

1 IDEOLOGIE EN CLASSIFICATIE IN DE HANDELINGEN
2 VAN DE TWEEDE KAMER
3 INGEDIEND VOOR GEDEELTELIJKE VOLTOOIING VAN GRAAD VAN
4 BACHELOR OF SCIENCE
5 JASPER VAN DER HEIDE
6 10732721
7 BACHELOR INFORMATIEKUNDE
8 FACULTEIT DER NATUURWETENSCHAPPEN, WISKUNDE EN
9 INFORMATICA
10 UNIVERSITEIT VAN AMSTERDAM
11 2018-06-28

12

	Begeleider	Tweede lezer
Titel, Naam	Dr Maarten Marx	Ir Loek Stolwijk
Affiliatie	UvA, FNWI, IvI	UvA, FNWI, IvI
Email	M.J.Marx@uva.nl	A.M.Stolwijk@uva.nl



UNIVERSITEIT VAN AMSTERDAM

14	Inhoudsopgave	
15	1 Introductie	3
16	2 Gerelateerd werk	4
17	2.1 Tekstclassificatie van parlementaire teksten	4
18	2.2 Classificatiemethoden	5
19	2.3 Invloed van partijnamen of sprekersnamen	6
20	2.4 Invloed van oppositie of regering	6
21	3 Methodologie	7
22	3.1 De data	7
23	3.2 Methoden	9
24	3.2.1 DV1: Beste classificatiemethode	9
25	3.2.2 DV2: Invloed van namen	11
26	3.2.3 DV3: Oppositie of regering	12
27	3.2.4 DV4: Links/rechts	14
28	3.2.5 DV5: Woordgebruik van sprekers	15
29	4 Resultaten	15
30	4.1 DV1: Beste classificatiemethode	15
31	4.2 DV2: Invloed van namen	17
32	4.3 DV3: Oppositie of regering	19
33	4.4 DV4: Links/rechts	22
34	4.5 DV5: Woordgebruik van sprekers	23
35	5 Discussie	24
36	5.1 DV1: Beste classificatiemethode	24
37	5.2 DV2: Invloed van namen	26
38	5.3 DV3: Oppositie of regering	26
39	5.4 DV4: Links/rechts	28
40	5.5 DV5: Woordgebruik van sprekers	28
41	5.6 Algemeen	29
42	6 Conclusies	29

43

Samenvatting

44

1 Introductie

Teksten van politieke partijen kunnen dienen als bron voor het bepalen van ideologische positie van andere teksten, aangezien zij zowel tekst bevatten als ook een bekende ideologie in de vorm van een partij van de spreker; de partij-affiliatie. Het classificeren op basis van tekst kan inzichten geven over de relatie tussen ideologie en woordgebruik. Deze informatie kan vervolgens toegepast worden op andere teksten die wellicht ideologisch van aard zijn. Bijvoorbeeld kan men aan de hand van deze informatie teksten uit kranten classificeren op basis van ideologie [1, 2].

In diverse landen zijn al onderzoeken gedaan naar het classificeren naar partij-affiliatie op basis van teksten van politici [1, 3]. Met deze tekstclassificatie naar partij-affiliatie proberen onderzoekers uit te vinden in hoeverre ideologie terug te vinden is in teksten van politici. De resultaten van de tekstclassificaties zijn in alle gevallen ruim boven de baseline. Diverse onderzoeken wijzen daarentegen ook naar redenen dat dit niet alleen het gevolg is van ideologie. Zo suggereren de resultaten van Hirst et al. [2] dat de partij-status (oppositie tegenover regering) van invloed is op de classificatie. Daarnaast laat dit onderzoek ook zien dat partijnamen een grote invloed hebben op de classificatie.

Een onderzoek gericht op het Nederlandse parlement is niet gevonden. Ook beperken veel onderzoeken zich vaak tot één classificatiemethode.

Dit onderzoek richt zich daarom op meerdere classificatiemethoden en daarnaast specifiek op de Nederlandse politiek. De onderzoeksvraag luidt daarom dus ook: "In hoeverre is classificatie op basis van partij-affiliatie aan de hand van spreekbeurten in de Tweede Kamer het gevolg van ideologie?"

Deze vraag wordt beantwoord door de antwoorden te vinden op de volgende deelvragen:

1. Wat is het beste classificatiemodel voor classificatie naar partij-affiliatie in de Tweede Kamer en wat is het resultaat van dit model?
2. In hoeverre is deze classificatie afhankelijk van achternamen van Kamerleden en partijen?
3. In hoeverre wordt deze classificatie bepaald door partij-status (oppositie of regering)?
4. In hoeverre wordt deze classificatie bepaald door links/rechts positie?
5. In hoeverre wordt deze classificatie bepaald door woordgebruik van sprekers?

Voor de eerste deelvraag zullen Support Vector Machine, Logistische Regressie en Naive Bayes met verschillende parameters vergeleken worden aan de hand van *accuracy* en F_1 score. Bij de tweede deelvraag wordt gekeken naar classificatie zonder achternamen van Kamerleden en partijnamen of met alleen achternamen van Kamerleden en partijnamen. De derde vraag bestaat uit meerdere experimenten, waarin gekeken zal worden naar de hoeveelheid misclassificaties binnen regering of oppositie tegenover tussen regering en oppositie. Daarnaast zal gekeken worden naar overlap in woordgebruik binnen regering en verschil in scores als een partij gewisseld is van partij-status.

89 **Overzicht van scriptie** Sectie 2 bevat vergelijkbare onderzoeken in andere
90 parlementen. Sectie 3 bevat de methodologie van de verschillende deelvragen.
91 Sectie 4 bevat de resultaten. Sectie 5 bevat de evaluatie van zowel de resultaten
92 als de methodologie. Sectie 6 bevat ten slotte het antwoord op de onderzoeks-
93 vraag.

94 2 Gerelateerd werk

95 Toespraken in parlementen worden veel gebruikt in tekstclassificatie, omdat
96 deze veel nette tekst bevatten en vaak gelabeld zijn. Labels zijn bijvoorbeeld
97 naam en partij van de spreker, maar ook daar uit afleidbare labels zoals geslacht,
98 leeftijd en partij-status (oppositie of regering).

99 In dit hoofdstuk zullen verschillende onderzoeken behandeld worden die
100 tekstclassificatie hebben toegepast op parlementaire teksten. Eerst zullen de
101 onderzoeken algemeen besproken worden. Vervolgens zal uitgebreider gekeken
102 worden naar de effecten van verschillende classificatiemethoden. In de latere
103 secties zullen aspecten besproken worden die in vergelijkbare onderzoeken ge-
104 noemd worden als van invloed op de classificatie.

105 2.1 Tekstclassificatie van parlementaire teksten

106 Diermeier et al. deden onderzoek naar het classificeren op basis van ideologische
107 positie in de Amerikaanse Senaat [4]. Ze trainden hun classificatie op de speeches
108 van de 25 meest liberale en de 25 meest conservatieve senatoren van het 101e
109 tot en met het 107e Congres en testten op dezelfde categorieën van het 108e
110 Congres. Een document was in dit onderzoek de verzameling van alle speeches
111 van een senator in een Congres. Deze classificatie resulteerde uiteindelijk in
112 een *accuracy* van 94% (baseline van 50%). Van de 50 senatoren in de test set,
113 kwamen er 44 al voor in de training set, doordat de training op voorgaande
114 Congressen was.

115 Later in het onderzoek vergeleken ze ook de 25 gematigd conservatieve
116 en de 25 gematigd liberale senatoren van dezelfde Congressen. Het resultaat
117 hiervan was 52% (baseline van 50%), dus nauwelijks beter dan de baseline.
118 Als verklaring voor dit verschil ten opzichte van de uitersten zeggen ze dat
119 gematigden een minder duidelijke ideologie hebben.

120 Yu et al. [5] richtte zich vervolgens op zowel het Amerikaanse Huis van
121 Afgevaardigden als de Senaat in 2005. Een document was in dit onderzoek de
122 verzameling van alle speeches van een senator in een Congres en het label de
123 partij. Voor het Huis van Afgevaardigden vonden ze een *accuracy* van 80.1%
124 (baseline van 51.5%) en voor de Senaat 86.0 % (baseline van 55.0%). Ze testten
125 hun classificaties ook op de andere kamer. Van Huis van Afgevaardigden naar
126 Senaat leverde dit een *accuracy* op van 88.0% (baseline van 55.0%) en andersom
127 67.6% (baseline van 51.5%). Hun verklaring voor dit verschil is dat het Huis
128 van Afgevaardigden sterker verdeeld is langs partijlijnen.

