Radboud University Nijmegen

Bachelor's thesis in Artificial Intelligence

Modeling and forecasting elections using topic models

Author:
Bas van Berkel

Supervisors:
Prof. Tom Heskes
Dr. Louis Vuurpijl

July 8, 2014

Abstract

After elections, people want to know who will win the election as quickly as possible. During election night forecasts are made, based on both polls and the results from early polling stations. In this research a forecasting model based on topic models is proposed. The model's forecasting performance is compared to a linear model's performance using the Dutch House of Representatives election. The model is also used to visualize and analyze voting profiles. The proposed model outperforms existing linear models with a lower mean absolute error if 10% or more of the polling stations are observed. For 2.5% or less observed polling stations the linear model has a lower mean absolute error. The proposed model is also able to give insight into voting behavior by visualizing voter profiles. Thus, the proposed model is useful for both forecasting and modeling elections.

Contents

1 Introduction

1.1 Election forecasting

Every few years there are elections for the House of Representatives. These days polls form a big part of elections. During election day exit polls are held, and during the counting of the results, preliminary results are calculated. After the election, maps with results per municipality and figures with voter shifts are printed in papers. This indicates an interest of the general public to know, what the election result will be as early as possible and what the underlying voting behavior is. A lot of this information is generated using polls and surveys. During election night television programs show incoming results and forecast the election results. This is known as election night forecasting. Especially during the early hours of the election night it is hard to predict an accurate outcome [4]. A first problem arises due to the fact that different regions and types of cities have a different voting behavior. This issue is increased by the fact that the order in which results are declared bymunicipalities does not necessarily result in a representative sample. In most elections small municipalities race to be the first to declare their results, whereas counting all votes in large cities might take longer. This can cause a biased result when only a small part of the results is observed. Another problem is that municipal borders change over time [4]. In the Netherlands the number of municipalities dropped from over 900 in 1963 [8] to just over 400 in 2012. Voting data is available for the Dutch House of Representatives elections since 1963. This data contains the number of votes for each party, municipality, and election.

1.2 Topic models

Topic models can be used to categorize documents based on topic or theme. Topic models try to uncover the topic of documents based on the words used in that document. Although topic models are designed for, and mainly used to, analyze large sets of text documents [1], it is also possible to use them in other applications. Past uses have been in bio-informatics [9], image analysis [15] and the analysis of social networks [11]. Multiple types of topic models can be identified: dynamic [3], correlated [2], supervised [10] and probabilistic [1] topic models, amongst others. The probabilistic topic modeling technique probabilistic Latent Semantic Analysis (pLSA) [7] is suitable for modeling elections because it is possible to apply pLSA to higher order data. For topic models a bag of words assumption is made, this means that the order in which words appear in the document is not taken into account by the model [1]. This bag of words assumption holds for elections, for the final result it does not matter in which order voters cast their vote.

The idea of pLSA is that each topic is a probability distribution over words in a vocabulary and each document is a probability distribution over topics. Documents can cover multiple topics [1]. When a word is in a given document, this is used as evidence for the document covering the topics which this word is part of. $P(d, w)$ denotes the probability that word w and document d co-occur. In the asymmetric formulation below, for each latent topic z, the probability that this is the topic of the given document is multiplied with the probability that the given word is generated by the topic. The sum over all topics results in the probability of the word occuring in the document. If this probability is combined with the probability of the document, we get the probability that word w and document d co-occur:

$$P(d, w) = P(d)P(w|d), P(w|d) = \sum_{z=1}^{Z} P(w|z)P(z|d).$$

To obtain the probability distribution over words of each topic and the distribution over topics of each document an expectation maximization (EM) algorithm is used. The EM-algorithm will be explained in the section on the proposed model.

In the proposed model individual votes correspond to word occurrences in pLSA. A party corresponds to a word in the vocabulary, a polling station to a document, and voting profiles to the latent topics.

1.3 Research question

In this research a topic model based on pLSA is proposed to model voter profiles and to forecast elections. The main questions are whether this model can give insight in voting profiles, their distribution within municipalities, and whether the proposed model can be used to forecast elections, given partial results of those elections. Before testing the model on forecasting performance it is important to find the optimal settings: the optimal number of topics given the number of observed municipal results and the number of previous elections known. To test the forecasting performance of the proposed model a linear model proposed by Pavía-Miralles [14] is used for comparison. Both models get the same information: number of previous elections known and number of observed municipal results in the current election.

2 Methods

First the proposed model and the linear model, which were used for comparison, are explained. In the second subsection the data is explained and in the last subsection the procedure of the expirement is described.

2.1 Model

2.1.1 Proposed model

The model proposed is an extended version of pLSA. For a single election the model is equivalent to pLSA. In the introduction pLSA is described using words and documents. In the proposed model individual votes correspond to word occurrences in pLSA. A party corresponds to a word in the vocabulary, a polling station to a document, and a voting profile to a topic. The main difference between pLSA and the proposed model is that there is another dimension of information, the dimension of the different elections. pLSA is extended in the sense that there is a topic distribution over parties per election. The distribution over voting profiles per polling station is assumed to be constant. Therefore, there is only one distribution over profiles per polling station, which is fixed over elections. Let e denote elections, where e_i is the i^{th} election. Let n denote voters spread over polling stations s with n_j voters in polling station s_j. There are parties w, where w_k denotes party k. There are voting profiles v, where v_m is voting profile m. $d_{kj}(i)$ denotes the actual number of votes on party k in polling station j during election i. We define $P(w_k|s_j, e_i)$ as the probability of a vote at party k in polling station j during election i:

$$P(w_k|s_j, e_i) = \sum_{m=1}^{V} P(w_k|v_m, e_i)P(v_m|s_j) = \sum_{m=1}^{V} \beta_{kmi}\theta_{mj}.$$

This is obtained by multiplying the probability that a voter in voting profile m votes on party k in election i: β_{kmi}, with the probability that a voter in polling station j is part of voting profile m: θ_{mj}. And summing over all voting profiles.

