

1 IDEOLOGIE EN CLASSIFICATIE IN DE HANDELINGEN  
2 VAN DE TWEEDE KAMER  
3  
4 SUBMITTED IN PARTIAL FULFILLMENT FOR THE DEGREE OF  
5 BACHELOR OF SCIENCE  
6  
7 JASPER VAN DER HEIDE  
8 10732721  
9  
10 BACHELOR INFORMATIEKUNDE  
11 FACULTEIT DER NATUURWETENSCHAPPEN, WISKUNDE EN  
INFORMATICA  
UNIVERSITEIT VAN AMSTERDAM  
2018-06-28

	<b>Begeleider</b>	<b>Tweede lezer</b>
12 <b>Titel, Naam</b>	Dr Maarten Marx	
<b>Affiliatie</b>	UvA, FNWI, IvI	
<b>Email</b>	maartenmarx@uva.nl .	



14	<b>Inhoudsopgave</b>	
15	<b>1 Introductie</b>	<b>3</b>
16	<b>2 Gerelateerd werk</b>	<b>3</b>
17	2.1 Classificatiemethoden . . . . .	4
18	2.2 Invloed van oppositie of regering . . . . .	5
19	<b>3 Methodologie</b>	<b>5</b>
20	3.1 De data . . . . .	5
21	3.2 Methoden . . . . .	6
22	3.2.1 Deelvraag 1 . . . . .	6
23	3.2.2 Deelvraag 2 . . . . .	8
24	3.2.3 Deelvraag 3 . . . . .	8
25	3.2.4 Deelvraag 4 . . . . .	9
26	<b>4 Evaluatie</b>	<b>9</b>
27	4.1 Resultaten . . . . .	9
28	4.1.1 Deelvraag 1 . . . . .	9
29	4.2 Discussie . . . . .	11
30	4.2.1 Deelvraag 1 . . . . .	11
31	4.2.2 Deelvraag 4 . . . . .	11
32	<b>5 Conclusies</b>	<b>11</b>
33	<b>A Slides</b>	<b>12</b>



## 36 1 Introductie

37 Teksten van politieke partijen kunnen dienen als bron voor het bepalen van  
38 ideologische positie van andere teksten, aangezien zij zowel tekst hebben als  
39 ook een bekende ideologie. Deze informatie kan vervolgens toegepast worden  
40 bij andere teksten die wellicht ideologisch van aard zijn. Bijvoorbeeld, aan de  
41 hand van deze informatie kan men teksten uit kranten classificeren op basis van  
42 ideologie[1, 2].

43 In diverse landen zijn al verschillende onderzoeken gedaan naar het clas-  
44 sificeren van partij-affiliatie op basis van teksten van politici[3, 1]. Mede omdat  
45 elk land een andere politiek stelsel en cultuur heeft, verschillen de resultaten.  
46 Daarnaast gebruikt elk onderzoek ook een andere methode voor het classificeren.  
47 Daarnaast vinden sommige onderzoeken dat deze classificatie minder het gevolg  
48 is van ideologie maar meer van bijvoorbeeld regering tegenover oppositie.[2]

49 Een onderzoek gericht op het Nederlandse parlement ontbreekt hierbij nog.  
50 Ook beperken veel onderzoeken zich vaak tot één classificatiemethode.

51 Dit onderzoek richt zich daarom op een breder scala aan mogelijke me-  
52 thoden en daarnaast specifiek op de Nederlandse politiek. De onderzoeksvraag  
53 luidt daarom dus ook: "In hoeverre is classificatie op basis van partij-affiliatie  
54 aan de hand van spreekbeurten in de Tweede Kamer het gevolg van ideologie?"

55 Deze vraag wordt beantwoord door de antwoorden te vinden op de vol-  
56 gende deelvragen:

- 57 1. Wat is het beste classificatiemodel voor classificatie van partij-affiliatie in  
58 de Tweede Kamer en wat is het resultaat van dit model?
- 59 2. In hoeverre is deze classificatie afhankelijk van partijnamen en namen van  
60 Kamerleden?
- 61 3. In hoeverre wordt deze classificatie bepaald door partij-status (d.w.z. op-  
62 positie of regering)?
- 63 4. In hoeverre is deze classificatie afhankelijk van of een partij rechts of links  
64 is?

