ECAI 2016
G.A. Kaminka et al. (Eds.)

doi:10.3233/978-1-61499-672-9-28

Validating Cross-Perspective Topic Modeling for Extracting Political Parties' Positions from Parliamentary Proceedings

Janneke M. van der Zwaan[1] and **Maarten Marx** and **Jaap Kamps**[2]

Abstract.
In the literature, different topic models have been introduced that target the task of viewpoint extraction. Because, generally, these studies do not present thorough validations of the models they introduce, it is not clear in advance which topic modeling technique will work best for our use case of extracting viewpoints of political parties from parliamentary proceedings. We argue that the usefulness of methods like topic modeling depend on whether they yield valid and reliable results on real world data. This means that there is a need for validation studies. In this paper, we present such a study for an existing topic model for viewpoint extraction called cross-perspective topic modeling [11]. The model is applied to Dutch parliamentary proceedings, and the resulting topics and opinions are validated using external data. The results of our validation show that the model yields valid topics (content and criterion validity), and opinions with content validity. We conclude that cross-perspective topic modeling is a promising technique for extracting political parties' positions from parliamentary proceedings. Second, by exploring a number of validation methods, we demonstrate that validating topic models is feasible, even without extensive domain knowledge.

1 Introduction

Over the last fifteen years, topic modeling has been established as a method for uncovering hidden structure in document collections. Traditionally, these methods learn probability distributions over words along a single dimension of topicality, and do not take into account other dimensions, such as sentiment, perspective, or theme [20]. More recently, various extensions to topic modeling have been proposed that do allow for multiple dimensions. In our work, we are interested in opinion or viewpoint extraction from text.

Different topic models have been proposed that target the task of viewpoint extraction, each of which is based on different assumptions about what an opinion or viewpoint consists of, and impose different requirements on the data. Section 2 provides an overview of these methods, and their particularities. Studies introducing new topic models usually only evaluate model fit based on held-out perplexity and provide anecdotal evidence that the results make sense. Systematic validations of new algorithms are rare. As a result of this, it is not clear in advance which topic modeling technique will work best for our use case of extracting viewpoints of political parties from parliamentary proceedings.

It can be argued that the usefulness of methods like topic modeling depend on whether they yield valid and reliable results on real world data. In order for researchers from other domains to trust and use methods such as topic modeling, insight into the strengths and limitations of individual techniques is indispensable. In domains applying these text mining techniques, e.g., political science, the need for validation is already recognized [14]. However, from a computer science perspective, validation of methods is also important, because insight into why methods are successful or not can help to improve existing methods and inform the design of new methods.

This paper presents a validation of one particular topic model for viewpoint extraction: the cross-perspective topic model [11]. The cross-perspective topic model learns topics and opinions from a corpus that is (or can be) divided in perspectives (e.g., political parties). Topics are learned from topic words (nouns) in the entire corpus, whereas opinions are learned from opinion words (adjectives) for each perspective separately. The model is applied to Dutch parliamentary proceedings from 1999 to 2012. The Dutch multiparty political system allows us to apply cross-perspective topic modeling to data consisting of more than just two or three perspectives. We demonstrate that cross-perspective topic modeling yields valid topics. While opinions extracted from the parliamentary proceedings are representative of the political parties' positions, these positions were uncorrelated with classical left/right rankings of the parties. These results indicate that cross-perspective topic modeling is a promising technique for extracting political parties' positions from parliamentary proceedings.

This paper is organized as follows. Section 2 introduces the related work by reviewing existing topic models for viewpoint extraction. In section 3, we provide an in-depth explanation of cross-perspective topic modeling. Section 4 presents the design of the validation study. Topic and opinion validity are assessed in section 5. The results are discussed in section 6. Finally, we present our conclusions in section 7.

[1] Netherlands eScience Center, The Netherlands, email: j.vanderzwaan@esciencecenter.nl
[2] University of Amsterdam, The Netherlands, email: {maartenmarx, j.kamps}@uva.nl

2 Related Work

Opinion mining or sentiment analysis is concerned with extracting subjective information from text. Pang and Lee provide a broad introduction to this diverse research area [19]. This section provides a review of topic models for opinion or viewpoint extraction, and discusses the need for validation studies.

2.1 Topic Models for Viewpoint Extraction

The Joint Topic and Perspective (JPT) model assumes lexical variation in ideological text can be attributed to the topic and the author's point of view [18]. Consequently, for each word in the vocabulary two weights are learned: a topical and an ideological weight. The model needs a collection of texts on the same topic that is divided into two contrasting perspectives. In essence, this method learns a single topic per perspective; so, there are two topics in total.

The Joint Topic Viewpoint (JTV) model proposed by Trabelsi and Zaïane assumes documents contain expressions of one or more divergent viewpoints [25]. Each term in a document is assigned a topic and a viewpoint. The model generates a probability distribution over terms for each topic-viewpoint pair. This means that 'objective' (i.e., substantive) and 'subjective' (i.e., viewpoint-specific) information are mixed. Although, theoretically, the number of perspectives can be > 2, for the experiments presented in the paper it is set to 2 (i.e., the viewpoints are 'supporting' and 'opposing').

Gottipati et al. propose a topic model to infer topics and positions (pro/con) by exploiting the hierarchical structure in which debates are organized on Debatepedia[3] [12]. The model learns to classify terms either as named entities, 'general' position terms, 'topic-specific' position terms, 'topic' terms, or 'background' terms. The positions are limited to pro and con.