The total proportion of votes $P(w_k|e_i)$ on party k in election i is calculated by the weighted sum of votes on each party k per polling station:

$$P(w_k|e_i) = \frac{\sum_{j=1}^{S} P(w_k|s_j, e_i)n_j}{N}.$$

Before being able to use the distributions over voting profiles per polling station θ_{mj}, and the distribution of parties over voting profiles β_{kmi}, their optimal values have to be found. The optimal values for these distributions can be extrapolated using the following expectation maximization algorithm which locally optimizes the log-likelihood. The algorithm's two steps are iteratively updated. In the expectation step posterior probabilities for the latent distributions are computed:

$$P(v_m|s_j, w_k, e_i) = \frac{P(w_k|v_m, e_i)P(v_m|s_j)}{\sum_{l=1}^{V} P(w_k|v_l, e_i)P(v_l|s_j)}.$$

In the maximization step β_{kmi} and θ_{mj} are updated:

$$\beta_{kmi} = P(w_k|v_m, e_i) \propto \sum_j d_{kj}(i)P(v_m|s_j, w_k, e_i),$$

$$\theta_{mj} = P(v_m|s_j) \propto \sum_{k,i} d_{kj}(i) P(v_m|s_j, w_k, e_i).$$

To forecast an election i the distribution over topics per polling station θ_{mj} is based on previous elections. The distribution over parties has to be estimated based on the observed polling stations s_j. This can be done with the EM-algorithm, however, in the maximization step only β_{kmi} should be updated.

2.1.2 Linear model

In 2005 Pavía-Miralles proposed a linear based model to estimate the proportion of votes each party obtains [14]. The model works with a linear prediction per party, based on the observed polling stations. $s_{jk}(i)$ denotes the result of party k at polling station j during election i. E is the current election. The results of the current election are predicted based on the assumption that:

$$P(w_k|s_j, e_E) = s_{jk}(E) = \alpha_k + \sum_{i=1}^{E-1} \beta_k(i) s_{jk}(i) + r_{kj}.$$

Here $P(w_k|s_j, e_i)$ and $s_{jk}(E)$ denote the probability of a vote at party k in polling station j during the last election E. α_k and $\beta_k(i)$ are party specific regression coefficients. The parameters α_k and $\beta_k(i)$ are estimated by applying non-negative least squares to the results of party k in the observed polling stations. Noise is modeled by r_{kj}..

The total proportion of votes $P(w_k|e_E)$ for party k during the last election V is calculated by the weighted sum of votes per polling station:

$$P(w_k|e_E) = \frac{\sum_{j=1}^{S} s_{jk}(E) n_j}{N}.$$

Non-negative least squares [5] is used, because the number of votes on a party is always positive. Another condition is that the sum of the proportions of votes each party obtains in a polling station should be 1. Thus, $\sum_{k=1}^{W} s_{jk}(V) = 1$ should hold. This is achieved by normalizing the estimated distribution of votes.

2.2 Data

In this research a data set containing the results of the elections for the Dutch House of Representatives (Tweede Kamer) is used. The Dutch House of Representatives election is a multiparty election where the number of seats obtained is proportional to the number of votes received. The data set contains election results per municipality since 1963. Thus, instead of polling stations, used in the description of the models in the previous section, municipalities are used.

The data was initially retrieved from the Dutch Electoral Council for research on the relation between weather and voter turnout [6]. And cleaned up by combining results from municipalities which merged between 1963 and 2012. The municipality borders of 2012 have been used in the research. Non existing data, such as the number of votes on parties not participating in an election have been set to 0. There are also a few missing elections in certain municipalities, this is because these municipalities did not exist during that election. Missing municipalities in some election years are: Dronten (1963-1971), Lelystad (1963-1971), Nunspeet (1963-1971), and Zeewolde (1963-1982). In total there are 415 municipalities in the final data set.

Combined over 16 elections, 18 parties, invalid votes, and other parties are identified. Table 1 provides an overview of the parties in the data set and the years these parties participated in the House of Representatives election.

Party	1963	'67	'71	'72	'77	'81	'82	'86	'89	'94	'98	2002	'03	'06	'10	'12
PvdA	y	y	y	y	y	y	y	y	y	y	y	y	y	y	y	y
CDA	y	y	y	y	y	y	y	y	y	y	y	y	y	y	y	y
VVD	y	y	y	y	y	y	y	y	y	y	y	y	y	y	y	y
D66	.	y	y	y	y	y	y	y	y	y	y	y	y	y	y	y
GL	y	y	y	y	y	y	y	y	y	y	y	y	y	y	y	y
SP	.	.	.	.	.	.	.	.	.	y	y	y	y	y	y	y
CU	y	y	y	y	y	y	y	y	y	y	y	y	y	y	y	y
SGP	y	y	y	y	y	y	y	y	y	y	y	y	y	y	y	y
PvdD	.	.	.	.	.	.	.	.	.	.	.	.	.	y	y	y
PVV	.	.	.	.	.	.	.	.	.	.	.	.	.	y	y	y
LPF	.	.	.	.	.	.	.	.	.	.	.	y	y	.	.	.
Ln	.	.	.	.	.	.	.	.	.	.	.	y	.	.	.	.
DS70	.	.	y	y	y	.	.	.	.	.	.	.	.	.	.	.
NMP	.	.	y	.	.	.	.	.	.	.	.	.	.	.	.	.
RKPN	.	.	.	y	.	.	.	.	.	.	.	.	.	.	.	.
BP	y	y	y	y	y	.	.	.	.	.	.	.	.	.	.	.
CPCD	.	.	.	.	.	.	y	.	y	y	.	.	.	.	.	.
AOV55	.	.	.	.	.	.	.	.	.	y	.	.	.	.	.	.
Other	y	y	y	y	y	y	y	y	y	y	y	y	y	y	y	y
Invalid	y	y	y	y	y	y	y	y	y	y	y	y	y	y	y	y

