

1 METHODEN VOOR HET VOORSPELLEN VAN
2 PARTIJ-AFFILIATIE IN DE TWEEDE KAMER

3 SUBMITTED IN PARTIAL FULFILLMENT FOR THE DEGREE OF
4 BACHELOR OF SCIENCE

5 JASPER VAN DER HEIDE
6 10732721

7 BACHELOR INFORMATIEKUNDE
8 FACULTY OF SCIENCE
9 UNIVERSITY OF AMSTERDAM
10 YOUR DATE OF DEFENCE IN THE FORMAT YYYY-MM-DD

11

	Internal Supervisor	Second Supervisor
Title, Name	Dr Maarten Marx	
Affiliation	UvA, FNWI, IvI	
Email	maartenmarx@uva.nl	.



13	Inhoudsopgave	
14	1 Introductie	3
15	2 Gerelateerd werk	3
16	2.1 Deelvraag 1	3
17	2.2 Deelvraag 3	4
18	3 Methodologie	4
19	3.1 De data	4
20	3.2 Methoden	6
21	3.2.1 Deelvraag 1	6
22	3.2.2 Deelvraag 2	6
23	3.2.3 Deelvraag 3	6
24	4 Evaluatie	7
25	5 Conclusies	8
26	A Slides	9

27

Samenvatting

28

29 1 Introductie

30 Teksten van politieke partijen kunnen bruikbaar zijn voor het bepalen van ide-
31 ologische positie van andere teksten, aangezien zij zowel een tekst leveren als
32 ook een bekende ideologie. Deze informatie kan vervolgens toegepast worden
33 bij andere teksten die wellicht ideologisch van aard zijn. Bijvoorbeeld, op ba-
34 sis van deze informatie kan men teksten uit kranten classificeren op basis van
35 ideologie.[1][2]

36 In diverse landen zijn al verschillende onderzoeken gedaan naar het clas-
37 sificeren van partij-affiliatie op basis van teksten van politici.[3][1] Mede omdat
38 elk land een andere politiek stelsel en cultuur heeft, verschillen de resultaten.
39 Daarnaast gebruikt elk onderzoek ook een andere methode voor het classificeren.

40 Een onderzoek gericht op het Nederlandse parlement ontbreekt hierbij nog.

41 Dit onderzoek richt zich daarom op een breder scala aan mogelijke metho-
42 den en daarnaast ook specifiek op de Nederlandse politiek. De onderzoeksvraag
43 luidt daarom dus ook: "In hoeverre is partij-affiliatie op basis van ideologie te
44 voorspellen aan de hand van spreekbeurten in de Tweede Kamer?"

45 Deze vraag wordt beantwoord door de antwoorden te vinden op de vol-
46 gende deelvragen:

- 47 1. Wat is het beste classificatiemodel voor classificatie van sprekers in de
48 Tweede Kamer en wat is het resultaat van dit model?
- 49 2. In hoeverre hebben partijnamen en namen van Kamerleden invloed op de
50 classificatie?
- 51 3. In hoeverre heeft partij-status invloed op de classificatie?
- 52 4. In hoeverre kan links en rechts onderscheiden worden in de classificatie?

53 **Overzicht van scriptie** In sectie 2 zal gerelateerd werk besproken worden,
54 met name vergelijkbare onderzoeken uit andere landen. In sectie 3 zal de meth-
55 odologie van de verschillende deelvragen behandeld worden. In sectie 4 zul-
56 len vervolgens de resultaten weergegeven worden. In sectie 5 zal een evaluatie
57 plaatsvinden van zowel de resultaten als de gehanteerde methodologie. In sectie
58 6 wordt ten slotte de onderzoeksvraag beantwoord.