129 Vervolgens herhaalden ze de classificaties op het Huis van Afgevaardigden
130 uit 2015, maar testten ditmaal op de Senaat elk jaar tussen 1989 en 2006 afzon-
131 derlijk. Hier zien zij een stijging in *accuracy* van 60% (baseline van 55.0%) in
132 1989 naar 87.0% (baseline van 55.0%) in 2006, maar met twee duidelijke dalen.

133 Ze presenteren twee mogelijke verklaringen voor de trend; het veranderen van
134 de onderwerpen en het sterker verdeeld worden van het Congres.

135 Als een vervolg op deze onderzoeken deden Hirst et al. een vergelijkbaar
136 onderzoek naar het Canadese Parlement [2]. Hierbij werd zowel gekeken naar de
137 Engelse als Franse teksten. Een document werd hier gezien als de samenvoeging
138 van alle spreekbeurten van een spreker. Afhankelijk van taal en dataset vinden
139 zij in dit onderzoek *accuracy* scores van 83.2% en hoger (baseline van 65.5%).

140 Het onderzoek bevat ook een classificatie van het Europees Parlement.
141 Hierbij voegen ze alle teksten van een parlements lid bij elkaar en delen die
142 op in stukken van gelijke grootte. Zij vinden voor documentgrootte van 267
143 woorden een *accuracy* van 44.0% oplopend tot 61.8% (baseline van 38-39%)
144 voor documentgrootte van 6666.

145 Bhand et al. [6] richtten zich op het classificeren van leden van het Ameri-
146 kaanse Congres in 2005, op basis van affiliatie (Republikeins of Democratisch).
147 Een document hierbij was in tegenstelling tot eerdergenoemde onderzoeken een
148 speech. Zij vonden hiervoor uiteindelijk een F_1 score van 0.68 (baseline niet
149 vermeld).

150 Ferreira [3] probeerde interventies van politici te classificeren op basis van
151 geslacht, leeftijdsgroep, partij-affiliatie en oriëntatie in het Portugese parlement
152 . In het geval van classificatie op basis van partij-affiliatie bereikte men een F_1
153 score van 0.90 (baseline niet vermeld, zes partijen).

154 Høyland et al. trainden een classificatie voor partij-affiliatie op basis van
155 teksten van het vijfde Europese Parlement (1999-2004) en testten vervolgens op
156 het zesde Europese Parlement (2004-2009) [7]. Alle teksten van een spreker zijn
157 samengevoegd tot één document. 40% van de sprekers in de test set zaten ook
158 in de training set. Hier verkregen zij een *macro* F_1 score van 0.464 (baseline
159 van 0.097) en *accuracy* van 0.551 (baseline van 0.410). Hun baseline is op basis
160 van altijd classificeren als grootste partij, terwijl voor F_1 score de baseline hoger
161 ligt als hiervoor gekozen wordt voor gokken gewogen bij grootte van een klasse.

162 2.2 Classificatiemethoden

163 Diermeier et al. [4] gebruikten Support Vector Machines. Verder maakten ze
164 gebruik van *tf-idf* met een minimale woordfrequentie van 50 en een minimale
165 documentfrequentie van 10 en *Part-Of-Speech tagging*.

166 Yu et al. [5] maakten gebruik van Support Vector Machines en Naive
167 Bayes, waarvan de varianten multinomial en Bernoulli. De features waren unig-
168 rams, met minimale woordfrequentie van drie en de top 50 meest voorkomende
169 woorden weggelaten. Voor de wegen van de features bij Support Vector Ma-
170 chines werd geëxperimenteerd met *boolean*, *tf-norm* en *tf-idf*. Het beste resultaat
171 was afhankelijk van welke kamer Voor het huis van afgevaardigden was het Sup-
172 port Vector Machines met als weging *tf-idf* en voor de Senaat Bernoulli Naive
173 Bayes.

174 Hirst et al. maakten gebruik van Support Vector Machines [2]. Ze experi-
175 menteerden met verschillende vormen van pre-processing, inclusief stemmen en
176 het verwijderen van woorden op basis van te hoge of te lage frequentie. Deze
177 variaties maakten in hun onderzoek geen grote verschillen en uiteindelijk is ge-
178 kozen voor het niet stemmen, het weglaten van woorden die in minder dan
179 vijf documenten voorkomen en resultaten van zowel met als zonder de top 500
180 meest frequente woorden. Daarnaast werd geëxperimenteerd met vier wegen

181 voor woorden: *boolean*, *tf*, *tf-norm* en *tf-idf*, waarvan *tf-idf* het beste resultaat
182 opleverde.

183 Bhand et al. gebruikten verschillende n-grams, inclusief verschillende ma-
184 nieren van *smoothing*[6]. Ze testten als weging voor features zowel *boolean* als
185 *tf*, waarbij ze vonden dat *boolean* betere resultaten opleverden. Voor classifica-
186 tiemodel experimenteerden ze met SVM en Naive Bayes . Voor het selecteren
187 van *features* experimenteerden ze met een minimale frequentie en selectie van
188 woorden op basis van hoogste mutual information. Uiteindelijk was het beste
189 model bij hen een SVM met uni- en bigrams en geselecteerd op basis van mutual
190 information.

191 Ferreira maakten gebruik van twee classificatiemethoden: Logistische re-
192 gressie en MIRA[3]. Logistische regressie werd aangevuld met *group Lasso* re-
193 gularisatie. Voor wegingen van woorden werd geëxperimenteerd met *tf*, *tf-idf*,
194 Δ -*tf-idf* en Δ -*BM-25*. Daarnaast wordt er gebruik gemaakt van woordcluste-
195 ring, *Concise Semantic Analysis* en stylometrische eigenschappen. Op *Part-Of-*
196 *Speech tagging* na hadden stylometrische eigenschappen een duidelijke negatieve
197 invloed op de classificatie.

198 Høyland et al. maakten gebruik van Support Vector Machine[7]. Als beste
199 waarde voor de regularisatieterm, de C-parameter, vonden zij 0.8. Daarnaast
200 gebruikten zij *dependency disambiguated stems* wat bij hen een F_1 score van
201 twee procent hoger opleverden dan normale stemming.

202 2.3 Invloed van partijnamen of sprekersnamen

203 Diermeier et al. lieten de namen van de sprekers en verwijzingen naar staten
204 die de senatoren representeren weg, omdat deze volgens hen de classificatie te
205 makkelijk zouden maken [4]. Hirst et al. vinden inderdaad dat partijnamen (en
206 het weglaten daarvan) bij het Europees Parlement een grote invloed hebben op
207 de classificatie [2]. Bij het Europees Parlement zien zij met name het gebruik
208 van de eigen partijnaam door een spreker, terwijl zij in het Canadese parlement
209 vooral zien dat de naam van de andere partij gebruikt wordt door een spreker.

210 2.4 Invloed van oppositie of regering

211 Hirst et al. vonden in hun onderzoek dat de classificatie van spreker in het
212 Canadese parlement op basis van partij-affiliatie meer zegt over de status van
213 de partij (regering of oppositie).[2] Zo vergeleken zij de top tien karakteristieke
214 woorden van de liberalen en conservatieven in het 36e parlement (liberalen in
215 de regering) en het 39e parlement (conservatieven in de regering. Hier vonden
216 zij dat vier van de tien woorden van de liberalen (regering) in het 36e parlement
217 bij het 39e parlement bij de conservatieven (regering) te vinden waren. An-
218 dersom gebeurde hetzelfde met één van de tien woorden van de conservatieven
219 (oppositie) in het 36e parlement naar liberalen (oppositie) in het 39e parlement.

220 In hetzelfde onderzoek trainden ze ook hun classificaties op het ene par-
221 lement en testten deze op het andere parlement. Hierbij vonden zij in beide
222 gevallen een *accuracy* ver onder de baseline. Daarnaast deden ze ook nog een
223 classificatie op de sprekers die in beide parlementen zaten en een andere classi-
224 ficatie op sprekers die niet in beide parlementen zaten. Bij de eerste classificatie
225 vonden ze *accuracy* scores rond de baseline, terwijl in de tweede situatie *accuracy*
226 scores gevonden werden ver boven de baseline.