Table 1: Overview of parties in the data set: the election years in which these parties have participated are marked y. Other indicates other (small) parties participating. *Invalid* stands for none of the above (blanco) and invalid votes.

2.3 Procedure

The proposed model has been tested both qualitatively and quantitatively. For the quantitative test the proposed model and the linear model were both tested on their forecasting performance. Besides forecasting, the proposed model can be used to visualize voting behavior. Because the proposed model has latent distributions over voting profiles. These latent topic distributions were visualized and analyzed.

2.3.1 Election forecasting

To test the election forecasting performance of the proposed model, the model was compared with the linear model described earlier, which was proposed by Pavía-Miralles [14]. The models were tested with increasing numbers of municipalities with observed results: 2 (0.5%), 4 (1%), 10 (2.5%), 42 (10%), and 104 (25%). This allows the performance at different stages of an actual election night to be simulated. The observed municipalities were selected using random sampling. Both models were tested with the same samples of observed municipalities. A problem which can arise is that too much data of previous elections causes under- or overfitting in one or both models. Therefore, the models were tested with different numbers of previous elections known. The dataset contains 16 elections in total. The last election was the election to forecast. The numbers of previous elections tested with were: 1, 3, 5, 8, 11, and 14.

The proposed model forecasts the election using a number of topics. Before comparing both models, the optimal number of topics had to be selected. A low number of topics could cause

underfitting, but too many topics could cause overfitting. A priori one would expect that the optimal number of topics might depend on the percentage of results observed in the current election and on the number of previous elections known. Therefore, the optimal number of topics was determined for all 30 combinations of observed municipalities and previous election results. The 2012 election was used to compare the two models, to prevent overfitting on a particular election the 2010 election was used to determine the optimal number of topics. Determining the optimal number of topics was done by running the proposed model 100 times with multiple numbers of topics for each combination of variables. The numbers of topics were compared on their mean absolute error at a national level. The mean absolute error is equivalent to Mosteller Measure 3 [16]: *"The average (without regard to sign) of the percentage point deviation for each candidate between his/her estimate and the actual vote."* [13]. Mosteller Measure 3 and 5 are recommended by Mitofsky (1998) to assess poll accuracy [12]. Mosteller Measure 5 only takes the top two contesting parties into account. Because the Dutch House of Representatives elections has more than two contesting parties, Mosteller Measure 3 is more suitable.

When the optimal number of topics was found for each combination of variables the two models were compared on the mean absolute error between their forecasts and the actual election results, both on a national and municipal level.

2.3.2 Voting profiles

The latent topics of the proposed model represent different voter profiles. When the model is fitted on data these voting profiles are estimated. An interesting aspect of the proposed model is that one can easily visualize and analyze these voting profiles. An example of a visualization of voting profiles can be found in section 3.3. When analyzing voting profiles one would expect that voters from one party can be found in the same profile across different elections, a continuation in profiles is expected. It is also possible to analyze which parties are together in voting profiles. Using the distribution over profiles of municipalities it is possible to identify municipalities with similar voting behavior.

3 Results and discussions

3.1 Optimal number of topics

Figure 1 shows the results of the search for the optimal number of topics, with 5 previous elections used to generate the topics. In figure 4 in Appendix A the full results for 1, 3, 5, 8, 11, and 14 previous elections can be found. Each subfigure of figure 4 has a different number of elections which are used to determine the distribution over topics per municipality. Each line represents a different number of observed municipalities. The x-axis states the number of topics used. For each combination of number of topics, observed municipalities, and previous elections one model is generated, which is used to forecast the 2010 election 100 times, each time using randomly selected observed municipalities. For all numbers of elections a similar pattern is visible. The standard error of the mean absolute error varies between 3% and 6% of the mean absolute error. This is clearly visible in figure 1 and 4. For all numbers of elections the mean absolute error seems to improve with an increasing number of topics, there are no signs of overfitting. Although there is no overfitting the mean absolute error does saturate with an increasing number of topics. For all numbers of observed elections the mean absolute error seems to saturate somewhere between 32 and 64 topics. Therefore, using 50 topics for the comparison of the proposed model with the linear model, for all combinations of observed municipalities with previous elections, seems enough to prevent and does not cause overfitting.