65 Daarom zal eerst gekeken worden naar classificatiemethoden en resultaten in  
66 vergelijkbare onderzoeken. Van deze classificatiemethoden zullen een aantal  
67 toegepast worden op teksten van de Tweede Kamer. Vervolgens zal door middel  
68 van de overige deelvragen bepaald worden in hoeverre dit een reflectie is van  
69 ideologie.

70 **Overzicht van scriptie** In sectie 2 zal gerelateerd werk besproken worden,  
71 met name vergelijkbare onderzoeken in andere landen. In sectie 3 zal de me-  
72 thodologie van de verschillende deelvragen behandeld worden. In sectie 4 zul-  
73 len vervolgens de resultaten weergegeven worden. In sectie 5 zal een evaluatie  
74 plaatsvinden van zowel de resultaten als de gehanteerde methodologie. In sectie  
75 6 wordt ten slotte de onderzoeksvraag beantwoord.

## 76 2 Gerelateerd werk

77 Diermeier et al. deden onderzoek naar het classificeren op basis van ideologi-  
78 sche positie in de Amerikaanse Senaat[4]. Ze trainden hun classificatie op de

speeches van de 25 meest liberale en de 25 meest conservatieve senatoren van het 101e tot en met het 107e congres en testten op de 25 meest liberale en de 25 meest conservatieve senatoren van het 108e congres. Later in het onderzoek vergeleken ze ook de 25 gematigd conservatieve en de 25 gematigd liberale senatoren. Dit onderzoek wist de ideologie van de senatoren te voorspellen met een 94 procent nauwkeurigheid voor de classificatie van de extremen, maar slechts een 52 procent nauwkeurigheid voor de classificatie van de gematigde senatoren.

Als een vervolg op dit onderzoek deden Graeme Hirst et al. een vergelijkbaar onderzoek naar zowel het Canadese Parlement [2]. Hierbij werd zowel gekeken naar de Engelse als Franse teksten. Een document werd hier gezien als de samenvoeging van alle spreekbeurten van een spreker. Afhankelijk van taal en dataset, vinden zij in dit onderzoek nauwkeurigheden van 83.2 procent en hoger. In tegenstelling tot het onderzoek van Diermeier et al., vinden zij minder dat de woorden van de sprekers een uiting zijn van ideologie.

Het onderzoek van Bhand et al. richtte zich op het classificeren van leden van het Amerikaanse congres in 2005, op basis van affiliatie (Republikeins of Democratisch)[5]. Zij vonden hiervoor uiteindelijk een  $F_1$  score van 0.684647.

Ferreira probeerde interventies van politici te classificeren op basis van geslacht, leeftijdsgroep, partij-affiliatie en oriëntatie in het Portugese parlement [3]. In alle classificaties kon men een  $F_1$  score van 0.87 of hoger bereiken.

In het onderzoek van Høyland et al. werd een classificatiemodel voor partij-affiliatie op basis van teksten getraind op het vijfde Europese Parlement (1999-2004) en getest op het zesde Europese Parlement[6]. Hier verkregen zij een *macro average*  $F_1$  score van 0.464.

## 2.1 Classificatiemethoden

In het onderzoek van Diermeier et al. werd gebruik gemaakt van support vector machines. Verder maakten ze gebruik van *tf-idf* met een minimale woordfrequentie van 50 en een documentfrequentie van 10, *Part-Of-Speech tagging* en werden alle eigennamen verwijderd.

In het onderzoek van Graeme Hirst et al. maakten ze gebruik van support vector machines[2]. Ze experimenteerden met verschillende vormen van pre-processing, inclusief stemmen en het verwijderen van woorden op basis van te hoge of te lage frequentie. Deze variaties maakten in hun onderzoek geen grote verschillen en uiteindelijk is gekozen voor het niet stemmen, het weglaten van woorden die in minder dan vijf documenten voorkomen en resultaten van zowel met als zonder de top 500 meest frequente woorden. Daarnaast werd geëxperimenteerd met vier wegen voor woorden: *boolean*, *tf*, *tf-norm* en *tf-idf*, waarvan *tf-idf* het beste resultaat opleverde.