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

229 3 Methodologie

230 3.1 De data

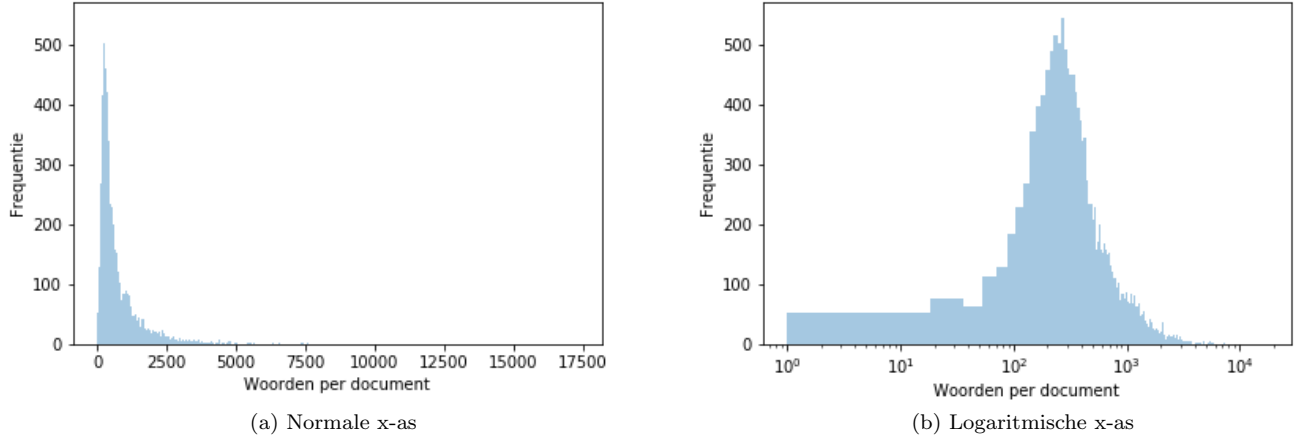
231 De data die gebruikt worden, zijn de Handelingen van de Tweede Kamer gedu-
232 rende het missionaire kabinet-Rutte II (5 november 2012 tot 22 maart 2017).
233 Er is gekozen voor dit kabinet, omdat de data hiervoor makkelijk verkrijgbaar
234 was, het kabinet lang zat - waardoor er veel data is - en het recent is waardoor
235 het makkelijker te interpreteren is. In dit kabinet zaten de PvdA en VVD. Deze
236 data zijn in xml-formaat van de website officielebekendmakingen.nl gehaald sa-
237 men met bijbehorende metadatabestanden. De bestanden van de Handelingen
238 bevatten voornamelijk informatie over spreekbeurten tijdens een debat, waar-
239 onder naam van een spreker, partij-affiliatie, inhoud van de spreekbeurt en het
240 soort spreekbeurt. Deze gegevens zijn samengevoegd tot één tabel.

241 Deze dataset bestaat uit een aantal soorten spreekbeurten; debatbijdra-
242 gen, interrupties en antwoorden. Een debatbijdrage is de eerste onafgebroken
243 spreekbeurt die een spreker geeft achter een spreekgestoelte, aangeduid in de
244 xml-file met het attribuut *nieuw*="ja". Dit kan een bijdrage in een debat zijn
245 of een vraag tijdens een vragen uur. Interrupties zijn de vragen die andere poli-
246 tici stellen vanachter de interruptiemicrofoon aan een spreker. De antwoorden
247 zijn vervolgens de reactie van een spreker achter het spreekgestoelte op een in-
248 terruptie. Aangezien een debatbijdrage geïnterrupteerd kan worden, kan deze
249 inhoudelijk doorlopen in een antwoord van een spreker. Gerelateerde onderzoe-
250 ken voegen vaak alle teksten van een spreker samen tot één document. Dit is
251 alleen niet mogelijk met de hoeveelheid kleine partijen in de Tweede Kamer,
252 die dan niet altijd in een training of test set zijn vertegenwoordigd. Daarom
253 is in dit onderzoek ervoor gekozen om een debatbijdrage met alle bijbehorende
254 antwoorden samen te voegen tot één document voor de classificatie.

255 Daarnaast zijn er verschillende soorten sprekers; de voorzitter, Tweede
256 Kamerleden, leden van het kabinet en gastsprekers. Hieruit is alleen gekozen
257 voor sprekers waarvan er een partij-affiliatie vermeld staat. Dit is niet het geval
258 voor leden van het kabinet, de voorzitter en gastsprekers (met uitzondering van
259 Nederlandse leden van het Europees Parlement).

260 Deze dataset bevat vervolgens naast de verkozen partijen na de Tweede
261 Kamerverkiezingen van 2012 ook afsplitsingen van die partijen (tien in totaal) en
262 bezoeken van vertegenwoordigingen van Nederlandse partijen uit het Europees
263 Parlement (tien in totaal). Omdat van beide categorieën relatief weinig data
264 is en er overlap zit met hun oorspronkelijke of gelieerde partij, zijn deze er
265 uit gehaald. 50PLUS is in 2014 [8] uiteengevallen in twee fracties die aanspraak
266 maakten op de partij-affiliatie 50PLUS. Vanaf dit moment zijn deze documenten
267 niet meer meegenomen om onduidelijkheid te voorkomen.

268 De documenten verschillen in grootte. De distributie van documentgrootte
269 lijkt op een lognormale verdeling, maar met een Kolmogorov-Smirnov test is hier
270 geen bewijs voor gevonden [9].



Figuur 1: Aantal woorden per document

Om toch de uitschieters er uit te halen, is aangenomen dat de distributie wel lognormaal verdeeld is en zijn daarmee de documenten buiten het betrouwbaarheidsinterval van 95% eruit gehaald. De documenten met een lengte van minimaal 28 en maximaal 1492 woorden bleven daarmee over. De gemiddelde documentlengte is daarna 498 woorden en de mediaan is 386 woorden.

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

	Totaal	Vragenuur	Debat
SP	2284	107	2177
CDA	1901	88	1813
D66	1889	133	1756
PvdA	1821	112	1709
PVV	1700	49	1651
VVD	1694	76	1618
ChristenUnie	1068	32	1036
GroenLinks	1068	47	1021
SGP	655	10	645
PvdD	432	14	418
50PLUS	387	12	375
Totaal	14899	680	14219

Deze 14899 documenten zijn verdeeld over 2984 debatten, waarbij elke vraag tijdens het vragenuur als één debat gezien wordt. Op basis van de aantallen is er voor classificatie een baseline *accuracy* van 0.15 (door altijd grootste partij te kiezen) en baseline F_1 score van 0.11 (door te gokken gewogen bij aantal documenten van een partij).

281 3.2 Methoden

282 3.2.1 DV1: Beste classificatiemethode

283 Om deze deelvraag te beantwoorden zullen een aantal classificatiemethoden ver-
284 geleken worden. Aangezien het niet mogelijk is om alle classificatiemethoden
285 te vergelijken, beperkt dit onderzoek zich tot classificatiemethoden die gebruikt
286 zijn in vergelijkbare onderzoeken, zoals besproken in sectie 2.2. Er is ervoor ge-
287 kozen om alleen gebruik te maken van methoden waarvan reeds implementaties
288 beschikbaar waren in scikit-learn. Voor alle methoden wordt gezocht naar de
289 beste parameters, ook wel bekend als een grid search. Deze grid search wordt
290 gedaan door 5-fold cross-validation, waarbij de training set steeds 80% is en de
291 test set 20% van de totale dataset. Een totaal aantal van 6480 combinaties van
292 methoden en parameters zijn getest. De hypothese is dat de scores lager zijn
293 dan die gevonden in het gerelateerd werk, omdat de documentgrootte kleiner is
294 en de baseline scores lager.

295 **Pre-processing** Voor pre-processing is gebruik gemaakt van tokenisation en
296 lowercasing. Voor tokenisation is de reguliere expressie
297 *w+* gebruikt, waardoor alles behalve letters en cijfers weggehaald wordt. Ver-
298 volgens is er gevarieerd tussen wel of geen gebruik maken van stemming. In
299 het geval van stemming is gebruik gemaakt van de Snowball Stemmer van de
300 Python NLTK module.

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

307 Voor dit onderzoek zijn de volgende wegingen voor woorden getest: *boolean*
308 (wel of niet aanwezig), *tf* (woordfrequentie), *tf-norm* (woordfrequentie genor-
309 maliseerd door documentlengte) en *tf-idf* (woordfrequentie gecompenseerd voor
310 documentfrequentie). Daarnaast wordt in dit onderzoek geëxperimenteerd met
311 een minimale of maximale woord- of documentfrequentie. Ook is gekeken naar
312 het effect van combinaties van de volgende n-grams; unigrams, bigrams en tri-
313 grams. N-grams zijn combinaties van N aantal opeenvolgende woorden. Bij
314 een unigram is elke feature gewoon één woord, terwijl bij een bigram dit twee
315 opvolgende woorden zijn. Dit kan nuttig zijn, want als bijvoorbeeld het woord
316 *asfalt* er in voorkomt, dan is het voor ideologie mogelijk relevant of er *minder*
317 *asfalt* of *meer asfalt* staat.