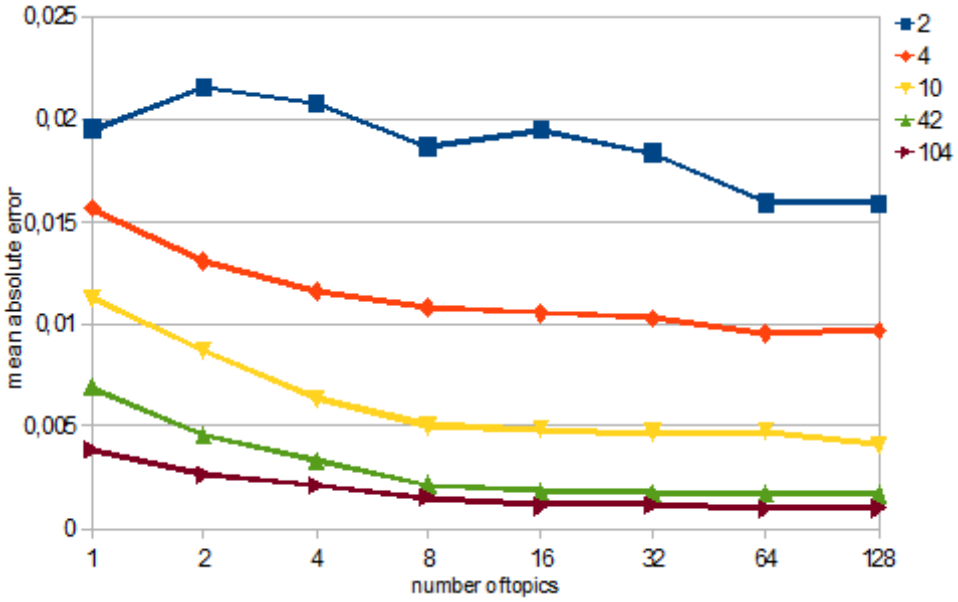


Figure 1: The mean absolute error for different numbers of topics and percentages of observed municipalities, with 5 previous elections

3.2 Forecasting performance

Table 2 and 3 show the forecasting performance of both the proposed and the linear model. For the proposed model one distribution over topics per municipality is generated for each combination of the number of observed municipalities with the number of previous elections known. Using this distribution over topics, which is based on the previous elections, the distri-

bution over parties per topic, for the new election, is estimated, based on results in the observed municipalities. This is repeated 500 times, using randomly selected municipalities. The same samples are used to forecast the election with the linear model. After each forecast, the mean absolute error between the forecast and the actual result is calculated, for the combined national result. This results in 500 paired mean absolute errors. A two-sided paired t-test is used to compare the results of both models. Standard deviations of both models are also shown to give insight into the overlap in the performance of both models.

2 observed municipalities						
Nr. of Elections	14	11	8	5	3	1
MAE prop	0.0178	0.0175	0.0172	0.0171	**0.0166**	0.0167
MAE lin	0.0098	0.0091	0.0087	0.0077	**0.0076**	0.0078
T-value MAE	23.44	25.18	24.43	28.68	25.10	25.82
σ prop	0.0067	0.0074	0.0072	0.0072	0.0071	0.0068
σ lin	0.0057	0.0049	0.0053	0.0044	0.0053	0.0059
MAE prop mun	0.0278	0.0272	0.0270	0.0267	0.0260	0.0263
MAE lin mun	0.0152	0.0143	0.0138	0.0124	0.0124	0.0117

4 observed municipalities						
Nr. of Elections	14	11	8	5	3	1
MAE prop	0.01101	0.01059	0.00979	0.00930	0.00916	**0.00897**
MAE lin	0.00559	0.00510	0.00498	0.00431	0.00387	**0.00369**
T-value MAE	25.41	26.87	26.71	30.38	31.20	32.378
σ prop	0.00441	0.00460	0.00388	0.00383	0.00405	0.00368
σ lin	0.00303	0.00239	0.00249	0.00192	0.00162	0.00150
MAE prop mun	0.02029	0.01897	0.01827	0.01716	0.01691	0.01664
MAE lin mun	0.01046	0.00987	0.00959	0.00868	0.00802	0.00732

10 observed municipalities						
Nr. of Elections	14	11	8	5	3	1
MAE prop	0.00480	0.00424	0.00408	0.00380	0.00369	**0.00353**
MAE lin	0.00256	0.00247	0.00246	0.00236	**0.00230**	0.00261
T-value MAE	25.73	24.13	22.31	21.70	19.61	11.77
σ prop	0.00201	0.00172	0.00170	0.00152	0.00163	0.00156
σ lin	0.00097	0.00099	0.00091	0.00086	0.00074	0.00108
MAE prop mun	0.01152	0.01055	0.00997	0.00987	0.00931	0.00867
MAE lin mun	0.00676	0.00670	0.00668	0.00658	0.00641	0.00657

Table 2: Comparing the proposed model and the linear model with 2, 4, and 10 observed municipalities and 50 topics. Where *prop* indicates the proposed model, *lin* the linear model, *mun* the results on a municipal level, and *MAE* indicates the mean (over 500 runs) of the mean absolute error between the forecast results and the actual results per party.

Table 2 shows the results for 2, 4, and 10 observed municipalities. With 2 (0.5%) observed municipalities the linear model outperforms the proposed model for all numbers of previous elections with $p < 0.001$ ($T > 3.4$, $df = 499$). The proposed model has a mean absolute error of around 0.017 for all numbers of previous elections, with a slight improvement when less elections are used to generate the distributions over voting profiles per municipality. The mean absolute

error of 0.017 translates to a mean error of 1.7% for each party, which translates to circa 2-3 seats. The mean absolute error of the linear model varies between 0.007 and 0.010. The linear model seems to slightly improve with less elections as well, 1 election the performance decreases. On the municipal level we see a similar pattern to the national level.

With 4 (1%) observed municipalities the linear model outperforms the proposed model for all numbers of previous elections with $p < 0.001$ ($T > 3.4$, $df = 499$). The proposed model has a slightly improving mean absolute error from 0.011 to 0.009. This pattern is similar to the pattern found with 2 observed municipalities and indicates that using a lot of previous elections causes underfitting. The linear model shows a similar improving mean absolute error, with a decreasing number of previous elections used.