The Topic-Aspect (TAM) model was designed to capture a text's underlying perspectives [20]. In addition to a mixture component to filter 'background' words, the model assigns words either to an aspect-neutral or aspect-dependent distribution. This means 'objective' and aspect-specific information is separated. The number of aspects (perspectives) is a parameter of the model. The model learns the perspectives from the data, so documents do not need to be labeled; in fact, it is assumed that documents contain mixtures of perspectives.

The Viewpoint and Opinion Discovery Unification Model (VODUM) uses heuristics to learn topics, viewpoints, and opinions from text [23]. A viewpoint is defined as a standpoint on a set of topics, and an opinion is a wording that is specific to a topic and a viewpoint. VODUM separates topic and opinion words based on their part of speech tag. In addition, words in the same sentence are assumed to belong to the same topic, and all text in a document is assumed to belong to the same viewpoint. These constraints help to improve model fit.

Generally, topic models for viewpoint extraction target slightly differing tasks, ranging from finding arguments or evidence for or against standpoints to discovering words indicative of viewpoints or perspectives, and are usually designed to exploit characteristics of the particular data used (e.g., document structure as with documents from Debatepedia). Although work that presents new viewpoint extraction topic models includes evaluations of the results produced by these models, the evaluations typically are limited. Quantitative evaluations usually assess model fit, based on held-out perplexity (i.e., [11, 18, 23, 25]). However, it is not surprising that topic models that exploit particularities of the data and/or tasks, generally result in models with a better fit to the data. Also, evaluations of model fit do not take into account the extent to which topics and viewpoints learned from data make sense at all. In fact, topics from a model with lower perplexity are not necessarily more meaningful to humans than topics from a model with higher perplexity [10]. We argue that in order to gain insight into the contents of document collections, topic modeling results must be semantically meaningful to humans. Therefore, it is necessary to evaluate performance in other ways than just calculating perplexity. In this study, we use additional, domain specific data to validate topic modeling results.

Qualitative evaluations of topic modeling results presented in the related work are anecdotal, and do not go beyond presenting anecdotal results for topics and opinions/viewpoints (i.e., [11, 12, 20, 23, 25]). Again we contend that more thorough evaluations are required in order for the results to be useful representations of documents in collections.

This paper presents a validation study of the results of cross-perspective topic modeling [11]. There were multiple reasons to select this method and not one of the others. The cross-perspective topic model yields explicit representations of viewpoints, which allows us to quantify differences between the viewpoints. This is helpful for representations that can be compared directly to external data. Also, the CPT model allows for more than two perspectives on a topic, whereas many other models aim to learn pro/con stances towards topics only. Although this results in a counterintuitive notion of what an opinion is (i.e., probability distribution over words vs. arguments for or against), it allows for a more nuanced representation of opinions and differences between opinions. Finally, CPT is conceptually simpler than some of the other models, which makes it easier to implement.

2.2 Assessing Validity

Grimmer and Stewart note that 'all automated [text mining] methods are based on incorrect models of language' [14]. While this does not imply that the results of these methods are therefore useless, it does mean that output of automated text mining methods must be validated before their results can be trusted. In the introduction, we argued that validation studies are relevant for both domain scientists that use these methods to explore text corpora, and computer scientists that work on developing new text mining methods and improving existing ones.

Validity refers to the extent to which a measure measures what it is intended to measure [9]. Table 1 lists different types of validity. According to Grimmer and Stewart, 'to validate the output of an unsupervised method, scholars must combine experimental, substantive, and statistical evidence to demonstrate that the measures are as conceptually valid as measures from an equivalent supervised model' [14].

[3] http://www.debatepedia.org/

Type	Description
Face	The extent to which results appear to be valid.
Content	The extent to which a method for measuring a latent construct represents all of its facets.
Criterion	Correlation between a measure and other measures that reflect the same concept.
Construct	The extent to which a measure behaves as expected in a theoretical context.

Table 1: Types of validity (adapted from [9]).

Quinn et al. performed a validation study of topic modeling on speech in the US Senate [21]. They show that, with the exception of 'procedural' topics, words from the same topic generally have common substantive meaning, that hierarchical clusterings of topics yield meaningful semantic relationships between topics, that topics found in speeches correlate with roll-calls and hearings, and that there is correlation between topics found in speeches and exogenous events. However, the different steps of the validation are mostly qualitative. In addition, the topic modeling method used is simpler than cross-perspective topic modeling; a speech can be assigned to a single topic only, and opinions are not taken into account.

In this paper, we address content validity and criterion validity of topics and opinions extracted using cross-perspective topic modeling.

3 Cross-Perspective Topic Modeling

The cross-perspective topic model [11] is an extended form of Latent Dirichlet Allocation (LDA) [6]. Topics are learned by doing LDA on the topic words (nouns) in the corpus. Opinions are learned from a separate LDA process using opinion words (adjectives, verbs, and adverbs). A topic is a probability distribution over topic words. An opinion is a probability distribution over opinion words. While the topics are shared among the entire corpus, opinions depend on the perspective a document belongs to. A document can only belong to a single perspective, and the division of the corpus in perspectives is fixed and must be known in advance.

The imaginary process for generating documents is: one first selects a topic, based on the topic mixture of that document. Then a topic word is drawn from the topic. This procedure is repeated until all topic words have been selected. Next, one selects an opinion based on the frequency of topic words associated with the topics in the document. The more words associated with a certain topic, the higher the chance that the corresponding opinion will be selected. The contents of the opinion (i.e., probabilities of opinion words) depend on the generator's perspective. Next, an opinion word is drawn from the selected opinion. This procedure is again repeated until all opinion words have been selected. More formally, this generative process can be described as follows.