318 **Support Vector Machines en Logistische Regressie** De meest voorko-
319 mende techniek in vergelijkbaar onderzoek is Support Vector Machine (SVM).
320 Een andere techniek die gebruikt wordt is logistische regressie. Beide kennen een
321 eigen implementatie in scikit-learn, maar deze implementaties zijn niet efficiënt
322 met grote datasets. Om deze reden is er in beide gevallen voor gekozen om
323 gebruik te maken van de functie *SGDClassifier*, die beide technieken leert met
324 *stochastic gradient descent learning*. Voor regularisatie is hier geëxperimenteerd
325 met L1 en L2 regularisatie, en een combinatie van beide genaamd Elasticnet.

De andere parameters zijn gelaten op de standaardwaarden van scikit-learn [10]. Een belangrijke onaangepaste waarde is die van maximaal aantal iteraties, waarvoor de scikit-learn standaard 5 is. Volgens scikit-learn convergeert de SGDClassifier rond de $10^6/n$ iteraties waar n het aantal documenten in de training set is. In het geval van deze dataset zou dat 84 iteraties zijn. Vanwege de grootte van de grid search was het voor dit onderzoek niet mogelijk het maximaal aantal iteraties te verhogen tijdens de grid search. De resultaten buiten de grid search zullen gebaseerd zijn op een maximaal aantal iteraties van 100.

Naive Bayes Een andere 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 in voor. Desalniettemin blijkt Naive Bayes effectief te zijn voor tekstclassificatie [6, 10]. Hiervoor zijn de functies van scikit-learn MultinomialNB en BernoulliNB gebruikt [6, 10].

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 worden berekend op basis van vier variabelen. Deze variabelen geven weer hoeveel documenten wel of niet bij een partij horen, en of deze wel of niet als dusdanig zijn geïdentificeerd [11].

	Behorend tot partij	Niet behorend tot partij
Geïdentificeerd als partij	<i>true positive (tp)</i>	<i>false positive (fp)</i>
Niet geïdentificeerd als partij	<i>false negative (fn)</i>	<i>true negative (tn)</i>

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

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

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

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

Accuracy is het percentage van documenten dat correct geïdentificeerd is. *Accuracy* wordt voor de hele classificatie gedaan en niet per klasse. *Precision* is het percentage van documenten geïdentificeerd als een partij, dat ook bij die partij hoort. *Recall* is het percentage documenten van documenten behorende tot een partij, dat ook als die partij geïdentificeerd is. F_1 is het harmonisch gemiddelde van *recall* en *precision*. *Precision*, *recall* en daarmee F_1 worden per partij berekend. Er zijn drie varianten om deze scores voor de hele classificatie te berekenen.

Allereerst is er *micro*, waarbij alle variabelen bij elkaar opgeteld worden en vervolgens de scores berekend. Dit leidt ertoe dat resultaten van partijen met

361 veel documenten belangrijker zijn. Als een classificatie kleine partijen groten-
362 deels fout classificeert, kan deze score alsnog hoog zijn. In het geval van meer
363 dan twee partijen is dit hetzelfde als *accuracy*.

364 Als tweede is er *macro*, waarbij alle scores per partij berekend worden en
365 daarvan het gemiddelde wordt genomen. Dit leidt er dan weer toe dat resultaten
366 van partijen met weinig documenten net zo belangrijk zijn. Hierdoor kan een
367 classificatie met een laag aantal correct geclassificeerde documenten hoog scoren
368 door vooral kleine partijen goed te classificeren.

369 Als laatste is er dan nog *gewogen*, deze berekent net als *macro* de scores per
370 partij, maar neemt hiervan het gemiddelde gewogen bij het aantal documenten
371 behorend tot een partij. Deze wijkt weinig af van de *micro* variant, tenzij er
372 uitschieters zijn bij partijen.

373 Aangezien *micro* al terugkomt in *accuracy* en het nadeel van *macro* te
374 groot is omdat de partijen nogal variëren in grootte, is gekozen voor *gewogen*
375 F_1 score naast *accuracy*.

376 3.2.2 DV2: Invloed van namen

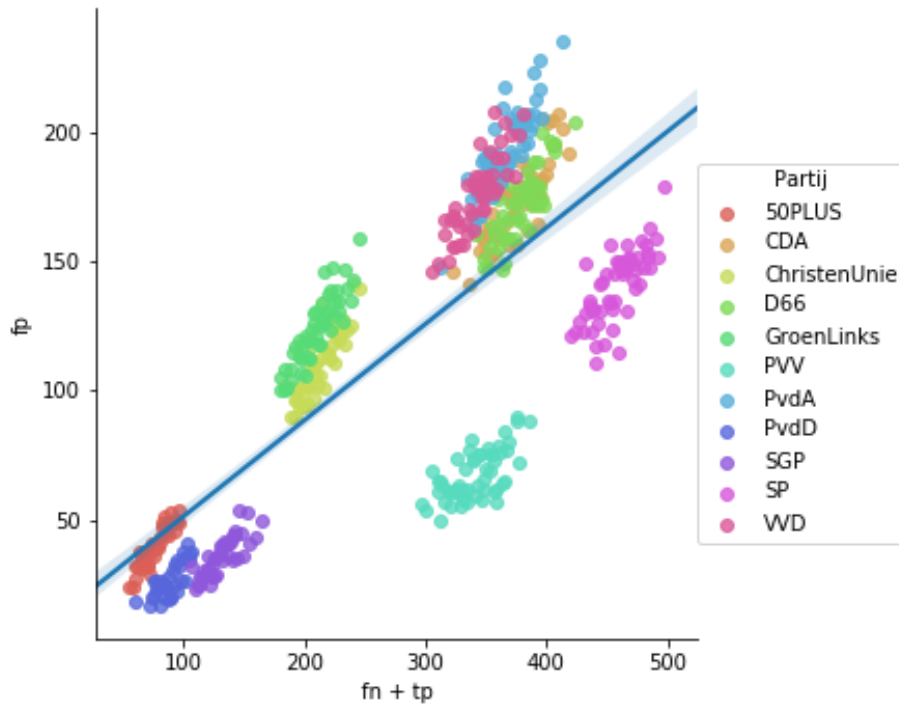
377 In Diermeier et al. [4] wordt aangenomen dat namen een groot effect hebben
378 op de classificatie en Hirst et al. [2] bevestigen dit voor het Europees Par-
379 lement. Aangezien hier bij deelvraag 1 niet voor is gekozen, wordt bij deze
380 deelvraag gekeken hoe groot het effect hiervan is, specifiek gericht op partij-
381 namen en achternamen van Kamerleden. Voor deze deelvraag wordt wederom
382 een classificatie gedaan met de classificatiemethode die resulteerde uit deelvraag
383 1. In deze classificatie worden alle partijnamen vervangen door *PARTIJNAAM*
384 en alle achternamen van Kamerleden vervangen door *KAMERLIDNAAM*. Deze
385 namen zijn uit de Handelingen gehaald. Voor partijnamen zijn ook lidwoorden
386 toegevoegd, voor achternamen van Kamerleden zijn ook verkortingen meegenomen.
387 Dit laatste omdat bijvoorbeeld *Van Haersma Buma* vaak aangesproken
388 wordt als *Buma*. Voornamen van Kamerleden worden zelden tot nooit gebruikt,
389 dus die zijn er niet uitgehaald. Een nadeel van deze aanpak is dat ook namen
390 van niet-Kamerleden of andere woorden weggehaald kunnen worden als deze
391 hetzelfde zijn als naam van een Kamerlid. Door gebruik van gevoeligheid voor
392 hoofdletters is geprobeerd dit te voorkomen. Een opvallend voorbeeld hiervan is
393 de naam Rutte, die zowel behoort tot het Kamerlid Arno Rutte als de premier
394 Mark Rutte. Steekproefgewijs is gekeken of er nog namen achter zijn gebleven,
395 maar die zijn niet gevonden.

396 Ook wordt gekeken naar classificatie met alleen partijnamen en achterna-
397 men van Kamerleden. Alle andere woorden worden weggehaald. Namen van
398 Kamerleden en partijen die niet aan elkaar geschreven worden, zoals *Partij van*
399 *de Arbeid*, worden aan elkaar geschreven zodat het één feature wordt. Doordat
400 alle andere woorden weggehaald zijn, worden de bi- en trigrams combinaties
401 van namen die zinnen uit elkaar kunnen staan, dus die niet meer informatie
402 geven dan unigrams. Daarom wordt er gebruikt van de classificatiemethode uit
403 deelvraag 1, maar dan met alleen unigrams. Hoge scores voor deze classificatie
404 geven aan dat met alleen namen classificatie goed te doen is en dat dit dus een
405 grote bijdrage heeft geleverd aan de resultaten uit deelvraag 1.