59 2 Gerelateerd werk

60 2.1 Deelvraag 1

61 Diermeier et al. deden onderzoek naar het classificeren op basis van ideologische
62 positie in de Amerikaanse Senaat[4]. Ze trainden hun classificatie op de speeches
63 van de 25 meest liberale en de 25 meest conservatieve senatoren van het 101e
64 tot en met het 107e congres en testten op de 25 meest liberale en de 25 meest
65 conservatieve senatoren van het 108e congres. Later in het onderzoek vergeleken
66 ze ook de 25 gematigd conservatieve en de 25 gematigd liberale senatoren. Voor
67 classificatie maakten ze gebruik van support vector machines. Verder maak-
68 ten ze gebruik van TF-IDF met een minimale woordfrequentie van 50 en een
69 documentfrequentie van 10, *Part-Of-Speech tagging* en werden alle eigennamen
70 verwijderd. Dit onderzoek wist de ideologie van de senatoren te voorspellen

71 met een 94 procent nauwkeurigheid voor de classificatie van de extremen, maar
72 slechts een 52 procent nauwkeurigheid voor de classificatie van de gematigde
73 senatoren.

74 Als een vervolg op dit onderzoek deden Graeme Hirst et al. een vergelijk-
75 baar onderzoek naar zowel het Canadese als het Europese Parlement[2]. In dit
76 onderzoek maken zij gebruik van support-vector machines. In tegenstelling tot
77 het onderzoek van Diermeier et al., vinden zij minder dat de woorden van de
78 sprekers een uiting zijn van ideologie. Daarentegen vinden zij wel een grotere
79 invloed van oppositie tegenover regering in de woorden van de sprekers.

80 Ferreira probeerde interventies van politici te classificeren op basis van ge-
81 slacht, leeftijdsgroep, partij-affiliatie en oriëntatie[3]. In dit onderzoek maakt hij
82 gebruik van twee classificatiemethoden, Logistische regressie en MIRA. Logis-
83 tische regressie werd aangevuld met *group Lasso* regularisatie. Voor wegingen
84 van woorden werd gebruikt gemaakt van woordfrequentie, TF-IDF, Δ -TF-IDF,
85 Δ -BM-25. Daarnaast wordt er gebruik gemaakt van woordclustering, *Concise*
86 *Semantic Analysis* en stylometrische eigenschappen. Op *Part-Of-Speech tagging*
87 na hadden stylometrische eigenschappen een duidelijke negatieve invloed op de
88 classificatie. In alle classificaties kon men aan de hand van logistische regressie
89 en *group Lasso* regularisatie een F1-score van 0.87 of hoger bereiken.

90 2.2 Deelvraag 3

91 Graeme Hirst et al. vonden in hun onderzoek dat de classificatie van spreker in
92 het Canadese parlement op basis van partij-affiliatie meer zegt over de status
93 van de partij (regering of oppositie).[2] Zo vergeleken zij de top tien karakteris-
94 tische woorden van de liberalen en conservatieven in het 36e parlement (liberalen
95 in de regering) en het 39e parlement (conservatieven in de regering). Hier vonden
96 zij dat vier van de tien woorden van de liberalen (regering) in het 36e parlement
97 bij het 39e parlement bij de conservatieven (regering) te vinden waren. An-
98 dersom gebeurde hetzelfde met één van de tien woorden van de conservatieven
99 (oppositie) in het 36e parlement naar liberalen (oppositie) in het 39e parlement.

100 In hetzelfde onderzoek traiden ze ook hun classifiers op het ene parle-
101 ment en testten deze op het andere parlement. Hierbij vonden zij in beide
102 gevallen een nauwkeurigheid ver onder de baseline. Daarnaast deden ze ook
103 nog een classificatie op de sprekers die in beide parlementen zaten en een an-
104 dere classificatie op sprekers die niet in beide parlementen zaten. Bij de eerste
105 classificatie vonden ze nauwkeurigheden rond de baseline, terwijl in de tweede
106 situatie nauwkeurigheden gevonden werden ver boven de baseline.

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

109 3 Methodologie

110 3.1 De data

111 De data die gebruikt worden, zijn de Handelingen van de Tweede Kamer gedu-
112 rende het missionaire kabinet Rutte II (5 november 2012 tot 22 maart 2017).
113 Deze data is in xml-formaat van de website officiële bekendmakingen.nl gehaald,
114 samen met corresponderende metadata xml-bestanden. De bestanden van de

115 Handelingen bevatten voornamelijk informatie over spreekbeurten tijdens een
 116 debat, waaronder naam van een spreker, partij-affiliatie, inhoud van de spreek-
 117 beurt en het soort spreekbeurt. Deze gegevens zijn samengevoegd tot een tabel
 118 en opgeslagen als csv-bestand.