1. Draw a perspective-independent multinomial topic word distribution ϕ from Dirichlet(β) for each topic z
2. Draw a perspective-specific multinomial opinion word distribution ϕ_o^i from Dirichlet(β_o^i) for each opinion x for perspective c^i
3. For each document d choose a topic mixture θ from Dirichlet(α)
4. For each topic word w in document d
   (a) Draw a topic z from Multinomial(θ)
   (b) Draw a word w from Multinomial(ϕ) conditional on z
5. For each opinion word o in document $d \in c^i$
   (a) Draw an opinion x from Uniform($z_{w_1}, z_{w_2}, \ldots, z_{w_{N_{w(d)}}}$)
   (b) Draw an opinion word o^i from Multinomial(ϕ_o^i) conditional on x^i

There are $2+C$ parameters that need to be estimated for the cross-perspective topic model: the document-topic distribution θ, the topic-word distribution ϕ, and, for every perspective, the opinion-word distribution ϕ_o^c. As Fang et al. [11], we chose to implement a Gibbs sampler to estimate the parameters [13]. Gibbs sampling is a type of Markov Chain Monte Carlo (MCMC) algorithm [1]. MCMC algorithms aim to construct a Markov chain that have the target posterior as the stationary distribution. In Gibbs sampling, new assignments of variables are sequentially sampled by drawing from the distributions conditioned on the current values of all other variables. To estimate parameters of opinions, additional Markov chains are introduced to simulate the generation of opinions. The sampling equations of the topic variable z for each topic word w_i is as follows. The notation used in these equations is explained in table 2.

$$p(z_i = k | w_i = v, \mathbf{z}_{-i}, \mathbf{w}_{-i}, \alpha, \beta) \propto \frac{n_{k(d)-i} + \alpha}{\sum_{k=1}^{K} n_{k(d)-i} + K\alpha} \times \frac{n_{v(k)-i} + \beta}{\sum_{v=1}^{V} n_{v(k)-i} + V\beta}$$

The sampling equation of opinion variable x^c for each opinion word o_i is

$$p(x_i^c = s | o_i = r, \mathbf{x}_{-i}^c, \mathbf{o}_{-i}, \beta_o) \propto \frac{n_{r(s)-i} + \beta_o^c}{\sum_{r=1}^{T} n_{r(s)-i} + T\beta_o^c} \times \frac{n_{s(d)}}{N_{w(d)}}$$

For every sample thus obtained, the relevant parameters are estimated using the following equations:

$$\theta_{kd} = \frac{n_{k(d)} + \alpha}{\sum_{k=1}^{K} n_{k(d)} + K\alpha}$$

$$\phi_{vk} = \frac{n_{v(k)} + \beta}{\sum_{v=1}^{V} n_{v(k)} + V\beta} \qquad \phi_{rs}^c = \frac{n_{r(s)} + \beta_o^c}{\sum_{r=1}^{T} n_{r(s)} + T\beta_o^c}$$

Our implementation of a Gibbs sampler for cross-perspective topic modeling is available online[4].

4 Study Design

As mentioned in section 2, most existing work on topic models for viewpoint extraction only addresses 'face validity' and model fit. This paper assesses content validity and criterion validity of topics and opinions extracted using cross-perspective topic modeling. In order to do so, we need to determine to what extent topics and opinions correspond to political subjects and political parties' ideological positions.

[4] https://github.com/NLeSC/cptm

Symbol	Description
w, o	Topic word and opinion word
d, v, r, k, s, c	Variable instances; d for document, v for topic word, r for opinion word, k for topic, s for opinion, c for perspective
D, K, C	The number of documents, topics, and perspectives
V, T	The size of the topic and opinion vocabulary
$\mathbf{z}, \mathbf{x}$	Topic and opinion
$\mathbf{w}_{-i}, \mathbf{z}_{-i}, \mathbf{o}_{-i}$	The vector values of $\mathbf{w}_i$, $\mathbf{z}_i$, and $\mathbf{o}_i$ on all dimensions except i
$N_{w(d)}$	The number of topic words in document d
$n_{k(d)[-i]}$	The number of times topic k has occurred in document d [except the current instance]
$n_{v(k)[-i]}$	The number of times word v is assigned to topic k [except the current instance]
$n_{r(s)[-i]}$	The number of times word r is assigned to opinion s [except the current instance]
$n_{s(d)}$	The number of times opinion s occurs in document d
θ	$D \times K$ matrix containing the document-topic distribution
ϕ	$K \times V$ matrix containing the topic-word distribution
ϕ_o^c	$K \times T$ matrix containing the opinion-word distribution for perspective c

Table 2: Notation used in the cross-perspective topic model (adapted from [11]).

This section presents the design of our validation study. After introducing the research questions, we provide a description of the dataset and experiments.

4.1 Research Questions

To assess validity of the topics, the following research questions need to be answered.

- Do the topics learned from the parliamentary proceedings cover all relevant political subjects? (content validity)
- Can the topics learned from the parliamentary proceedings be used to predict the political subject of texts? (criterion validity)

To assess content validity of the topics, we need to determine whether the topics extracted from the data cover all relevant political subjects. As a set of 'all relevant political subjects', we use the Comparative Agendas Project (CAP) main coding categories [5]. The 21 CAP main coding categories are listed in table 4. We check whether there is at least one topic covering each CAP coding category by training a text classifier on manually coded data, and use this classifier to predict CAP codes for the topics extracted from our data. To assess criterion validity, we apply our topic models to the manually coded data, and use the results to predict CAP codes. Performance of this 'classifier' is compared to two other text classifiers; a Naive Bayes classifier and Support Vector Machine (SVM). To make it a fair comparison, the classifiers are trained using the topic words (nouns) only.