3.2.3 DV3: Oppositie of regering

Om deze deelvraag te beantwoorden zal een analyse gedaan worden van de confusion matrix en zullen twee experimenten die gebaseerd zijn op experimenten uit Hirst et al. voor dezelfde vraag uitgevoerd worden op de dataset van de Tweede Kamer. Bij deze deelvraag zal de classificatiemethode uit deelvraag 2 gebruikt worden.

Als er een afhankelijkheid is van partij-status, dan is te verwachten dat het aantal misclassificaties minus verwachte waarde binnen regeringspartijen en binnen oppositiepartijen hoger ligt dan tussen oppositiepartijen en regeringspartijen. De verwachte waarde is afhankelijk van het aantal documenten van een partij in de training set [12]. Aangezien de test set uit dezelfde set als de training is gehaald, is de verwachte waarde ook afhankelijk van het aantal documenten van een partij in de test set. Uit de voorverkenning (op basis van resultaten uit deelvraag 1 en 2) blijkt deze correlatie tussen het aantal *false positives* van een partij en het aantal documenten behorend tot die partij.



Figuur 2: Het aantal *false positives* ten opzichte van het aantal documenten behorend tot die partij (*false negatives* en *true positives*). Dit is op basis van 100 classificaties met verschillende test en train set. De pearson correlatie is 0.77 en de p-waarde 5.40×10^{-101} .

Op basis van dit verband is het verwachte aantal documenten

$$V_{i,j} = fn_i * \frac{tp_j + fn_j}{tn_i + fp_i} \quad (5)$$

422 waar $i \neq j$ met i de echte partij waar een document bijhoort en j de (foutief)
423 voorspelde partij.

424 De error is dan het verschil van de verwachte waarde en het daadwerkelijk
425 aantal documenten

$$e_{i,j} = D_{i,j} - V_{i,j} \quad (6)$$

426 met opnieuw $i \neq j$ en i de voorspelde partij en j de echte partij waar een
427 document bijhoort.

428 Als dit een goede benadering is van de error, dan is het te verwachten
429 dat deze normaal verdeeld is [13]. Om te kijken of er een bias is, worden de
430 distributies binnen regeringspartijen en binnen oppositiepartijen vergeleken met
431 de distributie tussen beide groepen. Om de invloed van variantie door de wil-
432 lekeurige splitsing documenten voor trainen en testen te beperken, wordt de
433 classificatie 100 keer gedaan en worden deze errors bij elkaar in distributie ge-
434 nomen. In het geval dat de distributies normaal verdeeld zijn, zal de statistische
435 test plaatsvinden op basis van een eenzijdige t-toets. Als de distributies niet
436 normaal verdeeld zijn, zal dit plaatsvinden door een Mann-whitneytoets. Het
437 gekozen significantieniveau (α) is 0.05. De nulhypothese is dat er geen verschil
438 is tussen de verdelingen. De alternatieve hypothese is dan dat de distributie
439 van binnen oppositie of regering groter is dan die tussen een regerings- en oppo-
440 sitiepartij. Als de nulhypothese wordt verworpen, kan dus aangenomen worden
441 dat er een verschil is op basis van partij-status.

442 In het eerste experiment uit Hirst et al. zullen de meest karakteristieke
443 woorden per partij van de ene zittingsperiode vergeleken worden met de meest
444 karakteristieke woorden per partij van de andere zittingsperiode. Als de classi-
445 ficatie op basis van ideologie is in plaats van partij-status, is het te verwachten
446 dat de woorden bij een partij blijven en niet gekoppeld zijn aan in oppositie of
447 regering zitten.

448 In het tweede experiment uit Hirst et al. worden classificaties getraind op
449 een zittingsperiode en getest op een andere zittingsperiode. Als de classificatie
450 afhankelijk is van partij-status is de verwachting dat de scores van partijen die
451 gewisseld zijn van oppositie naar regering of andersom lagere scores krijgen dan
452 partijen die niet van partij-status zijn veranderd.

453 Als vergelijkingsmateriaal is voor deze experimenten een tweede dataset
454 nodig uit een ander kabinet. Hiervoor is het wenselijk dat dit kabinet bestaat uit
455 andere partijen dan kabinet-Rutte II. Daarnaast is het ook wenselijk als het niet
456 te ver terug is, zodat onderwerpen en taalgebruik enigszins overeenkomstig zijn.
457 Omdat kabinet-Rutte I een minderheidskabinet was met een bijzondere partij-
458 status voor de PVV, is ervoor gekozen om de Handelingen van de Tweede Kamer
459 tijdens het missionaire kabinet-Balkenende IV (22 februari 2007 tot 20 februari
460 2010) te gebruiken. Dit kabinet bestond uit CDA, PvdA en ChristenUnie.

461 De partij 50PLUS bestond nog niet gedurende kabinet-Balkenende IV,
462 dus documenten van deze partij zijn weggelaten. Verder heeft dezelfde verwer-
463 king van data plaatsgevonden, zoals beschreven in 3.1. Alleen de minimum- en
464 maximumlengte is overgenomen van de dataset van kabinet-Rutte II.

Tabel 2: Aantal documenten per partij gedurende het missionaire kabinet-Balkenende IV.

	Totaal	Vragenuur	Debat
CDA	1039	53	986
ChristenUnie	561	3	558
D66	518	22	496
GroenLinks	760	50	710
PVV	971	56	915
PvdA	903	22	881
PvdD	165	3	162
SGP	507	6	501
SP	1222	70	1152
VVD	1041	55	986
Totaal	7687	340	7347

3.2.4 DV4: Links/rechts

Als de classificatie afhankelijk is van links/rechts positie, dan is het te verwachten dat, net als bij deelvraag 3, de misclassificaties minus de verwachte waarde groter zijn als twee partijen dichterbij elkaar staan op de links/rechts as. Daarvoor zal wederom formule 5 gebruikt worden als verwachte waarde en formule 6 als error.

Er zijn verschillende methoden om partijen in te delen op een links/rechts as. Er is hier gekozen voor de indeling van het Manifesto Project [14]. Het Manifesto Project geeft scores op een heel aantal politieke posities, waaronder dus links/rechts, op basis van het verkiezingsprogramma van dat jaar, in dit geval dus van 2012.

Tabel 3: Links/rechts score per partij van het Manifesto Project voor de verkiezingsprogramma's van 2012.

Partij	Score van Manifesto Project
SP	-20.926
GroenLinks	-9.584
PvdA	-6.558
PvdD	-6.465
50PLUS	-6.311
D66	-0.778
ChristenUnie	10.203
PVV	15.642
CDA	17.701
VVD	22.629
SGP	26.6

476 3.2.5 DV5: Woordgebruik van sprekers

477 De vorige classificaties trainden op documenten en werden getest op andere
478 documenten, maar wel van dezelfde sprekers als uit de training set. Naast
479 de ideologie kan de classificatie daarom ook getraind zijn op het taalgebruik
480 van sprekers. Als een Kamerlid bijvoorbeeld een woord regelmatig in speeches
481 gebruikt, maar niet wordt gebruikt door zijn partijgenoten, wordt dit wel gezien
482 als een belangrijk woord voor de classificatie naar partij-affiliatie. Hirst et al.
483 [2] plaatsten al een soortgelijke kanttekening bij de resultaten van Diermeier et
484 al.

485 Om te kijken of dit effect er is, wordt er opnieuw een classificatie gedaan
486 met de methode uit deelvraag 2. Ditmaal worden alleen niet de individuele
487 documenten verdeeld over de training en test set, maar worden de Kamerleden,
488 met bijbehorende documenten, verdeeld over de training en test set. Als taalge-
489 bruik van een spreker in de training set voorheen invloed had op de classificatie,
490 zal dat nu geen effect meer hebben omdat er geen documenten van die spreker
491 meer voorkomen in de test set. De meest karakteristieke woorden uit de resulta-
492 ten van deelvraag 2 suggereren dat woordgebruik van Kamerleden invloed heeft
493 (zie tabel 5). De hypothese is daarom ook dat deze nieuwe classificatie lagere
494 scores vindt.