With 10 (2.5%) observed municipalities the linear model outperforms the proposed model for all numbers of previous elections with $p < 0.001$ ($T > 3.4$, $df = 499$). The proposed model again has an improving mean absolute error from 0.005 to 0.003, with a decreasing number of previous elections used. The linear model has a mean absolute error between 0.002 and 0.003 for all numbers of previous elections. The linear model has a similar improving mean absolute error as before, but with 1 election the performance decreases.

Table 3 shows the results for 42 and 104 observed municipalities. With 42 (10%) observed municipalities the linear model outperforms the proposed model for 14, and 11 known elections with $p < 0.001$ ($T > 3.4$, $df = 499$). The proposed model outperforms the linear model for 5, 3, and 1 known elections with $p < 0.001$ ($T < -3.4$, $df = 499$). For 8 known elections the proposed model does perform better, however, this is not significant with $p > 0.05$ ($T = -1.20$, $df = 499$). The proposed model again has an improving mean absolute error, from 0.0016 to 0.0011. The linear model's mean absolute error decreases with a lower number of previous elections from 0.00135 to 0.00208.

With 104 (25%) observed municipalities the linear model outperforms the proposed model for all numbers of previous elections with $p < 0.001$ ($T < -3.4$, $df = 499$). The proposed model again has an improving mean absolute error, from 0.0009 to 0.0006. a mean absolute error of 0.0006 translates to an error of 0.06% or $\frac{1}{11}th$ of a seat for each party. The linear model's mean aboslute error decreases from 0.00097 to 0.00195.

42 observed municipalities						
Nr. of Elections	14	11	8	5	3	1
MAE prop	0.00157	0.00148	0.00132	0.00122	0.00116	**0.00111**
MAE lin	0.00135	**0.00134**	0.00135	0.00137	0.00137	0.00208
T-value MAE	7.91	4.99	-1.20	-6.61	-9.38	-16.11
σ prop	0.00060	0.00057	0.00049	0.00046	0.00039	0.00041
σ lin	0.00044	0.00044	0.00043	0.00041	0.00043	0.00126
MAE prop mun	0.00682	0.00631	0.00595	0.00541	0.00509	0.00476
MAE lin mun	0.00593	0.00593	0.00593	0.00592	0.00592	0.00653

104 observed municipalities						
Nr. of Elections	14	11	8	5	3	1
MAE prop	0.00086	0.00086	0.00079	0.00070	0.00067	**0.00064**
MAE lin	**0.00097**	0.00101	0.00100	0.00101	0.00107	0.00195
T-value MAE	-5.98	-8.29	-12.19	-18.19	-24.46	-24.31
σ prop	0.00034	0.00033	0.00026	0.00023	0.00022	0.00021
σ lin	0.00026	0.00029	0.00028	0.00030	0.00029	0.00112
MAE prop mun	0.00578	0.00577	0.00532	0.00495	0.00471	0.00430
MAE lin mun	0.00589	0.00590	0.00590	0.00592	0.00595	0.00676

Table 3: Comparing the proposed model and the linear model with 42 (10%), and 104 (25%) observed municipalities and 50 topics. Where *prop* indicates the proposed model, *lin* the linear model, *mun* the results on a municipal level, and *MAE* indicates the mean (over 500 runs) of the mean absolute error between the forecast results and the actual results per party.

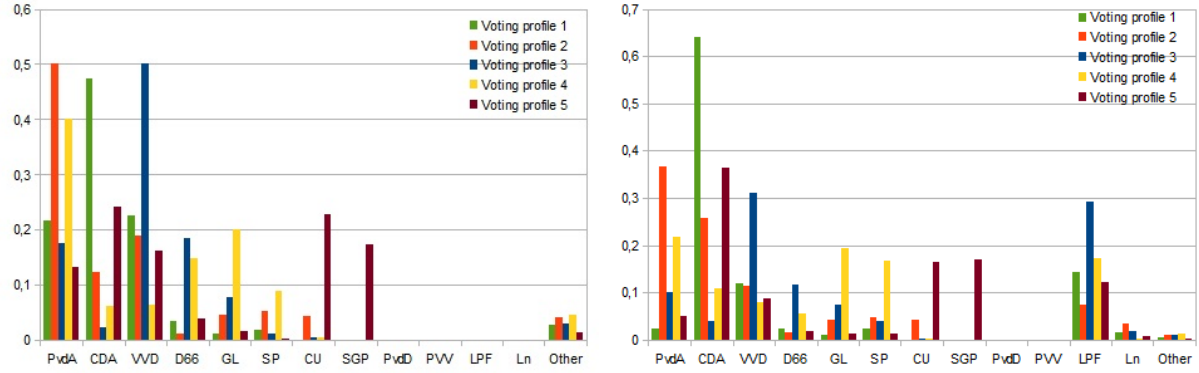
At all stages of an election night the proposed model has an optimal performance when only one previous election is used to compute the distribution over voting profiles per municipality. For the linear model this only holds with a small number of observed municipalities. Both models seem to be very sensitive to under- and overfitting the distribution over voting profiles of the municipalities, when too much previous elections are used. To find the optimal model at each stage of the election night we compare the optimal mean average errors of the model at each stage. With 0.5%, 1%, and 2.5% of the municipalities observed the linear model has a better forecasting performance. With 10% and 25% of the municipalities observed the proposed model has a better forecasting performance. However, the mean average errors of both models are very close, always within a factor 1.4 of each other. For other elections as the 2012 election similar patterns are observed.