To assess validity of the opinions, the following research questions are addressed.

- Are party opinions learned from the parliamentary proceedings representative of party manifestos? (content validity)
- Is there substantial correlation between party rankings generated from the opinions and rankings from domain experts? (criterion validity)

To assess content validity of the opinions, we need to determine whether the topics and associated opinions are representative of a party's ideological position. In order to do so, we estimate opinion word perplexity of party manifestos[5]. We can conclude the opinions have construct validity if there is a correspondence between the perspective that has lowest perplexity for a manifesto and the political party that published it. With regard to criterion validity, we test whether we can use the parties' opinions to rank them on a left/right scale. As a gold standard of the left/right scale we use expert rankings of political party ideology from the Chapel Hill Expert Survey (CHES) [4]. To generate rankings from our data, we apply principal components analysis (PCA) to the parties' opinions and project them on the first few principal components. We can conclude the opinions have criterion validity if there is substantial correlation between these and the CHES rankings.

4.2 Data and Experiments

The data used for validity assessment consists of Dutch parliamentary proceedings from the House of Parliament and Senate between September 21, 1999 and September 11, 2012[6]. It contains 20,594 documents in total. Each document contains speeches that are tagged with a political party. These tags are used to divide the corpus in perspectives. For this study, we made two divisions of the data. In the first division, there is a perspective for each political party. The *parties* dataset consists of 11 perspectives. For the second division, *parties/time*, the data is divided by party and government term. This set contains 59 perspectives. The numbers of documents per perspectives for the two datasets are listed in table 3. Figure 1 presents a timeline of the government terms and the period covered by the dataset. The light blue lines represent the government terms, while the dark line represents the time period covered by the parliamentary proceedings.

Party name	*parties*	*parties/time*					
		K.II	B.I	B.II	B.III	B.IV	R.I
CDA	6416	1165	296	1715	240	1788	1212
CU	2783	332	138	673	77	828	735
D66	4151	941	236	1010	118	813	1033
GL	4960	986	262	1335	163	1176	1038
LPF	846	13	151	666	14	2	-
PvdA	6590	1134	300	1864	263	1685	1344
PvdD	667	-	-	-	20	285	362
PVV	2179	-	-	-	57	1105	1017
SGP	2669	531	139	767	120	712	400
SP	5506	611	199	1307	203	1867	1319
VVD	5976	1054	252	1552	227	1770	1121

Table 3: Number of documents per perspective in the *parties* and *parties/time* datasets.

[5] Party manifestos were obtained through The Manifesto Project (MP) [17]. The MP provides manually coded versions of party manifestos. For the validation, we only used the texts.

[6] http://ode.politicalmashup.nl/data/summarise/folia/

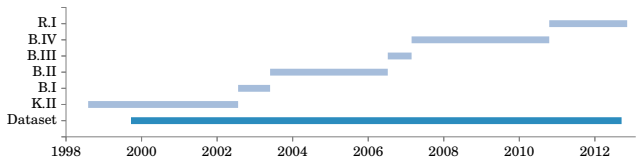

Figure 1: Timeline of period covered by the government terms and parliamentary proceedings dataset.

All text was part-of-speech (POS) tagged and lemmatized using Frog [26]. Nouns are saved as topic words. Because preliminary experiments showed that verbs and adverbs mostly add noise to the opinions, only adjectives are used as opinion words. The topic and opinion vocabularies were filtered. Terms occurring less than six times, the top 100 most frequent terms, and the top 100 terms that occur in the most documents were removed. The topic vocabulary contains 38,145 terms and the opinion vocabulary contains 6245 terms.

The cross-perspective topic model has $2 + C$ Dirichlet hyper parameters: α, β and β^i; $i \in$ perspectives. α affects the number of topics found in a document (lower α leads to fewer topics per document), whereas β affects the number of words a topic consists of (lower β leads to topics with fewer words). Because the values of these parameters mostly affect the convergence of Gibbs sampling and not the results [13], we fix them to $\alpha = 50/K$, and $\beta = \beta^i = 0.02$ (cf. [13]). Based on previous experience with the Dutch parliamentary proceedings, the number of topics (K) is set to 100. Gibbs sampling is done for 200 iterations, and the final θ, topics, and opinions are calculated by taking the average of every tenth iteration starting from iteration 80.

5 Results

In this section, we answer the research questions formulated in section 4. Sections 5.1 and 5.2 address topic and opinion validity respectively.

5.1 Topic Validity

This section addresses content and criterion validity of the topics. For the assessment of content validity, topics are divided into two sets: high quality topics and low quality topics. Topic quality is determined by calculating topic coherence measure *NPMI* [7, 22] using the Dutch Wikipedia as a reference corpus [27]. Topics are considered to be of high quality if their *NPMI* score is above the mean.

5.1.1 Content Validity

In order to determine whether all political subjects are covered by our two sets of 100 topics, we train a text classifier that predicts CAP main categories based on manually coded data. The dataset we use are manually coded parliamentary questions[7] [8] [24]. To train text classifiers, we selected the parliamentary questions texts from September 21, 1999 to September 11, 2012. One text that was not coded properly was removed. The resulting dataset consists of 834 texts. Table 4 lists the percentages of texts coded with each CAP main coding category. There are three CAP codes that do not occur in this dataset: 9, 18, and 23. These are excluded from analysis. Note that CAP codes 11 and 22 do not exist and are therefore also excluded.