495 4 Resultaten

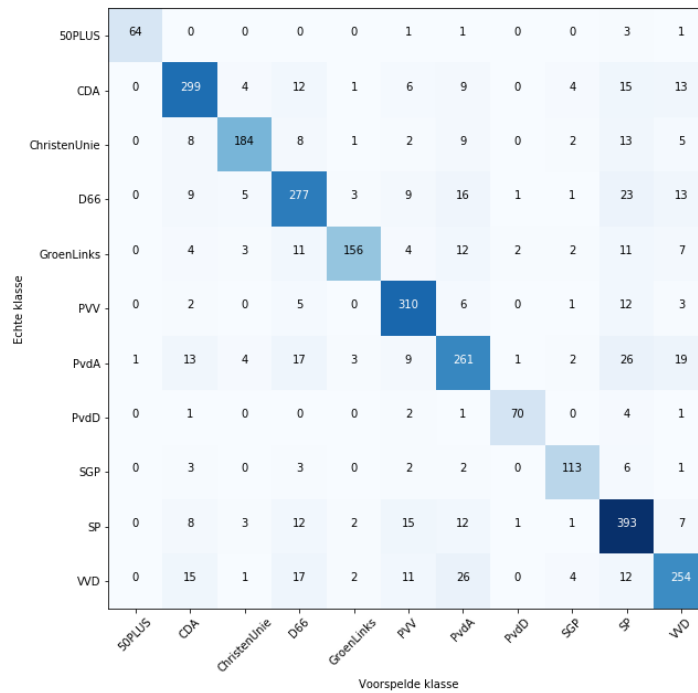
496 4.1 DV1: Beste classificatiemethode

497 Het beste resultaat werd bereikt met Support Vector Machines gebruikmakend
498 van *stochastic gradient descent learning* en Elasticnet regularisatie. De woorden
499 waren hierbij gestemd. De features waren zowel unigrams, bigrams als trigrams.
500 Geen features zijn hierin weggelaten door minimale of maximale documentfre-
501 quenties. Het maximum aantal iteraties was 5 voor de grid search, maar alle
502 resultaten zijn op basis van 100.

503 Tabel 4 laat de scores zien per partij met het aantal documenten in de
504 test set. De *accuracy* voor deze classificatie is 0.80. De F_1 scores per partij
505 liggen tussen de 0.7 en 0.9. De partijen met een sterke focus op één onderwerp,
506 50PLUS, PVV en PvdD, als ook de SGP hebben hoge scores, terwijl de coa-
507 litiepartijen, VVD en PvdA, lagere scores hebben. Figuur 3 laat zien waar de
508 fouten in deze classificatie zitten. De meest karakteristieke features per partij
509 zijn te zien in tabel 5. Met meest karakteristiek worden de n-grams bedoeld die
510 de hoogste coëfficiënt hebben in de classificatie en die dus relatief het meeste
511 belangrijk zijn voor de classificatie van een partij. Hierin is te zien dat vrijwel
512 alle n-grams achternamen van Kamerleden of partijnamen bevatten.

Tabel 4: Classificatie scores per partij van beste classificatiemethode (SVM). Gemiddelde van vijf splitsingen van training en test set. Maximum aantal iteraties is 100.

	Precision	Recall	F_1 score	Documenten
50PLUS	0.97	0.86	0.91	78
PvdD	0.89	0.88	0.89	83
SGP	0.89	0.86	0.88	137
PVV	0.83	0.89	0.86	339
ChristenUnie	0.86	0.78	0.82	220
CDA	0.83	0.80	0.81	376
GroenLinks	0.89	0.73	0.81	203
SP	0.75	0.86	0.80	448
D66	0.76	0.76	0.76	385
VVD	0.75	0.72	0.74	340
PvdA	0.73	0.73	0.73	371
Totaal	0.80	0.80	0.80	2980



Figuur 3: Confusion matrix van beste classificatiemethode (SVM). Gemiddelde van vijf splitsingen van training en test set.

Tabel 5: Meest karakteristieke n-grams per partij op basis van beste classificatie gedurende kabinet-Rutte II. N-grams die niet achternamen van Kamerleden of partijnamen bevatten, zijn dikgedrukt.

50PLUS	CDA	ChristenUnie	D66	GroenLinks
50plus	cda	de christenunie	d66	groenlinks
het lid krol	het cda	christenunie	mijn fractie	lid van tongeren
lid krol naar	cda fractie	lid dik	leden van veldhoven	lid voortman naar
lid krol	de cda fractie	het lid dik	van veldhoven	het lid voortman
krol naar mij	de cda	lid dik faber	veldhoven	lid voortman
krol naar	lid omtzigt naar	dik faber	lid van veldhoven	voortman
krol	lid omtzigt	faber	lid van meenen	tongeren
van 50plus	het lid omtzigt	leden voordewind	d66 wil	van tongeren
gepensioneerden	het cda is	de leden voordewind	d66 is	tongeren naar mij
ouderen	cda is	dik	de leden schouw	van tongeren naar

Tabel 5: Meest karakteristieke n-grams per partij op basis van beste classificatie gedurende kabinet-Rutte II. N-grams die niet achternamen van Kamerleden of partijnamen bevatten, zijn dikgedrukt. (*Vervolg*)

PVV	PvdA	PvdD	SGP	SP	VVD
pvv	de pvda	lid ouwehand	sgp	sp	de vvd
de pvv	pvda	het lid ouwehand	de sgp	de sp	vvd
nederland	de partij van	lid ouwehand naar	sgp fractie	sp fractie	de vvd fractie
islamitische	van de arbeid	ouwehand naar	de sgp fractie	de sp fractie	vvd fractie
miljard	de arbeid	ouwehand naar mij	de leden dijkgraaf	van gerven	de vvd is
het lid graus	partij van de	ouwehand	leden dijkgraaf	gerven	vvd is
lid graus	partij van	dieren	leden van der	lid van gerven	voor de vvd
graus	arbeid	voor de dieren	mevrouw de voorzitter	smaling	wat de vvd
lid graus naar	de partij	de dieren	mevrouw de	leijten naar mij	vvd vindt
klever	pvda fractie	thieme	de leden bisschop	leijten naar	de vvd vindt

513 4.2 DV2: Invloed van namen

514 In tabel 5 was al te zien dat de meest karakteristieke n-grams voornamelijk ach-
515 ternamen van Kamerleden of partijnamen bevatten. In tabel 6 zijn de scores te
516 zien voor een classificatie met alleen achternamen van Kamerleden en partijna-
517 men. De *accuracy* is 0.61. De scores zijn gedaald ten opzichte van de resultaten
518 van deelvraag 1, maar ruim hoger dan de baseline scores.

Tabel 6: Classificatierapport van beste classificatie met alleen achternamen van Kamerleden en partijnamen. Hiervoor is alleen gebruikgemaakt van unigrams. Gemiddelde van vijf splitsingen van training en test set.

	Precision	Recall	F_1 score
50PLUS	0.82	0.88	0.85
PvdD	0.68	0.78	0.69
GroenLinks	0.71	0.66	0.68
PVV	0.66	0.71	0.67
CDA	0.67	0.65	0.66
ChristenUnie	0.66	0.58	0.62
SP	0.61	0.64	0.62
VVD	0.68	0.57	0.62
SGP	0.69	0.54	0.60
D66	0.56	0.53	0.54
PvdA	0.56	0.51	0.52
Totaal	0.64	0.62	0.62

519 In tabel 7 zijn de F_1 scores te zien van classificatie met achternamen van
520 Kamerleden en partijnamen vervangen. De *accuracy* hiervan is 0.58. De scores
521 zijn aanzienlijk lager dan die uit deelvraag 1 en ook nog lager dan van de clas-
522 sificatie met alleen namen. Wel zijn de scores nog ruim hoger dan de baseline.
523 In tabel 8 is vervolgens te zien welke n-grams het meest karakteristiek zijn per
524 partij voor deze classificatie.

Tabel 7: Classificatie scores per partij van beste classificatie zonder achternamen van Kamerleden en partijnamen met het relatieve verschil ten opzichte van tabel 4. Gemiddelde van vijf splitsingen van training en test set.

	Precision	Recall	F_1 score	ΔF_1 score (%)
SGP	0.71	0.73	0.72	-18
PvdD	0.75	0.70	0.72	-19
PVV	0.63	0.80	0.70	-19
ChristenUnie	0.68	0.46	0.55	-21
CDA	0.52	0.53	0.52	-23
SP	0.54	0.71	0.61	-24
D66	0.55	0.55	0.55	-28
VVD	0.54	0.49	0.52	-30
50PLUS	0.86	0.49	0.62	-32
PvdA	0.51	0.48	0.50	-32
GroenLinks	0.64	0.38	0.48	-41
Totaal	0.59	0.58	0.57	-29

Tabel 8: Meest karakteristieke n-grams per partij op basis van classificatie uit deelvraag 1 zonder achternamen van Kamerleden en partijnamen gedurende kabinet-Rutte II.