3.3 Visualizing voting behavior

The proposed model can be used to visualize voting behavior. The model forecasts elections using latent topic distributions. The first distribution, is the distribution over parties per topic which is visualized for the elections from 1998 to 2012 in figure 2. The second distribution, is the distribution over topics per municipality. In figure 3 a selection of 10 municipalities, and the national result are visualized.

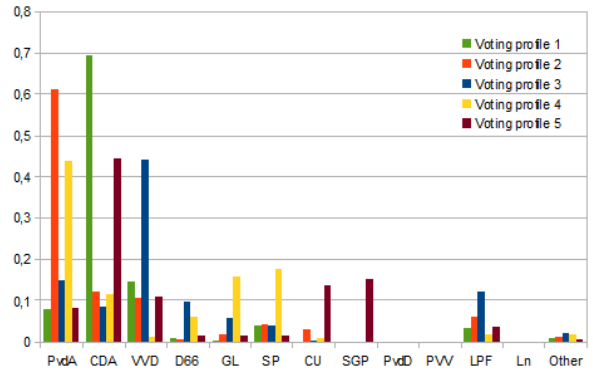
The 5 topics in figure 2 are quite constant over elections, the same parties seem to be a meaningful factor in a topic over multiple elections. The first voting profile shows a strong preference for CDA in the earlier elections. In the 2010 and 2012 elections the probabilities of voting VDD, SP and PVV rise. The second voting profile is mainly a PvdA topic. The third

voting profile shows a strong preference for VDD, and in 2002 and 2003 for the LPF as well. The fourth voting profile shows a left wing and D66 preference, with PvdA, GL, SP and D66 as main parties. The fifth voting profile has high probabilities for protestant parties such as CU and SGP. VVD and CDA are also represented in this voting profile. Voting profiles 1 and 3 each account for 25% of the national voters, voting profiles 2 and 4 account for 20%, and voting profile 5 accounts for 10% of the voters. These voting profiles, and the changes within profiles, can be used to find specific voting patterns and analyze voting. An example is the CDA that lost a lot of voters in recent elections, this can be seen in all voting profiles, but, the percentage of voters lost is higher in profiles 1 and 2, compared to profile 5.

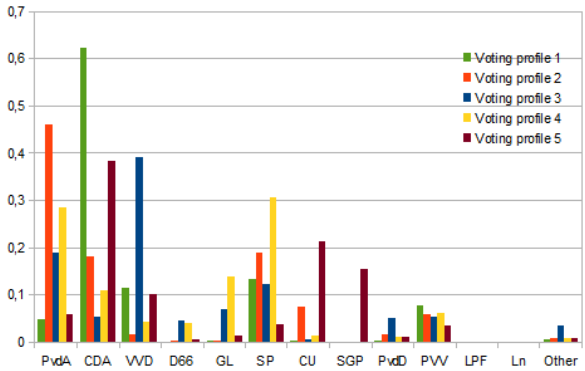


(a) 1998

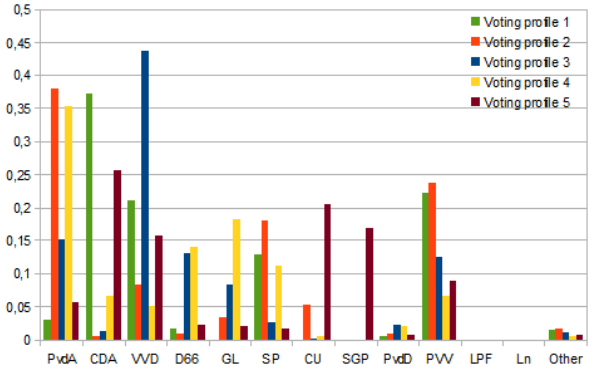
(b) 2002



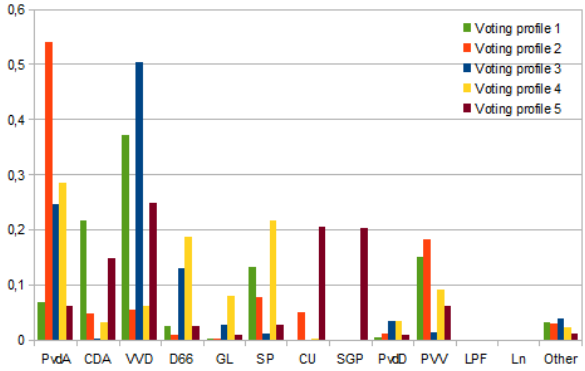
(c) 2003



(d) 2006



(e) 2010



(f) 2012

Figure 2: The distribution over parties per voting profile for the elections from 1998 to 2012. The y-axis indicates the proportion of the given party in the given voting profile.

In figure 3 the distribution over the 5 voting profiles from figure 2 is shown for 10 distinctive municipalities. Large university cities such as Amsterdam, Utrecht and Nijmegen have a high percentage of voters in profile 4, the left wing-D66 profile. Amsterdam and Utrecht also have quite a high percentage in profile 3, the VVD profile. Most municipalities in North-Brabant and Limburg such as Oirschot and Venlo have a high percentage of voters in topic 1, the CDA profile. Municipalities in Groningen, such as Appingedam have a high percentage of voters in topic 2, the PvdA profile. A distinctive municipality is Urk, in Urk there is a 100% probability that voters are part of profile 5. Profile 5, the only profile in which we find the protestant parties, is found mostly in the Bible Belt. Voting profile 5 is very distinctive in the fact that municipalities either have a very low or very high probability for this profile. Voting profile 3 is quite equally distributed over municipalities, although it is a bit higher in the high income municipalities in South-Holland and North-Holland such as Naarden.