In het onderzoek van Bhand et al. gebruikten ze verschillende n-grams, inclusief verschillende manieren van *smoothing*[5]. Zij gebruikte als weging altijd de aanwezigheid van een woord. Als classificatiemodellen experimenteerden ze support vector machines en naive bayes classificatie. Voor het selecteren van *features* experimenteerden ze met een simpele minimale frequentie en het gebruik van een top aantal woorden op basis van mutual information. Uiteindelijk was het beste model bij hen een met support vector machine, met uni- en bigrams, gekozen op basis van mutual information.

In het onderzoek van Ferreira werd gebruik gemaakt van twee classificatiemethoden: Logistische regressie en MIRA[3]. Logistische regressie werd

127 aangevuld met *group Lasso* regularisatie. Voor wegingen van woorden werd  
128 geëxperimenteerd met *tf*, *tf-idf*,  $\Delta$ -*tf-idf* en  $\Delta$ -*BM-25*. Daarnaast wordt er ge-  
129 bruik gemaakt van woordclustering, *Concise Semantic Analysis* en stylome-  
130 trische eigenschappen. Op *Part-Of-Speech tagging* na hadden stylometrische  
131 eigenschappen een duidelijke negatieve invloed op de classificatie.

132 In het onderzoek van Høyland et al. werd gebruik gemaakt van een multi  
133 class support vector machine[6]. Als beste waarde voor de regularisatieterm,  
134 de C-parameter, vonden zij 0.8. Daarnaast gebruikten zij *dependency disambi-*  
135 *guated stems* wat bij hen een  $F_1$  score van twee procent hoger opleverden dan  
136 normale stemming.

## 137 2.2 Invloed van oppositie of regering

138 Graeme Hirst et al. vonden in hun onderzoek dat de classificatie van spreker in  
139 het Canadese parlement op basis van partij-affiliatie meer zegt over de status  
140 van de partij (regering of oppositie).[2] Zo vergeleken zij de top tien karakteris-  
141 tieke woorden van de liberalen en conservatieven in het 36e parlement (liberalen  
142 in de regering) en het 39e parlement (conservatieven in de regering. Hier vonden  
143 zij dat vier van de tien woorden van de liberalen (regering) in het 36e parlement  
144 bij het 39e parlement bij de conservatieven (regering) te vinden waren. An-  
145 dersom gebeurde hetzelfde met één van de tien woorden van de conservatieven  
146 (oppositie) in het 36e parlement naar liberalen (oppositie) in het 39e parlement.

147 In hetzelfde onderzoek trainden ze ook hun classifiers op het ene parle-  
148 ment en testten deze op het andere parlement. Hierbij vonden zij in beide  
149 gevallen een nauwkeurigheid ver onder de baseline. Daarnaast deden ze ook  
150 nog een classificatie op de sprekers die in beide parlementen zaten en een an-  
151 dere classificatie op sprekers die niet in beide parlementen zaten. Bij de eerste  
152 classificatie vonden ze nauwkeurigheden rond de baseline, terwijl in de tweede  
153 situatie nauwkeurigheden gevonden werden ver boven de baseline.

154 Deze resultaten leidden de onderzoekers tot de conclusie dat de classificatie  
155 voornamelijk het gevolg is van de status van de partij en minder van ideologie.

## 156 3 Methodologie

### 157 3.1 De data

158 De data die gebruikt worden, zijn de Handelingen van de Tweede Kamer gedu-  
159 rende het missionaire kabinet-Rutte II (5 november 2012 tot 22 maart 2017).  
160 Er is gekozen voor dit kabinet, omdat de data hiervoor makkelijk verkrijgbaar  
161 was, het kabinet lang zat, waardoor er veel data is, en het recent is waardoor  
162 het makkelijker te interpreteren is. Deze data zijn in xml-formaat van de web-  
163 site officiële bekendmakingen.nl gehaald, samen met corresponderende metadata  
164 xml-bestanden. De bestanden van de Handelingen bevatten voornamelijk infor-  
165 matie over spreekbeurten tijdens een debat, waaronder naam van een spreker,  
166 partij-affiliatie, inhoud van de spreekbeurt en het soort spreekbeurt. Deze ge-  
167 gevens zijn samengevoegd tot een tabel en opgeslagen als csv-bestand.

168 Deze dataset bestaat uit een aantal soorten spreekbeurten, zoals speeches,  
169 interrupties en antwoorden. Daarnaast ook door verschillende soorten sprekers,  
170 zoals de voorzitter, Tweede Kamerleden, leden van het kabinet en gastsprekers.