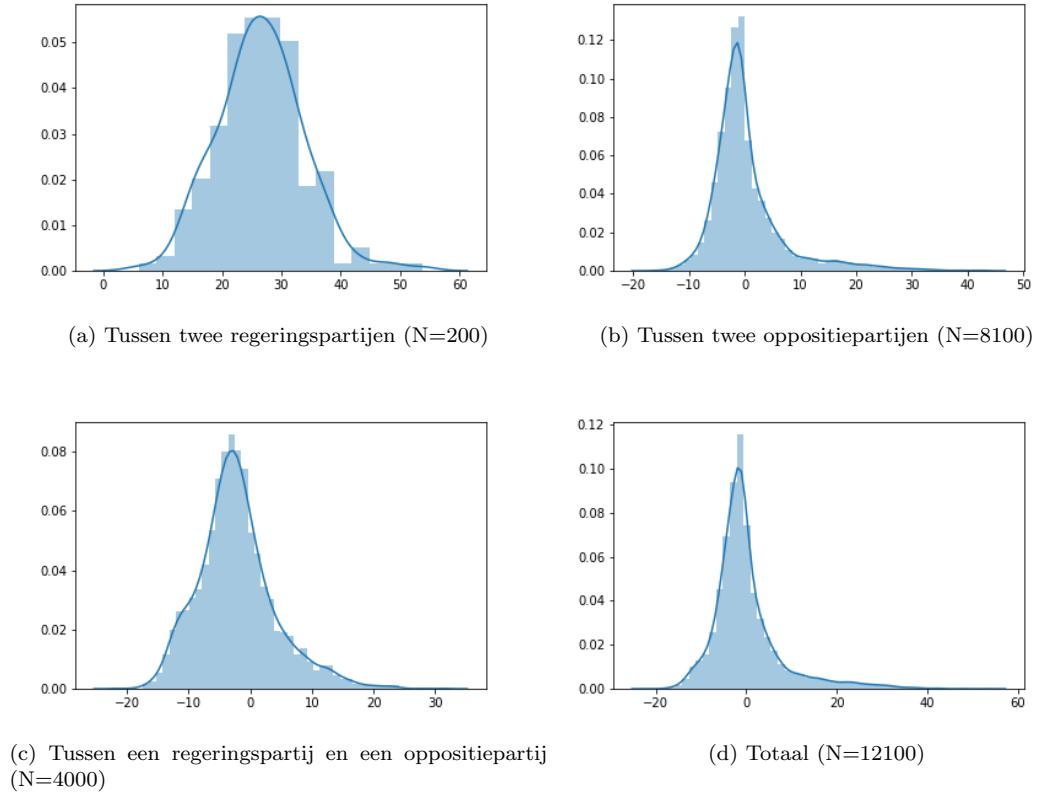
50PLUS	CDA	ChristenUnie	D66	GroenLinks
gepensioneerden	PARTIJ fractie	gezinnen	mijn fractie	zou
ouderen	inwoners	inderdaad	mijn	kamer hierover te
oudere	regering	ik hoop	natuurlijk	schone energie
koopkrachtontwikkeling	PARTIJ	voedselverspilling	fractie	persoonsgebonden
plussers	hier	hoop	buitengewoon	schone
werkenden	fractie	zullen	belangrijk	belastingontwijking
overwegende dat	de regering	mensenhandel	het kabinet	in elk geval
overwegende	echt	ik hoop dat	daarom	hierover te
50 plussers	wij	rechtsstaat	minister	elk geval
voor gepensioneerden	de	horeca	vandaag	hierover te informeren

Tabel 8: Meest relevante n-grams per partij op basis van classificatie uit deelvraag 1 zonder achternamen van Kamerleden en partijnamen gedurende kabinet-Rutte II. (*Vervolg*)

PVV	PvdA	PvdD	SGP	SP	VVD
islamitische	en	dieren	mevrouw de voorzitter	zegt	PARTIJ
nederland	jongeren	milieu	mevrouw de	mening dat	volgens mij
miljard	kinderen	dierenwelzijn	beantwoording	voorstellen	aruba
natuurlijk	lagere overheden	natuur	punt	van mening dat	speelveld
islam	goede	de bio	voor de beantwoording	bezuinigingen	aangegeven
de islam	circulaire	bio	de beantwoording	de bevolking	volgens
al	tevens	industrie	wel	mensen	en
asielzoekers	circulaire economie	bio industrie	allerlei	huurders	PARTIJ fractie
verzorgingshuizen	mijn partij	klimaatverandering	bewindslieden	van mening	essentieel
brussel	regering tevens	de bio industrie	nadrukkelijk	bevolking	regelgeving

525 4.3 DV3: Oppositie of regering

526 In figuur 4 zijn de distributies van de errors, zoals gedefinieerd in formule 6 te
527 zien van combinaties van regerings- en oppositiepartijen.



Figuur 4: Genormaliseerde distributie van de error uit formule 6 voor de verschillende combinaties.

528 Voor alle distributies kan de nulhypothese verworpen worden dat deze
529 normaal verdeeld zijn, hoewel dit wel verwacht was. In tabel 9 is vervolgens te
530 zien dat er een significant verschil is tussen de distributies binnen regering en
531 oppositie tegenover de distributie tussen regering en oppositiepartij.

Tabel 9: Uitslagen van eenzijdige Mann-whitneytoets tussen de distributie tussen een regeringspartij en oppositiepartij en twee distributies. α is 0.05.

	p-waarde	U-waarde
Tussen twee regeringspartijen	7.04×10^{-124}	717042
Tussen twee oppositiepartijen	4.4×10^{-108}	16328471

532 In tabel 10 zijn de meest karakteristieke n-grams te zien voor classificatie
533 van kabinet-Balkenende IV. Hierin zijn geen opvallende overlappen te zien van
534 regeringspartijen met de classificatie van kabinet-Rutte II in tabel 8.

Tabel 10: Meest karakteristieke n-grams per partij op basis van classificatie uit deelvraag 2 gedurende kabinet-Balkenende IV.

CDA	ChristenUnie	D66	GroenLinks	PVV
PARTIJ fractie	fractie van PARTIJ	premier	PARTIJfractie	burgers
wij	de fractie van	de premier	fractie van PARTIJ	onze
fractie	de fractie	ik hoop	de fractie	niet
wij hebben	fractie van	arbeidsmarkt	de fractie van	deze
KAMERLID	mijn fractie	de arbeidsmarkt	fractie van	immigratie
dank	geweest	hoop	politieke	natuurlijk
PARTIJ fractie heeft	moment	hij	premier	politie
zorgvuldige	termijn	schone energie	deal	de burgers
ons	verschillende	plannen	ik	burger
buitengewoon	beantwoording	kunnen	en	door

Tabel 10: Meest karakteristieke n-grams per partij op basis van classificatie uit deelvraag 2 gedurende kabinet-Balkenende IV. (*Vervolg*)

PvdA	PvdD	SGP	SP	VVD
wij	dieren	mijn fractie	zegt	PARTIJ
vrouwen	natuur	wel	mensen	onze fractie
belangrijk	bio industrie	beantwoording	problemen	PARTIJ fractie
kinderen	de bio industrie	de bewindslieden	niet	fractie
of	de bio	bewindslieden	vandaar	ondernemers
punt	bio	de voorzitter	is	je
goed	veehouderij	toch	de mensen	want
onderzoek	dierenwelzijn	mijn	bureaucratie	awbz
roc	industrie	een	verdrag	eens
vragen	de natuur	natuurlijk	waarom	antwoorden