CAP	Description	% texts
1	Domestic Macroeconomic Issues	3.36
2	Civil Rights, Minority Issues, and Civil Liberties	8.63
3	Health	10.55
4	Agriculture	2.88
5	Labor and Employment	9.47
6	Education	8.03
7	Environment	2.40
8	Energy	2.52
9	Immigration and Refugee Issues	0.00
10	Transportation	5.52
12	Law, Crime, and Family Issues	12.95
13	Social Welfare	7.19
14	Community Development and Housing Issues	4.08
15	Banking, Finance, and Domestic Commerce	3.48
16	Defense	3.36
17	Space, Science, Technology, and Communications	1.68
18	Foreign Trade	0.00
19	International Affairs and Foreign Aid	7.67
20	Government Operations	5.64
21	Public Lands, Water Management, and Territorial Issues	0.60
23	Cultural Policy Issues	0.00

Table 4: CAP main codes and percentages of texts in the parliamentary questions dataset.

Topic words (nouns) extracted from the texts were used to train two classifiers: a Naive Bayes classifier and SVM. Results reported in table 5 were obtained through 5-fold cross validation. Performance of the SVM is significantly better than performance of the Naive Bayes classifier (Welch's two-sided t-test, $p < 0.05$). Based on these results the SVM was selected to map topics to coding categories.

Text classification was performed on the top 10 topic words of our two sets of 100 topics. Figures 2 and 3 show the numbers of topics that were mapped to the different CAP coding categories for all topics and the high quality topics. The figures show that all CAP coding categories are covered by the topics.

[7] In addition to the parliamentary questions data, there is a second dataset available: the Queen's speeches [8]. However, a text classification experiment using the Queen's speeches resulted in very low performance (Naive Bayes $F_1 \approx 0.06$; SVM $F_1 \approx 0.07$). This can be explained by the fact that the Queen's speeches data consists of coded sentences that are to short to learn anything from (20.59 words on average; std = 10.40). We therefore decided not to use the Queen's speeches data to assess content validity.

[8] The parliamentary questions data has the disadvantage that the texts are also part of the parliamentary proceedings. However, given that there is no other suitable manually coded dataset available, and that these texts make up a very small part of the parliamentary proceedings, this data was used to assess content validity.

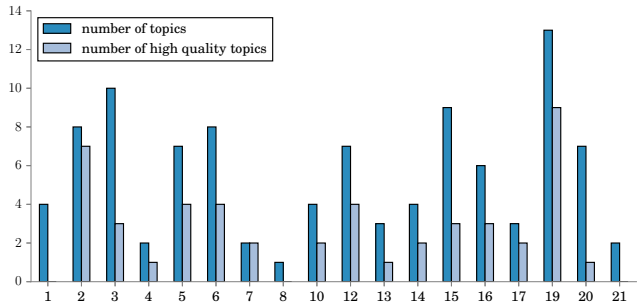

Figure 2: Number of topics and high quality topics from the *parties* dataset that is predicted for each main CAP code.

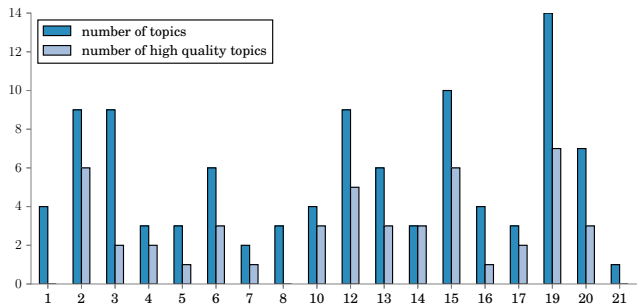

Figure 3: Number of topics and high quality topics from the *parties/time* dataset that is predicted for each main CAP code.

With regard to the high quality topics: the *parties* dataset does not contain high quality topics for CAP codes 1, 8, and 21; in the *parties/time* dataset there are no high quality topics for CAP code 21. As shown in table 4, these coding categories are relatively rare in the parliamentary questions data (0.60% - 3.36%). Based on these results, we conclude that topics extracted using cross-perspective topic modeling have content validity.

5.1.2 Criterion Validity

To assess criterion validity, we use the mapping between topics and CAP coding categories described in the previous section to predict CAP codes of the manually coded parliamentary questions texts. In order to do so, we estimate θ for the parliamentary questions texts using ϕ^{topic} obtained through the experiments. Then, the most important topic found for a text is mapped to a CAP category. Because the results of topic modeling are probabilistic, this procedure was repeated 10 times. Classification performance is calculated using accuracy and F_1. Table 5 presents the results of the Naive Bayes classifier (baseline), SVM, and the topic model classifiers.

All differences in performance are statistically significant at $p < 0.05$, except the differences between the two topic model classifiers. The results for the topic model classifiers are higher than the results obtained with the Naive Bayes classifier (baseline) and lower than the performance of the SVM. When interpreting the results, it is important to keep in mind that the Naive Bayes and SVM are algorithms for supervised learning, whereas topic modeling is unsupervised.

	Accuracy	F_1
Naive Bayes	0.447 ± 0.026	0.377 ± 0.028
SVM	0.603 ± 0.030	0.585 ± 0.032
Topic models *parties*	0.537 ± 0.009	0.529 ± 0.008
Topic models *parties/time*	0.550 ± 0.008	0.548 ± 0.008

Table 5: Machine learning performance of parliamentary questions data

We are not so much interested in classifier performance per se, but instead investigate whether there is meaningful correlation between the most important topic in a text and manually assigned CAP codes. Because the SVM was used to map topics to CAP codes, its performance effectively is an upper bound for the performance of the topic model classifiers. The closer performance of the topic model classifiers is to performance of the SVM, the better. Based on the results found, we conclude that the topics have criterion validity.

