Linder, F., Padmakumar, V., Liu, M., Nagler, J., Tucker, J. A., & Bonneau, R. (2020). A comparison of methods in political science text classification: Transfer learning language models for politics. *Available at SSRN*.

Thomassen, J., & Schmitt, H. (1997). Policy representation. *European Journal of Political Research*, *32*(2), 165–184.

Tresch, A., Lefevere, J., & Walgrave, S. (2017). How parties' issue emphasis strategies vary across communication channels: The 2009 regional election campaign in belgium. *Acta Politica*, 1–23.

Van Dalen, A., Fazekas, Z., Klemmensen, R., & Hansen, K. M. (2015). Policy considerations on facebook: Agendas, coherence, and communication patterns in the 2011 danish parliamentary elections. *Journal of Information Technology & Politics*, *12*(3), 303–324.

Van Ditmars, M. M., Maggini, N., & van Spanje, J. (2020). Small winners and big losers: strategic party behaviour in the 2017 dutch general election. *West European Politics*, *43*(3), 543–564.

Van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning*, *109*(2), 373–440.

Van Erkel, P., Thijssen, P., & Van Aelst, P. (2014). Vier ideologische dimensies, één breuklijn. *Stichting Gerrit Kreveld: Samenleving en Politiek*, 1–12.

Van Santen, R., Helfer, L., & Van Aelst, P. (2015). When politics becomes news: An analysis of parliamentary questions and press coverage in three west european countries. *Acta Politica*, *50*(1), 45–63.

Vargo, C. J., Guo, L., McCombs, M., & Shaw, D. L. (2014). Network issue agendas on twitter during the 2012 us presidential election. *Journal of Communication*, *64*(2), 296–316.

Wagner, M., & Meyer, T. M. (2014). Which issues do parties emphasise? salience strategies and party organisation in multiparty systems. *West European Politics*, *37*(5), 1019–1045.

Walgrave, S., Tresch, A., & Lefevere, J. (2015). The conceptualisation and measurement of issue ownership. *West European Politics*, *38*(4), 778–796.

Wu, L., Morstatter, F., & Liu, H. (2018). Slangsd: building, expanding and using a sentiment dictionary of slang words for short-text sentiment classification. *Language Resources and Evaluation*, *52*(3), 839–852.

Zirn, C., Glavaš, G., Nanni, F., Eichorts, J., & Stuckenschmidt, H. (2016). Classifying topics and detecting topic shifts in political manifestos.

Appendix A

Evaluation of the CAP dictionary per issue

Table A1

Intercoder reliability (ICR), measured by Cohen's Kappa, per issue for the annotated tweets. ICR is fair to moderate for most issues. The issues with low (to zero) ICR have low (to no) occurrence.

Issue	ICR
Macroeconomics	0.61
Human rights	0.43
Health	0.52
Agriculture	0.79
Labor and employment	0.60
Immigration	0.69
Education	0.69
Environment	0.68
Energy	0.73
Transportation	0.64
Law and crime	0.47
Social welfare	0.31
Community development	0.30
Banking and finance	0.56
Defense	0.51
Science and technology	0.18
Foreign trade	0.00
International affairs	0.40
Government operations	0.41
Public lands and water	0.00
Culture and arts	0.41

Table A2

Evaluation of the CAP dictionary in terms of precision, recall and F1 score per issue.

Issue	Precision	Recall	F1 score
Macroeconomics	43%	13%	20%
Human rights	44%	19%	27%
Health	63%	21%	31%
Agriculture	66%	20%	31%
Labor and employment	54%	28%	37%
Immigration	74%	22%	34%
Education	72%	36%	48%
Environment	65%	35%	45%
Energy	66%	36%	47%
Transportation	77%	48%	59%
Law and crime	61%	16%	25%
Social welfare	66%	4%	8%
Community development	29%	12%	17%
Banking and finance	60%	8%	14%
Defense	69%	13%	21%
Science and technology	50%	1%	2%
Foreign trade	0%	0%	0%
International affairs	27%	7%	11%
Government operations	46%	3%	5%
Public lands and water	0%	0%	0%
Culture and arts	60%	5%	9%

Table A3

Number and percentage of tweets that was assigned a certain CAP issue using the CAP dictionary.