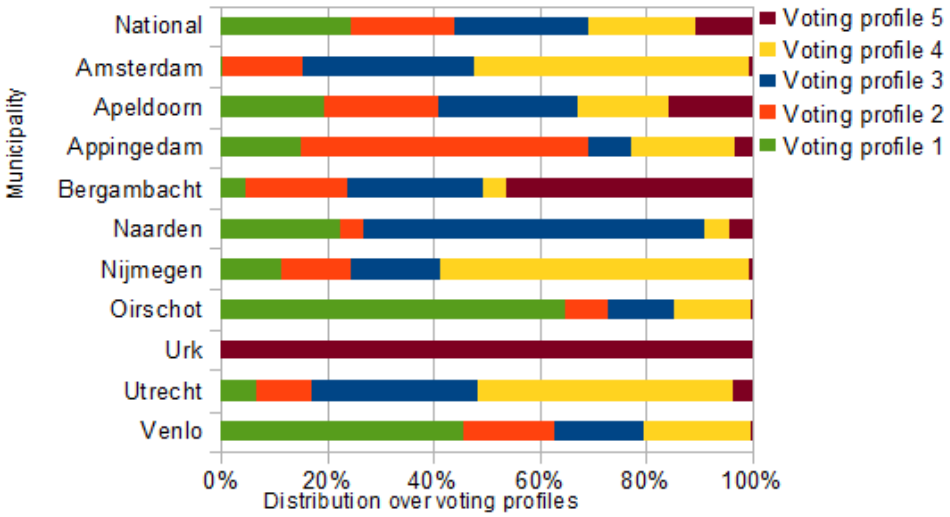


Figure 3: The distribution over the voting profiles in figure 2 for ten municipalities, and the national distribution.

4 Conclusions

The main question was whether the proposed model can give insight into voting behavior, and whether the model can be used to forecast elections. As shown in section 3.1 on the number of topics, there are no signs of overfitting when a high number of topics is used. A high number of previous elections did, however, seem to cause underfitting. Thus, the proposed higher order pLSA model performs best with a low number of previous elections used to forecast elections. The proposed and linear model were both tested on their forecasting performance with different numbers of previous elections known and different numbers of observed municipalities. With 0.5-2.5% of the municipalities observed, the linear model outperformed the proposed model, however, with 10-25% of the municipalities observed the proposed model was optimal. Thus, the proposed model can be used to forecast elections, it is even able to outperform existing models. The proposed model can also be used to give insight into voting behavior. When a small number of topics is used to model the election it is possible to observe voting profiles, similar to those we would expect based on left and right wing distributions, and based on the ideas of parties. It is also possible to observe voter shifts between parties in the visualization. Thus, it is possible to use the model to both gain insight in voting behaviour and forecast election results.

4.1 Future research

Although the forecasting performance of the proposed model is promising it might be possible to extend and improve the model. Firstly, the results show clear signs of underfitting with an increasing number of previous elections. The proposed model performs optimal with only one previous election used to generate the topic distributions for municipalities. It might be possible to improve performance by adding weights for election, thus allowing recent elections to get a higher weight. Secondly, in early stages of election night forecasting exit polls are more reliable than the few results from observed polling stations. The proposed model can also be applied to exit polls, thus allowing for earlier forecasts based on more information. Exit poll results of each polling station can be used in the model as the actual result of that polling station. However, two problems can be expected. Firstly polls are not actual results, voters do not have to, and therefore will not always, state their actual vote. Secondly it is important that the voters asked are a random sample of the voters of the given polling station. Polling at a certain time might cause a bias in the polling result.

Besides improvements, future research can also focus on testing the model in different situations. In the Dutch House of Representatives election results are declared on a municipal level. It would be interesting to see how the model performs when results are declared on a polling station level, which for example happens in Spain [14]. It would also be interesting to test the model on elections where a single member is chosen per district. Because the results are forecast per polling station, only the way in which the results are calculated has to be changed.

5 Acknowledgments

I would like to thank my supervisor, Prof. Tom Heskes, for his advice and valuable suggestions during my thesis project. I would also like to thank my internal supervisor, Dr. Louis Vuurpijl, for his constructive suggestions on this thesis. My grateful thanks are also extended to Prof. Rob Eisinga, who provided me with valuable data and insight into the field of election forecasting. Finally, I wish to thank Dr. Jakob Verbeek, for allowing his matlab implementation of pLSA to be used non commercially.

References

[1] David M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[2] David M. Blei and John D. Lafferty. Correlated topic models. *Advances in neural information processing systems*, 18:147, 2006.

[3] David M. Blei and John D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM, 2006.

[4] Philip Brown and Clive Payne. Election night forecasting. *Journal of the Royal Statistical Society. Series A*, 138(4):463–498, 1975.

[5] Donghui Chen and Robert J. Plemmons. Nonnegativity constraints in numerical analysis. In *Symposium on the Birth of Numerical Analysis*, pages 109–140, 2009.

[6] Rob Eisinga, Manfred te Grotenhuis, and Ben Pelzer. Weather conditions and voter turnout in dutch national parliament elections, 1971-2010. *International Journal of Biometeorology*, 56(4):783–786, 2011.

[7] Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pages 289–296, San Francisco, CA, 1999. Morgan Kaufmann.

[8] Trudy Lisci-Wessels. Ontwikkeling van het aantal gemeenten sinds 1900. Bevolkingstrends, 1e kwartaal 2004, Centraal Bureau voor Statistiek, 2004.

[9] Marco Masseroli, Davide Chicco, and Pietro Pinoli. Probabilistic latent semantic analysis for prediction of gene ontology annotations. In *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, June 2012.