119 Deze dataset bevat naast de verkozen partijen van de 2012 Tweede Ka-
 120 merverkiezingen, ook afsplitsingen van die partijen (tien in totaal) en bezoeken
 121 van vertegenwoordigingen van die partijen uit de Eerste Kamer (10 in totaal).
 122 Omdat van beide categoriën relatief weinig data is en er overlap zit met hun
 123 oorspronkelijke partij, zijn deze er uit gehaald.

Tabel 1: Spreekbeurten per partij gedurende 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

124 3.2 Methoden

125 3.2.1 Deelvraag 1

126 Om deze deelvraag te beantwoorden zullen een aantal classificatiemethoden
127 vergeleken worden. Aangezien het onmogelijk om alle classificatiemethoden
128 te vergeleken, beperkt dit onderzoek zich tot classificatiemethoden die goede
129 resultaten hebben opgeleverd in andere onderzoeken, genoemd in 2.1. De clas-
130 sificatiemethoden zullen opgesplitst worden in feature engineering en machine
131 learning algoritmes. De volgende methoden zijn hiervoor gekozen:

132 Feature Engineering

133 1. TF-IDF

134 Machine Learning Algorithms

135 1. Support-Vector Machines

136 2. Logistische Regressie

137 Voor de classificatiemethoden wordt waar mogelijk gebruik gemaakt van
138 functies van de Python module sklearn, aangevuld met zelf geschreven code als
139 dit niet reeds beschikbaar is. Bij al deze classificatiemethoden wordt gevarieerd
140 met meerdere parameters door middel van een gridsearch. Hierbij wordt ge-
141 bruikt gemaakt van 5-fold cross-validation. De uitslagen worden beoordeeld op
142 basis van gewogen f1-scores.

143 3.2.2 Deelvraag 2

144 Voor deze deelvraag wordt wederom een classificatie gedaan met de classifica-
145 tiemethode die resulteerde uit deelvraag 1. In deze classificatie worden alle
146 partijnamen vervangen door de tag PARTIJNAAM en alle namen van Kamerle-
147 den vervangen door de KAMERLIDNAAM. Deze resultaten worden vervolgens
148 vergeleken met de resultaten uit deelvraag 1.

149 3.2.3 Deelvraag 3

150 Om deze deelvraag te beantwoorden zullen de drie experimenten die Graeme
151 Hirst et al. uitvoerden voor dezelfde vraag gereproduceerd worden op de dataset
152 van de Tweede Kamer. Bij deze deelvraag zal de beste classifier uit deelvraag
153 1 gebruikt worden.

154 Als vergelijkingsmateriaal is voor deze experiment een tweede dataset no-
155 dig uit een ander kabinet. Het kabinet moet bestaan uit andere partijen dan
156 kabinet-Rutte II. Er moet voor het derde experiment variatie zijn in de Kamer-
157 leden tussen de twee kabinetten, maar ook voldoende Kamerleden die in beide
158 perioden in de kamer zaten. Daarnaast is het ook wenselijk als het niet te ver te-
159 rug is, zodat onderwerpen en taalgebruik enigszins overeenkomstig zijn. Omdat
160 kabinet-Rutte I een minderheidskabinet was met een bijzondere partij-status
161 voor de PVV, is ervoor gekozen om de Tweede Kamer tijdens het missionaire
162 kabinet-Balkenende IV (22 februari 2007 tot 20 februari 2010) te gebruiken.

163 Er is voor dit parlement gekozen aangezien deze redelijk recent is (dus
164 met redelijk overeenkomstig taalgebruik) en er een andere coalitie is met andere
165 premier. Voor dit experiment was het beter als er een kabinet was zonder de

166 partijen van kabinet-Rutte II, aangezien er dan minder overlap is, maar dit is
167 sinds de Tweede Oorlog niet langer dan een jaar gebeurd.