Issue	Number of tweets	Percentage of tweets
Macroeconomics	911	2%
Human rights	517	1%
Health	780	1%
Agriculture	393	1%
Labor and employment	1265	2%
Immigration	1170	2%
Education	1772	3%
Environment	767	1%
Energy	598	1%
Transportation	2277	4%
Law and crime	1138	2%
Social welfare	521	1%
Community development	277	0%
Banking and finance	256	0%
Defense	291	1%
Science and technology	55	0%
Foreign trade	104	0%
International affairs	735	1%
Government operations	454	1%
Public lands and water	17	0%
Culture and arts	116	0%
No issue	43245	76%

Appendix B

Extended dictionary with word embedding

The dictionary maps keywords to their respective political issues and aims to be very precise, with keywords having a very distinct meaning and low probability to be present in one of the other issues. For analysis of short social media texts such as tweets, in which very few words are present, this precision is less important and coverage with the expert dictionary is of more concern. To extend the indicator words in the original dictionary, we use word embeddings trained on a large corpus of political social media data (Kreutz & Daelemans, 2018). The word embeddings encode a numerical vector per word, which contains the point-wise mutual information (PMI) with other words in the corpus. Using these vectors, we can find candidate words that are semantically similar to the keywords already present in the dictionary, using a cosine-similarity of 0.6 or higher. The candidates were then manually inspected and filtered to contain only words that extend coverage of the expert issues without clearly impairing their delineation. Using word embeddings in this way, we were able to extend the keywords from an average of 87 per expert issue to 157 per expert issue and consequently, 85% of the tweets could be assigned at least one issue (compared to only 24 % for the original dictionary).

The extended dictionary was also tested on a random subset of 9,280 tweets that was manually coded for the 21 CAP issues. Accuracy of the extended dictionary is 20%, recall is 35% and precision is 39%. Although recall and coverage could be increased by extending the dictionary, the precision is much lower than that of the original dictionary. For this reason we decided to apply the original dictionary in this research, despite the low coverage. This shows that accurately extending the existing dictionaries is still a difficult challenge.

Appendix C

Most discriminative features for the data-driven representations

Table C1

The most discriminative features when using the BoW approach and their three most related CAP issues. Named entities are printed in capital letters.

Party	Most discriminative features	Corresponding CAP issues
Groen	itcanbedifferent, greenworks, lowerhouse, glyphosate, meyremalmaci, climate ambition, morehealthy, changecongress, antwerpcandoit, advertising, widening change, concrete stop, cyclist, green, screening, whistle blower, forest, air pollution, pesticide, climate generation, c'est (French), youthforclimate, climate top, incomprehensible, air, fairer, hormone disruptor, glyphosate, takeNUMBER, hood, practical test, longlivepolitics, e.g., climate policy, serious, audit, kristofcalvo, position, flemish, terzaketv, majority, complete, unworthy, unbelievable, poverty line	1. Environment 2. / 3. /
sp.a	security of care, wetakecare, municipal works, schaarbeek, bredene, securityforall, stopthedebtindustry, flemish government, goleft, security, johncrombez, debt industry, assets, molenbeek, proposal, nuclear weapon, new battle, weapon embargo, fail, deposit (for packaging), future budget, beach, care crisis, plastic, resolution, veviba (meat company), replacement, nuclear, youarewhatyoueat, sp.a, history of bredene, throwback, contest, vanovertveldt, voted out, water bill, litter, saving, reynders, feed, plow, reading, profit, crazy, weak	1. Social welfare 2. Environment 3. Macroeconomics
CD&V	thewayforward, quality of life, socialeurope, gtgen, justice, wbeke, bike is king, social, climate court, cd&v, traffic jam idea, safe traffic, consultation, peeterskrisNUMBER, residential care centers, teacher, jokeschauvliege, mobility budget, info, koengeensNUMBER, belgians, homeinthecity, crevits, improve, tooth, renew, brexit, worker, inheritance law, social right, electrical, movingsafely, callNUMBER, economic, elderly care, belgiangovernment, simpler, school construction, climate pact, quitting principle, close to you, recommendation, opening, school year, servaisv	1. Social welfare 2. Transportation 3. Education
Open Vld	justdoit, positiveforward, vilvoorde, must (ENG), pedestrian son, liberal, liberal, Sint-truiden, basic income, etc., europe, reform, ambitious, plenary, children (ENG), united (ENG), read, lost, survive, miscellaneous, facebook, subway, would (ENG), unsupported, entrepeneur, agriculture, proud (ENG), think (ENG), could (ENG), need (ENG), dry, closer, iameuropean, city hall, right (ENG), unity (ENG), futureofeurope (ENG), humanright (ENG), minor, Brussels, entrepreneur, strategy (ENG), weareeurope (ENG), speech (ENG), task	1. International affairs 2. Macroeconomics 3. Banking and finance
NVA	member of parliament, good news, pride heritage, prisoner, herental, left, vdag, marrakesh coalition, meanwhile, heritage, lgbthistorymonth, member of the European parliament, works of change, minority government, animal welfare, tg, flanders, prime minister, budgetNUMBER, self-determination, rajoy, flemishNUMBER, civil service, marrakesh coalition, structure, homeland, policeman, transit migration, change, via, migrant, gene, union, factor, restriction, catalan, repression, hear, yourpowerfulmanagement, say, excellent, steenokkerzeel, restoration, maybe, prosperity	1. Immigration 2. Government operations 3. Law and crime
VB	immigration, tomvangriek, islamization, vlaparl, immigration pact, mass immigration, islam, alien, immigration stop, immigrant, mosque, cordon, mosque, community, population, illegal, immigration policy, asylum seeker, multicultural, border, flandersoursagain, concerning, URL, real, scum, immigrant, cause, country, people, people, terrorist, stop immigration, liberty, independence, ourpeoplefirst, protect our people, muslim, headscarf, so-called, government, even, elite, pact, madness	1. Immigration 2. Government operations 3. /