5.2 Opinion Validity

This section addresses content and criterion validity of the opinions.

5.2.1 Content Validity

To assess content validity of the opinions, we use party manifestos as implicit, but complete representations of political parties' viewpoints on the topics they find most important. To assess whose opinion is expressed in party manifesto d, we need to calculate

$$\underset{i \in C}{\operatorname{argmax}}\, p(d|o^i)$$

Assuming equal probabilities for each perspective (political party), $p(d|o^i) \propto p(o|d)$. $p(o|d)$ is calculated as the opinion word perplexity for party manifesto document d:

$$perplexity(d) = exp - \frac{log(p(\mathbf{o}))}{N_o}$$

where

$$p(\mathbf{o}) = \prod_{i=1}^{N_o} \sum_{k=1}^{K} p(o_i|z_i = k) p(z_i = k|d)$$

In this equation $\mathbf{o}$ is the set of opinion words in document d; N_o is the number of opinion words in document d; $p(o_i|z_i = k)$ is learned from the original experiments on the parliamentary proceedings; and $p(z_i = k|d)$ is estimated from the party manifestos using parameters estimated in the original experiments.

Dutch party manifestos from 2006 onwards are available in the Party Manifesto Project dataset. We downloaded manifestos for the elections in 2006, 2010, and 2012 for all parties except LPF, which is not present in the Party Manifesto Project dataset. Documents were subjected to the same preprocessing procedure as the parliamentary proceedings, and per document opinion word perplexity was calculated as described above.

As shown in table 6, the party with lowest perplexity for the *parties* dataset is correct for 66.67% of the party manifestos. The confusion matrix in figure 4 shows that mistakes

are made all over the political spectrum. First, the opinions of two confessional parties CDA and CU can't be distinguished. Also, SGP, which is a more conservative confessional party is confused with CU for one of the three manifestos. On the left side of the political spectrum, PvdD is confused with GL and/or SP, which are all left-wing parties. Parties closer to the center of the political spectrum, D66, VVD, and PvdA, are also confused. These results are in line with a common preconception of Dutch politics that although there is a multiparty system, the differences between individual parties are small.

	Accuracy
parties	0.667
parties/time previous	0.300
parties/time next	0.100
parties/time parties	0.567

Table 6: Accuracy of predicting the parties of party manifestos.

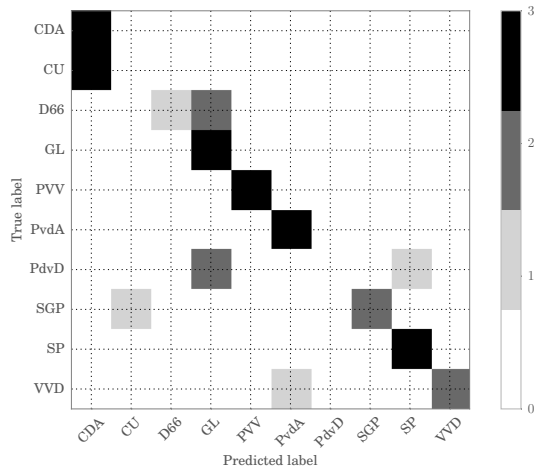

Figure 4: Confusion matrix of predictions for party manifestos based on perplexity calculated using parameters learned from the *parties* data.

For the *parties/time* data, the time parameter has to be taken into account. Because it is not clear in advance whether party manifestos represent a party's viewpoints of the previous or next government term, accuracy was calculated for both these possibilities. The results in table 6 show that political parties' opinions of the government term before an election are more similar to the party manifestos than the opinions of the next government term (accuracy of 0.300 and 0.100 respectively). Figure 5 shows the confusion matrix of the *parties/time* previous results. The black squares on the diagonal show that there is correlation between the actual party and government term and what is predicted. The results are distorted by two vertical lines, one over B.IV-CDA, the other over B.IV-PvdA. Showing the dominance of center parties (CDA and PvdA), and the government term B.IV, which is the longest government term in the time period we have party manifestos for (2006–2012). These results can be explained by the fact that there are more documents available for government term B.IV than for B.III.

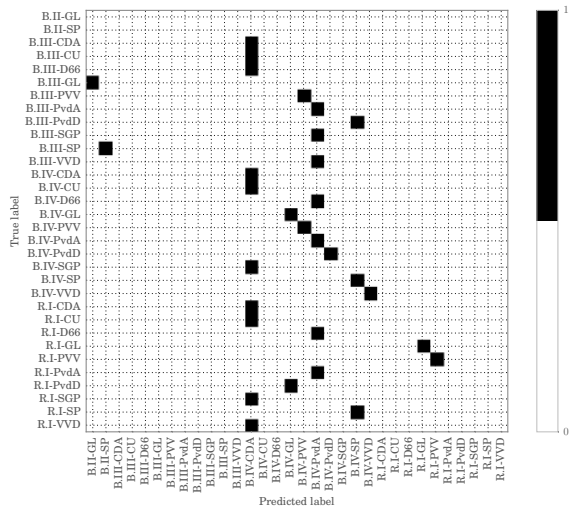

Figure 5: Confusion matrix of predictions for party manifestos based on perplexity calculated using parameters learned from the *parties/time* data.