171 Uit deze dataset is gekozen voor de eerste spreekbeurt nadat een spreker achter  
 172 het spreekgestoelte is gaan staan, aangezien deze vaak langer zijn dan de andere  
 173 spreekbeurten en naar verwachting meer zeggen over ideologie. In de oorspron-  
 174 kelijke xml-bestanden hadden deze spreekbeurten het attribuut *nieuw*="ja".  
 175 Daarnaast is alleen gekozen voor sprekers waarvan er een partij-affiliatie ver-  
 176 meld staat, dit is niet het geval voor leden van het kabinet, de voorzitter en  
 177 gastsprekers (met uitzondering van Nederlandse leden van het Europees Parle-  
 178 ment).

179 Deze dataset bevat vervolgens naast de verkozen partijen van de 2012  
 180 Tweede Kamerverkiezingen, ook afsplitsingen van die partijen (tien in totaal) en  
 181 bezoeken van vertegenwoordigingen van Nederlandse partijen uit het Europees  
 182 Parlement (tien in totaal). Omdat van beide categoriën relatief weinig data is  
 183 en er overlap zit met hun oorspronkelijke partij, zijn deze er uit gehaald. Op  
 184 basis van de aantallen is er voor classificatie een baseline nauwkeurigheid van  
 185 0.148 en baseline  $F_1$  score van \*\*\*\*, verkregen door altijd de grootste partij, de  
 186 SP te kiezen.

Tabel 1: Aantal spreekbeurten per partij gedurende het missionaire kabinet-  
 Rutte II.

50PLUS	413
CDA	2216
ChristenUnie	1223
D66	2211
GroenLinks	1193
PVV	1880
PvdA	2269
PvdD	480
SGP	770
SP	2573
VVD	2157

## 187 3.2 Methoden

### 188 3.2.1 Deelvraag 1

189 Om deze deelvraag te beantwoorden zullen een aantal classificatiemethoden ver-  
 190 geleken worden. Aangezien het onmogelijk is om alle classificatiemethoden te  
 191 vergelijken, beperkt dit onderzoek zich tot classificatiemethoden die gebruikt  
 192 zijn in vergelijkbare onderzoeken, zoals besproken in 2.1. Daarnaast is omwille  
 193 van de tijd ervoor gekozen om alleen gebruik te maken van methoden waar-  
 194 van reeds implementaties beschikbaar waren in Python. Hieronder worden de  
 195 verschillende onderdelen besproken.

196 **Pre-processing** Voor pre-processing is gebruik gemaakt van tokenisation en  
 197 lowercasing. Voor tokenisation is de reguliere expressie  
 198  $w+$  gebruikt, die daarmee alleen de letters van het alfabet overhoudt. Deze  
 199 woorden zijn vervolgens allemaal omgezet in kleine letters. Vervolgens is er  
 200 gevarieerd tussen wel of geen gebruik maken van stemming. In het geval van

stemming is gebruik gemaakt van de Snowball Stemmer via de Python NLTK module.

**Bag-of-words model** Bag-of-words model is de meest gebruikte representatie van data in vergelijkbare onderzoeken. Bij het bag-of-words model wordt elk document gerepresenteerd door een vector, waarbij elke kolom een woord voorstelt met een bijbehorende waarde. Voornaamste beperking van dit model is dat het geen rekening houdt met de volgorde van woorden, wat een groot effect kan hebben op de betekenis van een document.

Voor dit onderzoek zijn de volgende wegingen voor woorden getest: *boolean* (wel of niet aanwezig), *tf* (woordfrequentie), *tf-norm* (woordfrequentie genormaliseerd door documentlengte) en *tf-idf*. Daarnaast wordt in dit onderzoek geëxperimenteerd met een minimale of maximale woord- of documentfrequentie. Ook is gekeken naar het effect van combinaties van unigrams, bigrams en trigrams.