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

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

180 In het derde experiment zullen twee classificaties vergeleken worden. De
181 eerste op Kamerleden die in beide parlementen zaten en een classificatie op
182 Kamerleden die maar in één van de twee parlementen hebben gezeten.

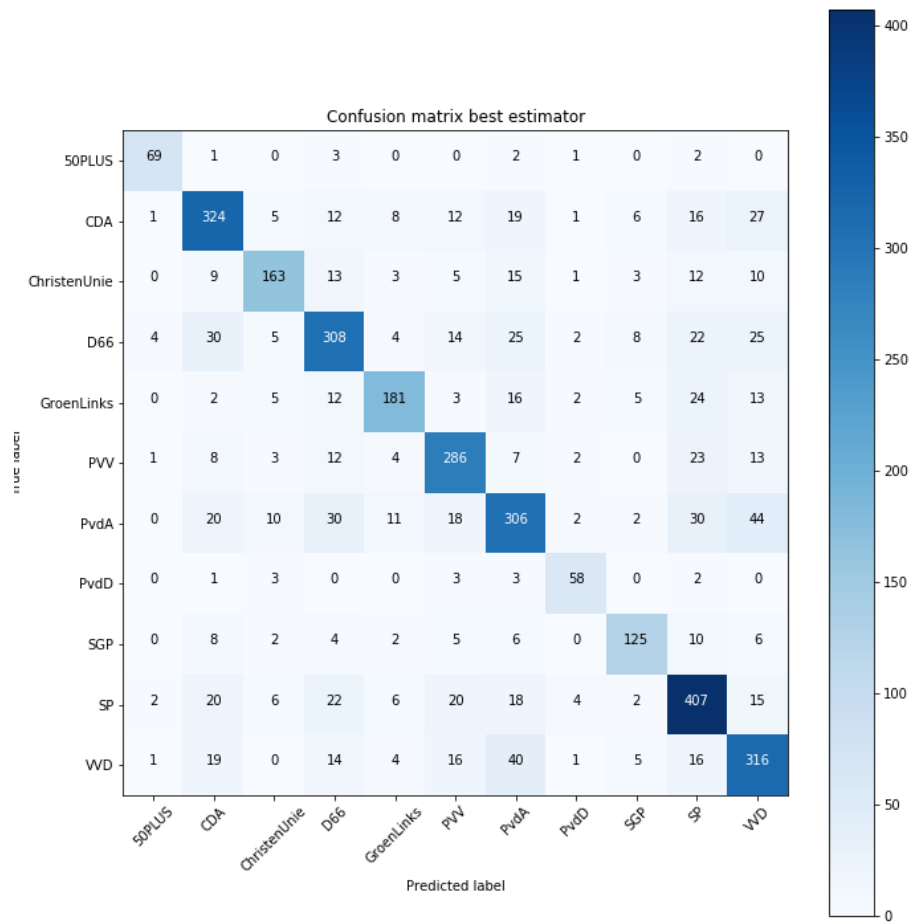
183 4 Evaluatie

184 Met een subsectie voor elke deelvraag.

185 In hoeverre is je vraag beantwoord?

186 Een mooie graphic/visualisatie is hier heel gewenst.

187 Hou het kort maar krachtig.



5 Conclusies

Hierin beantwoord je jouw hoofdvraag op basis van het eerder vergaarde bewijs.

Referenties

- [1] Felix Bießmann. Automating political bias prediction. *CoRR*, abs/1608.02195, 2016.
- [2] Graeme Hirst, Yaroslav Riabinin, Jory Graham, and Magali Boizot-roche. Text to ideology or text to party status? *.
- [3] Vasco Ferreira. Using textual transcripts of parliamentary interventions for profiling portuguese politicians. 2016.
- [4] Daniel Diermeier, Jean-François Godbout, Bei Yu, and Stefan Kaufmann. Language and ideology in congress. *British Journal of Political Science*, 42(1):31–55, 2012.

- 202 [5] Bei Yu, Stefan Kaufmann, and Daniel Diermeier. Classifying party affiliation
203 from political speech. *Journal of Information Technology & Politics*, 5(1):33–
204 48, 2008.

205 **A Slides**