When removing the time dimension from the predictions, accuracy increases to 0.567. This is slightly lower than accuracy for the *parties* opinion perplexity experiment. The confusion matrix is very similar to figure 4 and is not displayed. Based on the results found, we conclude that the opinions have content validity.

5.2.2 Criterion Validity

To assess criterion validity of the opinions, we use the opinions to generate rankings of perspectives and compare these rankings to rankings of political parties in the CHES dataset [4]. The CHES dataset is based on expert knowledge, and contains estimates of political party positions on different subjects, including European integration, ideology, and policy issues for national parties in different European countries. The survey is repeated every few years. For this study, we use data from 1999, 2002, 2006, and 2010. Traditionally, one of the most important scales political parties are ranked on is the left/right spectrum. The CHES dataset contains two variables that are relevant to this scale: *lrgen*, which measures ideological stance (left/right spectrum), and *lrecon*, which measures ideological stance on economic issues. Rankings generated from the opinions are compared to rankings based on these two variables.

Rankings of the different perspectives based on the opinions learned from the *parties* and *parties/time* data are generated by doing PCA on the opinions. For each dataset, we create rankings of perspectives by projecting the opinions on the first 5 principal components. CHES rankings for for *parties* are generated by averaging *lrgen* and *lrecon* over parties. For the *parties/time* data, years are mapped to government terms, and averaged over party/government term combinations.

To compare the rankings, we calculate Kendall's Tau [15] and Spearman's r [16]. The results for the *parties* dataset are presented in table 7. There are very few significant results.

We conclude that there is no linear relation between opinions from the *parties* and CHES *lrgen* or *lrecon*. Table 8 presents the results for the *parties/time* data. These results are very similar to the results for *parties*. Again we conclude that a linear relation between opinions from the *parties* and CHES *lrgen* or *lrecon* does not exist. This means that there is no criterion validity with regard to the left/right distinction.

	PC 1	PC 2	PC 3	PC 4	PC 5
Kendall's Tau					
lrgen	0.382	0.236	0.127	-0.273	0.127
lrecon	0.164	0.164	0.055	-0.636†	-0.091
Spearman's r					
lrgen	0.364	0.136	-0.882†	0.055	0.164
lrecon	0.209	0.173	-0.827†	0.427	0.200

Table 7: Correlation between opinions learned from the *parties* dataset projected on the first five PCA components and CHES *lrgen* and *lrecon* († statistically significant at $p < 0.05$).

	PC 1	PC 2	PC 3	PC 4	PC 5
Kendall's Tau					
lrgen	-0.191	0.094	0.037	-0.077	0.009
lrecon	-0.048	0.031	0.066	-0.060	0.140
Spearman's r					
lrgen	0.153	0.123	0.042	0.010	0.537†
lrecon	0.093	0.045	0.076	0.038	0.643†

Table 8: Correlation between opinions learned from the *parties/time* dataset projected on the first five PCA components and CHES *lrgen* and *lrecon* († statistically significant at $p < 0.05$).

6 Discussion

In this paper, we explore a number of validation methods, and demonstrate that validating the results of topic modeling is feasible, even without extensive domain knowledge. The results of our study reveal that cross-perspective topic modeling is a promising technique for extracting political parties' positions from parliamentary proceedings. We have shown that the topics have content and criterion validity, and the opinions have content validity. However, in order to be able to apply cross-perspective topic modeling, the data must be divided (or at least dividable) into perspectives. Because not all datasets meet these requirements, the CPT model certainly does not solve all viewpoint extraction tasks.

For criterion validity of the opinions, we tried to use opinions to rank political parties on a left/right scale. The results indicate that the differences between the opinions of political parties are more complicated. There are two possible solutions to this problem. First, there might other valid domain-specific interpretations of the rankings generated by applying PCA to the opinions, such as standpoints towards European integration, or socio-cultural liberal-conservative dimensions. Unfortunately, because for Dutch political parties this data is not available in the CHES data, we have been unable to test these hypotheses. Generally speaking, solving the criterion validity problem of political parties' positions, requires additional domain knowledge.

Another way that might help to correct the rankings generated from the opinions is by improving the quality of topics and opinions, as it is known topics learned from the parliamentary proceedings are noisy [3]. In the original paper, Fang et al. used an elaborate method involving supervised machine learning to select sentences containing opinion words [11]. The resources required to do this for Dutch data do not exist, and would therefore need separate validation. However, higher quality opinions could lead to better results. Another possibility to improve opinion quality is to impose constraints on the dataset, as done in VODUM [23]. Especially the constraint that words in the same sentence are assigned to the same topic might help to reduce noise in topics and opinions. Finally, topic and opinion quality could be improved by applying postprocessing techniques. For example, parsimonization might be used to select high quality topic and opinion terms [2].

In addition to content and criterion validity, there is also construct validity. Construct validity refers to the extent to which a measure behaves as expected in a theoretical context. Assessing this aspect of validity requires extensive domain knowledge, which is why we did not include construct validity in our study. What we have shown, however, is that extensive domain knowledge is not required in order to validate a topic model. What is required, of course, is the availability of external data to validate against.

The validation of new topic modeling methods is also impeded by the fact that researchers who introduce new models rarely provide implementations of these models. Of the work discussed in section 2, only an implementation of VODUM [23] was made available. To facilitate validation studies, providing access to source code of new algorithms would be very helpful.

7 Conclusion

This paper presented a validation study of cross-perspective topic modeling [11] using Dutch parliamentary proceedings. The results show that the method yields valid topics (content and criterion validity). While opinions were found to be representative of the political parties' positions as expressed in party manifestos (content validity), we were unable to find correlation between opinions and positions on the left/right political spectrum (criterion validity). Further work is required to determine whether differences between opinions correlate with other politically meaningful dimensions. We also propose to investigate the effect of improving topic and opinion quality on the validation results.