**Support Vector Machines en Logistische Regressie** De meest voorkomende techniek in vergelijkbaar onderzoek is Support Vector Machine (SVM). Een andere techniek die gebruikt wordt is logistische regressie. Beide kennen een eigen implementatie in sklearn, maar gezien de grootte van de dataset, duurt dit te lang met een gridsearch. Om deze reden is er in beide gevallen voor gekozen om gebruik te maken van de functie *SGDClassifier*, die beide technieken leert met *stochastic gradient descent learning*. Er is hiervoor gevarieerd met de regularisatie, learning rate en maximum aantal iteraties. De andere parameters zijn gelaten op de standaardwaarden van scikit-learn[7].

**Naive Bayes** Een simpelere techniek die gebruikt wordt voor politieke tekstclassificatie is Naive Bayes. Dit algoritme neemt aan dat elke *feature* onafhankelijk is ten op zichte van de rest. Dit is bij tekstclassificatie vaak niet het geval omdat het gebruik van sommige woorden gepaard kan gaan met het gebruik van andere woorden. Daarnaast is het gebruik van meerdere n-grams in een classificatie schending van de aanname, want als bijvoorbeeld een bigram er in voorkomt dan komen ook beide unigrams er sowieso in voor. Desalniettemin blijkt Naive Bayes effectief te zijn voor tekstclassificatie[7, 5].

Er zijn twee frequent gebruikte aannames voor de distributies in tekstclassificatie; *Multinomial* en *Bernoulli*. In gerelateerde werken wordt niet gespecificeerd welke gebruikt wordt. Om deze reden zijn ze allebei gebruikt. Hiervoor zijn respectievelijk de functies van scikit-learn *MultinomialNB* en *BernoulliNB* gebruikt.[7, 5]

**Beoordelen van kwaliteit** De meest gebruikte methoden om kwaliteit van politieke tekstclassificatie te beoordelen zijn accuracy en  $F_1$  score, die opgebouwd is uit recall en precision. Deze scores zijn opgebouwd uit het aantal correct positief (*tp*), foutief positief (*fp*), correct negatief (*tn*) en foutief negatief (*fn*) geclassificeerde waarden.

$$Precision = \frac{tp}{tp + fp} \quad (1)$$



$$Recall = \frac{tp}{tp + tn} \quad (2)$$

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \quad (3)$$

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (4)$$

Deze waarden worden per klasse bepaald en daar wordt vervolgens een gemiddelde van genomen, gewogen bij documenten behorende tot die klasse. [8, 7].

Voor de classificatiemethoden wordt waar mogelijk gebruik gemaakt van functies van de Python module scikit-learn[7], aangevuld met zelf geschreven code als dit niet reeds beschikbaar is. Bij al deze classificatiemethoden wordt gevarieerd met meerdere parameters door middel van een gridsearch. Hierbij wordt gebruikt gemaakt van 5-fold cross-validation. Daardoor wordt de data gespleten in vijf delen, waarvan steeds één deel als testset wordt gebruikt en de rest voor training.

### 3.2.2 Deelvraag 2

In het onderzoek van Diermeier et al. worden alle eigennamen weggelaten zodat, volgens hen, namen van personen en partijen niet de classificatie domineren. Aangezien hier bij deelvraag 1 niet voor is gekozen, wordt bij deze deelvraag gekeken hoe groot het effect hiervan is, specifiek gericht op partijnamen en namen van kamerleden. Voor deze deelvraag wordt wederom een classificatie gedaan met de classificatiemethode die resulteerde uit deelvraag 1. In deze classificatie worden alle partijnamen vervangen door de tag PARTIJNAAM en alle namen van Kamerleden vervangen door de KAMERLIDNAAM. Deze resultaten worden vervolgens vergeleken met de resultaten uit deelvraag 1.

### 3.2.3 Deelvraag 3

Om deze deelvraag te beantwoorden zullen de twee experimenten die Graeme Hirst et al. uitvoerden voor dezelfde vraag gereproduceerd worden op de dataset van de Tweede Kamer. Bij deze deelvraag zal de beste classifier uit deelvraag 1 gebruikt worden.

Als vergelijkingsmateriaal is voor deze experiment een tweede dataset nodig uit een ander kabinet. Hiervoor is het wenselijk dat dit kabinet bestaat uit andere partijen dan kabinet-Rutte II. Daarnaast is het ook wenselijk als het niet te ver terug is, zodat onderwerpen en taalgebruik enigszins overeenkomstig zijn. Omdat kabinet-Rutte I een minderheidskabinet was met een bijzondere partij-status voor de PVV, is ervoor gekozen om de Handelingen van de Tweede Kamer tijdens het missionaire kabinet-Balkenende IV (22 februari 2007 tot 20 februari 2010) te gebruiken.