[10] Jon D. Mcauliffe and David M. Blei. Supervised topic models. In *Advances in neural information processing systems*, pages 121–128, 2008.

[11] Andrew McCallum, Andres Corrada-Emmanue, and Xuerui Wang. Topic and role discovery in social networks. *Journal of Articial Intelligence Research*, 30:249–272, 2007.

[12] Warren J. Mitofsky. Review: was 1996 a worse year for polls than 1948? *Public Opinion Quarterly*, 62:230–249, 1998.

[13] Frederick Mosteller, Hyman Herbert, Phillip J. Marks, Eli S. McCarthy, and David B. Truman. *The pre-election polls of 1948: Report to the committee on analysis of pre-election polls and forecasts*. Social Science Research Council, New York, 1949.

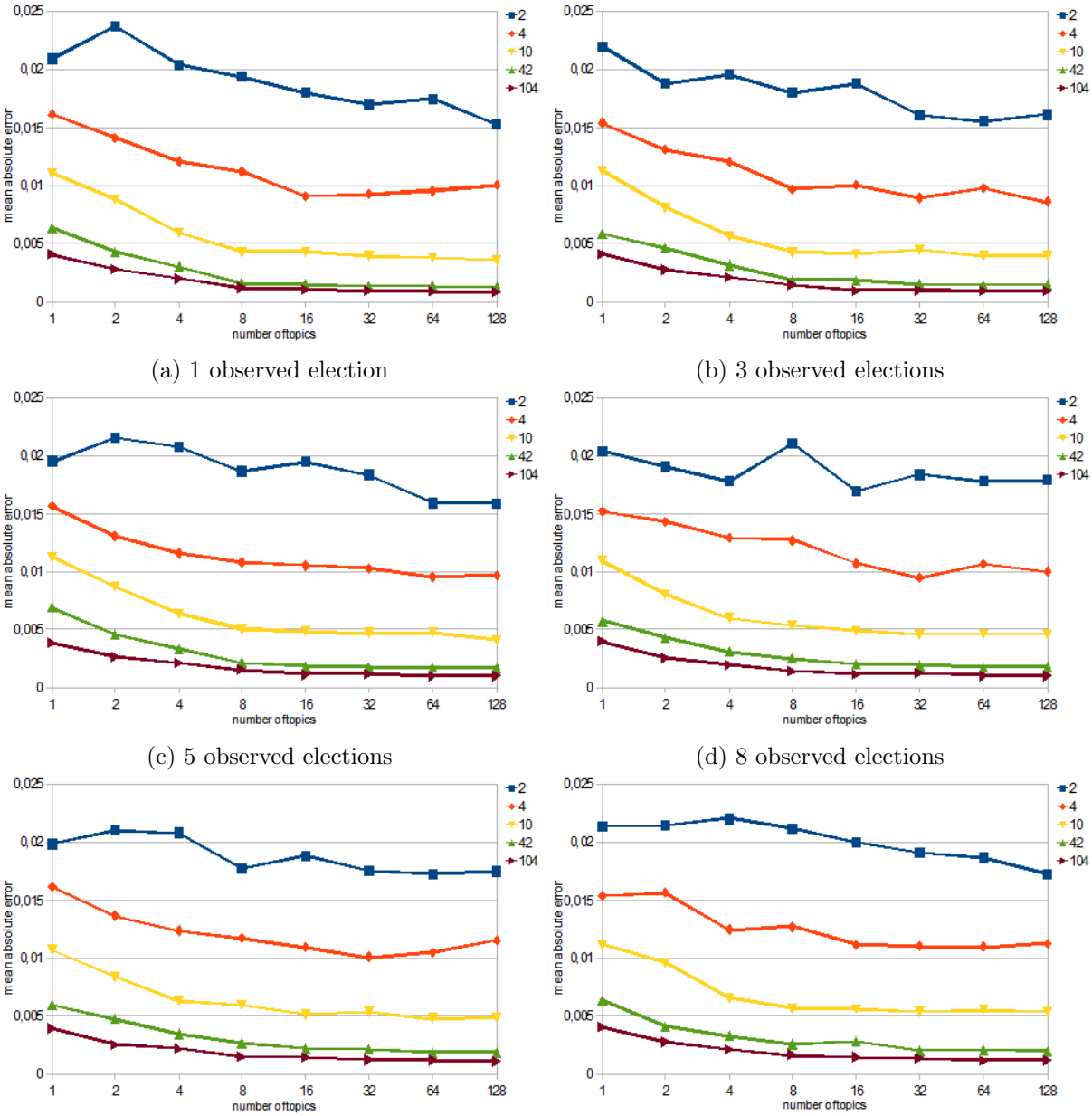
[14] Jose M. Pavía-Miralles. Forecasts from nonrandom samples: The election night case. *Journal of the American Statistical Association*, 100(472):1113–1122, 2005.

[15] Josef Sivic, Bryan C. Russell, Alexei A. Efros, Andrew Zisserman, and William T. Freeman. Discovering object categories in image collections. Technical Report MIT-CSAIL-TR-2005-012, Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, February 2005.

[16] Malcolm J. Wright, David P. Farrar, and Deborah F. Russell. Polling accuracy in a multiparty election. *International Journal of Public Opinion Research*, 26(1):113–124, 2013.

6 Appendix A

The full results of the proposed model for different numbers of topics, previous elections and observed municipalities. The standard deviations of the mean absolute error are between 26



(a) 1 observed election

(b) 3 observed elections

(c) 5 observed elections

(d) 8 observed elections

(e) 11 observed elections

(f) 14 observed elections

Figure 4: The mean absolute error for different numbers of topics and different percentages of observed municipalities