Table C2

The most discriminative features when using the topic modeling representation and their most related CAP issues. Named entities are printed in capital letters.

Party	Most discriminative features	Corresponding CAP issues
Groen	1. itcanbedifferent, green, deochtend (radio program), climate, air, work green, meyremalmaci, lower house, kristofcalco, your, poverty, wouterdevriendt, climate policy, plan, honest	1. Environment
	2. WORK_OF_ART, NUMBERday, flemish parliament, so, vtmnieuws, tomvangriek, antwerpcandoit, zaak, koengeensNUMBER, according to, petermertenq, wbeke, youthforclimate, URL, get	2. /
	3. incomprehensible, guess, advertisement, even, a lot, muyters, only, soil, online, mother tongue, abuse, flight, unacceptable, rent deposit, just	3. /
sp.a	1. MENTION, URL, deposit (for packaging), strong, colleague, among others, rightfully, gasses, hearing, deochtend (radio program), proposal, later, success, member of parliament, tonight	1. Environment
	2. government, federal, parliament, decision, follow, fall, decided, run, flemish, previous, next, prime minister, on behalf of, opposition	2. Government operations
	3. care, for, affordable, security of care, wellbeing, qualitative, quality of life, elderly, informal care, person, elderly care, retirement home, qualitatively, support, quality	3. Social welfare
CD&V	1. thewayforward, quality of life, care, thanks to, municipality, job, air quality, plenty, bike is king, reformation, neighbourhood, mobility budget, further, healthy, ambitious	1. Environment
	2.important, put, step, further, forwards, step, busy, role, direction, because, again, shoulder, follow, measurement, look	2. /
	3. information , URL, from, discuss, during, free, school year, website, from now on, to, dual, number/grade, correct, subscribe, learn	3. Education
Open Vld	1.europe, need, new, peopl, social, future, work, today, right, together, must, world, maak, fight, meeting (all in English)	1. International affairs
	2. MENTION, URL, deposit (for packaging), strong, colleague, among others, rightfully, gasses, hearing, deochtend (radio program), proposal, later, success, member of parliament, tonight	2. Environment
	3. PERSON, URL, prime minister, plus, important (ENG), right (ENG), brussels, migration (ENG), conversation, must, police, ORGANIZATION, question (ENG), us (ENG), one (ENG)	3. Immigration
NVA	1. NATIONALITY, URL, meeting, captured, economy, according to, member of parliament, speak, president, colleague, level, political, citizen, violence, nationalities	1. Immigration
	2. say, member of parliament, dare, when, come on, no, enough, debt, alone, nuclear plant, little, money, MENTION, often, opinion	2. Energy
	3. via, URL, MENTION, member of parliament, representative, save, fiscal, migrant, sail, money, finance, asylum seeker, information, free, security of care	3. Immigration
VB	1. ULR, action, and, due to, youngsters, again, care, ready, draw, petition, live, part, right, thanks to, help	1. Human rights
	2. country, border, safe, criminal, population, origin, illegal, deportation, alien, greatest, when, migrant, deportation, hard, nationality	2. Immigration
	3. our, society, protect, safety, propose, economy, society, values, prosperity, and, earn, pride, norm, farmer, resolut	3. Social welfare