In het eerste experiment zullen de tien meest karakteristieke woorden per partij van het ene parlement vergeleken worden met de tien meest karakteristieke woorden per partij van het andere parlement. Als de classificatie op basis van ideologie is in plaats van partij-status, is het te verwachten dat de woorden bij een partij blijven en niet gekoppeld zijn aan in oppositie of regering zitten.

In het tweede experiment worden classifiers getraind op het ene parlement en getest op het andere parlement. Als de classificatie op basis van ideologie is in plaats van partij-status, is de verwachting dat er nog steeds aanzienlijke voorspellingen gedaan worden, aangezien de ideologie naar verwachting redelijk stabiel is binnen tien jaar (hoewel woordgebruik varieert). Als de scores aanzienlijk lager zijn, kan dit het gevolg zijn van het veranderen van partij-status van partijen.

#### 3.2.4 Deelvraag 4

Voor deze deelvraag vergelijken we de resultaten van de eerdere classificatie per partij met een binaire classificatie op basis van rechts en links. Hiervoor wordt wederom de dataset van kabinet-Rutte 2 gebruikt, met het model wat resulteerde uit deelvraag 1.

Voor deze vraag moet vastgesteld worden welke partijen links en rechts zijn. Omdat dit lastig te bepalen is en er meerdere indelingen zijn, wordt hier gebruik gemaakt van twee verschillende indelingen. De indeling op basis van het Kieskompas van Andre Krouwel voor de Kamerverkiezing 2012 en de indeling volgens het Manifesto Project[9] gebaseerd op verkiezingsprogramma's voor de Kamerverkiezing van 2012. In beide gevallen is de nullijn van het politieke spectrum gebruikt om te bepalen of een partij links of rechts is.

Tabel 2: Rechts (R) of link (L) indeling per partij op basis van het Kieskompas en het Manifesto Project.

Partij	Kieskompas	Manifesto Project
SP	L	L
PvdA	L	L
GroenLinks	L	L
PvdD	L	L
50PLUS	L	L
D66	R	L
PVV	-	R
ChristenUnie	R	R
SGP	R	R
VVD	R	R
CDA	R	R