The second contribution of this paper is that we show validation studies are feasible, even without extensive domain knowledge. We contend that in order for topic models to be useful, the results must be semantically meaningful to humans. Because anecdotal qualitative evaluations and/or assessments of model fit fail to capture this essential aspect, validation of results is required before researchers from other domains will apply these methods.

Acknowledgements

The authors would like to thank Kostas Gemenis and Andreas Warntjen for valuable suggestions with regard to this study.

References

[1] C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan, 'An introduction to MCMC for machine learning', *Machine learning*, **50**(1-2), 5–43, (2003).

[2] H. Azarbonyad, M. Dehghani, T. Kenter, M. Marx, J. Kamps, and M. de Rijke, 'Measuring Topical Diversity of Text Documents Using Hierarchical Parsimonization', (Under submission).

[3] H. Azarbonyad, F. Saan, M. Dehghani, M. Marx, and J. Kamps, 'Are Topically Diverse Documents Also Interesting?', in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pp. 215–221, (2015).

[4] R. Bakker, E. Edwards, L. Hooghe, S. Jolly, J. Koedam, F. Kostelka, G. Marks, J. Polk, J. Rovny, G. Schumacher, M. Steenbergen, M. Vachudova, and K. Zilovic. 1999-2014 Chapel Hill Expert Survey Trend File. http://chesdata.eu/, 2015.

[5] S. Bevan. Gone Fishing: The Creation of the Comparative Agendas Project Master Codebook. http://sbevan.com/cap-master-codebook.html, 2014.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, 'Latent dirichlet allocation', *the Journal of machine Learning research*, **3**, 993–1022, (2003).

[7] G. Bouma, 'Normalized (pointwise) mutual information in collocation extraction', in *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology (GSCL '09)*, pp. 31–40, (2009).

[8] G. Breeman, D. Lowery, C. Poppelaars, S. L. Resodihardjo, A. Timmermans, and J. de Vries, 'Political attention in a coalition system: Analysing Queen's speeches in the Netherlands 1945–2007', *Acta Politica*, **44**(1), (2009).

[9] E. G. Carmines and R. A. Zeller, *Reliability and validity assessment*, Sage publications, 1979.

[10] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei, 'Reading tea leaves: How humans interpret topic models', in *Advances in Neural Information Processing Systems*, pp. 288–296, (2009).

[11] Y. Fang, L. Si, N. Somasundaram, and Z. Yu, 'Mining contrastive opinions on political texts using cross-perspective topic model', in *the fifth ACM international conference on Web search and data mining*, pp. 63–72, (2012).

[12] S. Gottopati, M. Qiu, Y. Sim, J. Jiang, and N. Smith, 'Learning topics and positions from debatepedia', in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1858–1868, (2013).

[13] T. L. Griffiths and M. Steyvers, 'Finding scientific topics', *Proceedings of the National Academy of Sciences*, **101**(1), 5228–5235, (2004).

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

[15] M. G. Kendall, 'A new measure of rank correlation', *Biometrika*, **30**(1), 81–93, (1938).

[16] E. L. Lehmann and H. J. M. D'Abrera, *Nonparametrics: Statistical Methods Based on Ranks*, Springer, 1998.

[17] P. Lehmann, T. Matthieß, N. Merz, S. Regel, and A. Werner. Manifesto Corpus. Version: 2015-5. https://manifestoproject.wzb.eu/, 2015.

[18] W. H. Lin, E. Xing, and A. Hauptmann, 'A joint topic and perspective model for ideological discourse', in *Machine Learning and Knowledge Discovery in Databases*, 17–32, Springer Berlin Heidelberg, (2008).

[19] B. Pang and L. Lee, 'Opinion mining and sentiment analysis', *Foundations and trends in information retrieval*, **2**(1-2), 1–135, (2008).

[20] M. Paul and R. Girju, 'A two-dimensional topic-aspect model for discovering multi-faceted topics', in *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10)*, pp. 545–549, (2010).

[21] K. M. Quinn and B. L. Monroe, 'How to analyze political attention with minimal assumptions and costs', *American Journal of Political Science*, **54**, 209–228, (2010).

[22] M. Röder, A. Both, and A. Hinneburg, 'Exploring the space of topic coherence measures', in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pp. 399–408, (2015).

[23] T. Thonet and G. Cabanac, 'VODUM: a Topic Model Unifying Viewpoint, Topic and Opinion Discovery', in *Advances in Information Retrieval*, 533–545, Springer, (2016).

[24] A. Timmermans and G. Breeman, 'Morality issues in the Netherlands: coalition politics under pressure', in *Morality Politics in Western Europe*, 35–61, Palgrave Macmillan UK, (2012).

[25] A. Trabelsi and O. R. Zaiane, 'Mining contentious documents using an unsupervised topic model based approach', in *Data Mining (ICDM), 2014 IEEE International Conference on*, pp. 550–559, (2014).

[26] A. van den Bosch, B. Busser, S. Canisius, and W. Daelemans, 'An efficient memory-based morphosyntactic tagger and parser for Dutch', *LOT Occasional Series*, **7**, 191–206, (2007).

[27] J. M. van der Zwaan, M. Marx, and J. Kamps. Palmetto position storing Lucene index of Dutch Wikipedia. http://dx.doi.org/10.5281/zenodo.46377, 2016.