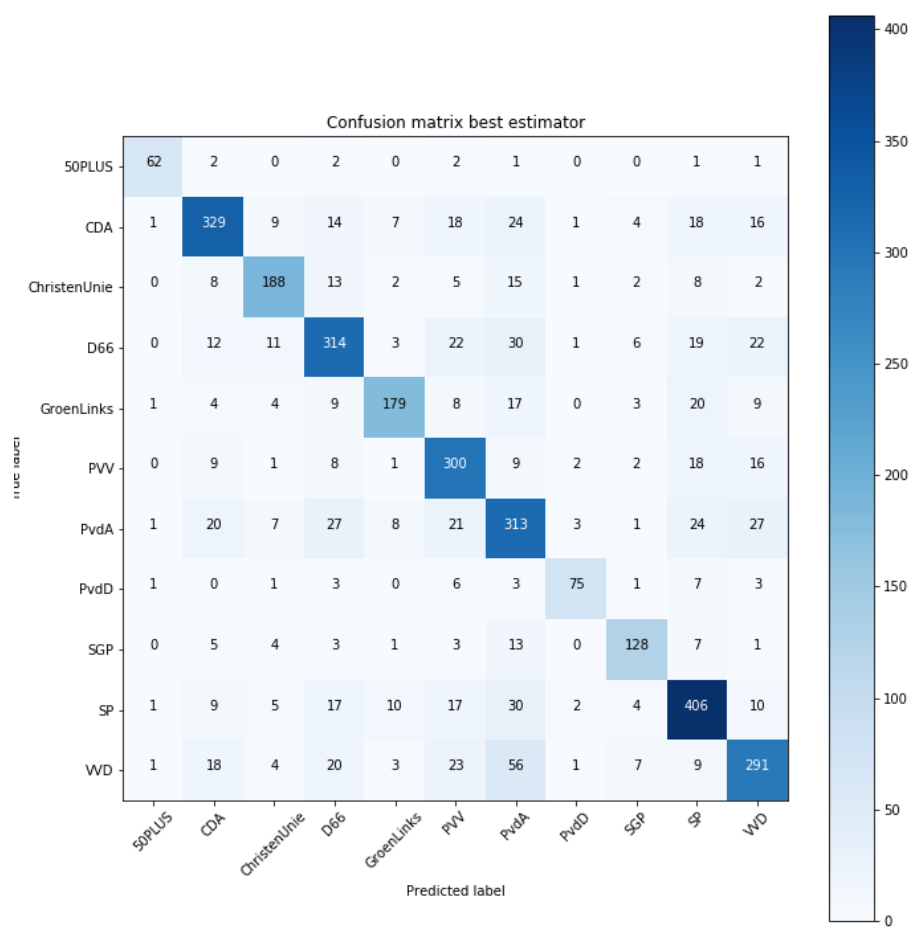
## 4 Evaluatie

### 4.1 Resultaten

#### 4.1.1 Deelvraag 1

Het beste resultaat werd bereikt met SVM gebruikmakend van *stochastic gradient descent learning*

Figuur 1: Confusion matrix van beste classificatie.



## 306 4.2 Discussie

### 307 4.2.1 Deelvraag 1

308 Dit onderzoek heeft zich beperkt tot methoden genoemd in vergelijkbare onder-  
309 zoeken én waarvan de implementatie beschikbaar is in Python. Een aantal me-  
310 thoden die in gerelateerde literatuur leidden tot goede classificaties zijn daarom  
311 niet getest. Ook nieuwe methoden die nog niet gebruikt zijn in een vergelijkbaar  
312 onderzoek voor politieke tekst classificatie zijn daarom niet getest. Omdat dus  
313 niet alle opties getest zijn, kan geen uitsluitel gegeven worden dat dit daad-  
314 werkelijk het classificatiemodel is. Voor vervolgonderzoek kan daarom gekeken  
315 worden naar meer verschillende methoden.

### 316 4.2.2 Deelvraag 4

317 Er zijn verschillende visies op links en rechts, en de indeling van de partijen,  
318 ook buiten de twee methoden gekozen in dit onderzoek.

## 319 5 Conclusies

## 320 Referenties

- 321 [1] Felix Bießmann. Automating political bias prediction. *CoRR*,  
322 abs/1608.02195, 2016.
- 323 [2] Graeme Hirst, Yaroslav Riabinin, Jory Graham, and Magali Boizot-roche.  
324 Text to ideology or text to party status? \*.
- 325 [3] Vasco Ferreira. Using textual transcripts of parliamentary interventions for  
326 profiling portuguese politicians. 2016.
- 327 [4] Daniel Diermeier, Jean-François Godbout, Bei Yu, and Stefan Kaufmann.  
328 Language and ideology in congress. *British Journal of Political Science*,  
329 42(1):31–55, 2012.
- 330 [5] Conal Sathi Maneesh Bhand, Dan Robinson. Text classifiers for political  
331 ideologies, 2009.
- 332 [6] Bjørn Høyland, Jean-François Godbout, Emanuele Lapponi, and Erik Vell-  
333 dal. Predicting party affiliations from european parliament debates. In  
334 *Proceedings of the ACL 2014 Workshop on Language Technologies and*  
335 *Computational Social Science*, pages 56–60. Association for Computatio-  
336 nal Linguistics, 2014.
- 337 [7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Gri-  
338 sel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas,  
339 A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay.  
340 Scikit-learn: Machine learning in Python. *Journal of Machine Learning*  
341 *Research*, 12:2825–2830, 2011.
- 342 [8] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Intro-*  
343 *duction to Information Retrieval*. Cambridge University Press, New York,  
344 NY, USA, 2008.

- 345 [9] Andrea Volkens, Pola Lehmann, Theres Matthieß, Nicolas Merz, Sven Re-  
346 gel, and Bernhard Weßels. The manifesto data collection. manifesto project  
347 (mrg/cmp/marpor). version 2017b, 2017.
- 348 [10] Bei Yu, Stefan Kaufmann, and Daniel Diermeier. Classifying party affilia-  
349 tion from political speech. *Journal of Information Technology & Politics*,  
350 5(1):33–48, 2008.

## 351 **A Slides**