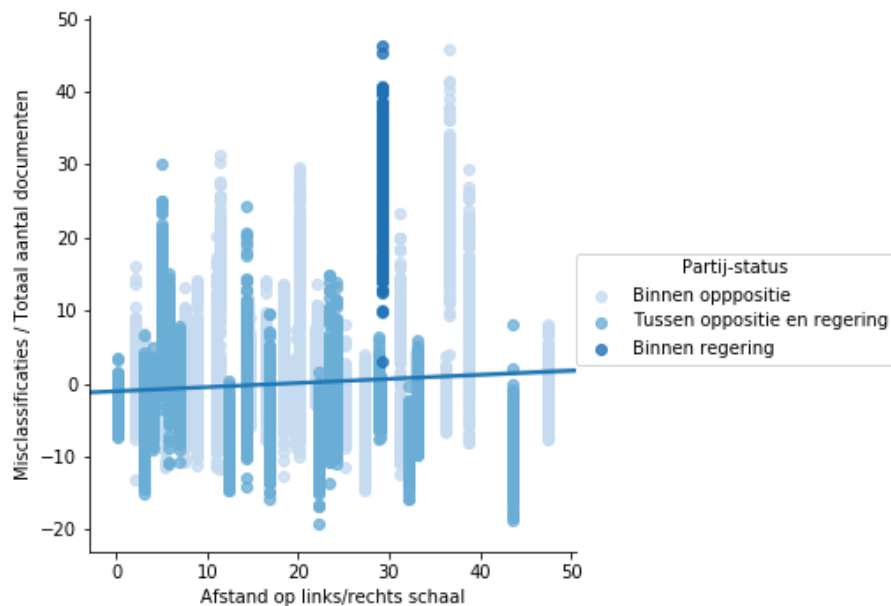
535 In tabel 11 zijn de resultaten van de classificatiescores te zien waarbij de
 536 classificatie getraind is op een zittingsperiode, maar getest op een andere. De
 537 resultaten zijn sterk gedaald, maar nog boven de baseline. De daling verschilt
 538 enorm per partij en zittingsperiode met dalingen van F_1 scores tussen 12 en
 539 92%.

Tabel 11: F_1 scores van de classificatie getraind op ene zittingsperiode en getest op andere zittingsperiode. Scores van een classificatie getraind en getest op kabinet-Rutte II zonder 50PLUS zijn bijgevoegd ter referentie, als ook de relatieve daling. De classificatiemethode uit deelvraag 1 is gebruikt zonder achternamen van Kamerleden en partijnamen. Partijen met een asterisk zijn gewisseld van partij-status.

		Training set → Test set			
Rutte II		Balkenende IV → Rutte II Baseline = 0.11		Rutte II → Balkenende IV Baseline = 0.12	
	F_1	F_1	ΔF_1 score (%)	F_1	ΔF_1 score (%)
SGP	0.74	0.56	-24	0.49	-34
PvdD	0.73	0.64	-12	0.45	-38
PVV	0.70	0.50	-29	0.60	-14
SP	0.61	0.41	-33	0.53	-13
ChristenUnie*	0.55	0.37	-33	0.22	-60
D66	0.54	0.16	-70	0.28	-48
CDA*	0.53	0.28	-47	0.43	-19
PvdA	0.52	0.29	-44	0.27	-48
VVD*	0.51	0.18	-65	0.10	-80
GroenLinks	0.49	0.31	-37	0.04	-92
Totaal	0.58	0.34	-41	0.35	-40

540 4.4 DV4: Links/rechts

541 In tabel 5 is de error te zien ten opzichte van de afstand op de links/rechts
542 schaal.



Figuur 5: Error ten opzichte van de afstand op de links/rechts schaal van twee partijen. Gebaseerd op 100 classificaties met verschillende test en train set. De pearson correlatie is 0.09 en de p-waarde 2.39×10^{-20} .

De correlatie is daarmee significant op het significantieniveau van 0.05, maar wel positief. Uit deelvraag 3 bleek dat de error binnen oppositie of regering significant afweek van de error tussen regering en oppositie. Dit effect lijkt ook zichtbaar in tabel 5. Daarom wordt er in tabel 12 ook gekeken naar de correlatie tussen afstand op de links/rechts schaal en error binnen oppositie en tussen regerings- en oppositiepartij.

Tabel 12: Nog invullen

	Pearson correlatie	p-waarde
Tussen oppositie- en regeringspartij	-0.29	3.44×10^{-69}
Tussen twee oppositiepartijen	0.18	1.76×10^{-55}

4.5 DV5: Woordgebruik van sprekers

In tabel 13 staan de scores van classificatie waarbij de Kamerleden verdeeld zijn over de training en test set. De scores zijn hierbij amper hoger dan de baseline.

Tabel 13: Classificatierapport van beste classificatie met de Kamerleden verdeeld over training en test set. Gemiddelde van vijf splitsingen van training en test set.

	Precision	Recall	F_1 score	ΔF_1 score (%)
50PLUS	0.29	0.06	0.09	
CDA	0.12	0.20	0.14	
ChristenUnie	0.08	0.14	0.09	
D66	0.22	0.22	0.22	
GroenLinks	0.16	0.04	0.05	
PVV	0.29	0.50	0.37	
PvdA	0.25	0.19	0.21	
PvdD	0.46	0.17	0.22	
SGP	0.17	0.05	0.07	
SP	0.34	0.33	0.33	
VVD	0.31	0.26	0.24	
Totaal	0.31	0.24	0.24	

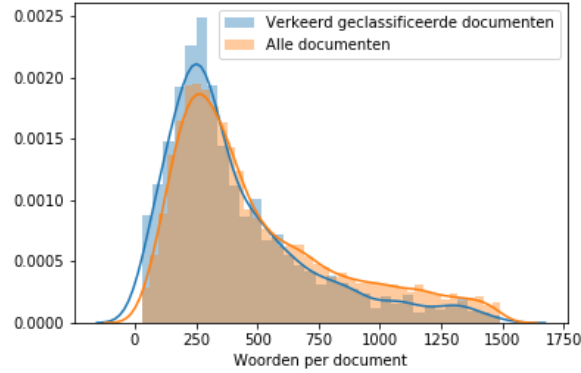
5 Discussie

5.1 DV1: Beste classificatiemethode

Het onderzoek behaalt resultaten in lijn der verwachting op basis van gerelateerd werk en daarnaast ruim boven de baseline scores. De lage scores voor de coalitiepartijen steunen de hypothese van een afhankelijkheid van partij-status, zoals besproken wordt in deelvraag 3. Het bijna alleen voorkomen van namen van partijen en Kamerleden in de meest karakteristieke n-grams per partij in tabel 5 steunt daarnaast het vermoeden dat deze classificatie sterk afhankelijk is van die namen, zoals besproken wordt in deelvraag 2.

Dit onderzoek heeft zich beperkt tot methoden genoemd in vergelijkbare onderzoeken en waarvan de implementatie beschikbaar is in scikit-learn. Een aantal methoden die in gerelateerde literatuur leidden tot goede classificaties zijn daarom niet getest. Ook nieuwe methoden die nog niet gebruikt zijn in een vergelijkbaar onderzoek voor politieke tekst classificatie zijn daarom niet getest. Daarnaast richtte zich dit ook maar op een beperkt aantal parameterwaarden. Voor vervolgonderzoek kan daarom dit onderdeel uitgebreid worden. Het effect van het beperkte maximum iteraties was bij de beste classificatiemethode 2%.

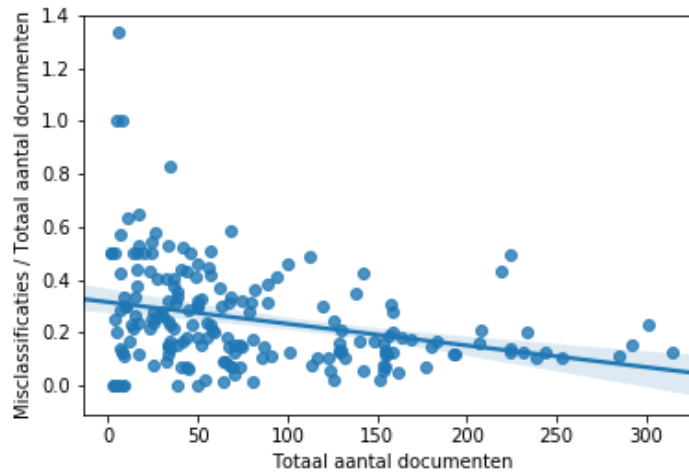
Het onderzoek van Hirst et al. vond dat resultaten afhankelijk kunnen zijn van documentgrootte. Alle documenten in dit onderzoek zijn kleiner dan de grootste documentgrootte uit het onderzoek van Hirst et al. en ook de minimale documentgrootte ligt lager dan de kleinste documentgrootte uit dat onderzoek. Het effect wat zij vinden tussen documentgrootte van 267 en 6666 is een verschil in *accuracy* van 19.8%. Dit onderzoek vindt inderdaad dat kleinere documenten vaker foutief geclassificeerd worden.



Figuur 6: Genormaliseerde distributie van documentlengtes van foutief geclas-
sificeerde documenten en alle documenten. Totaal van 5-fold cross-validation,
waardoor documenten vaker voor kunnen komen. Mediaan documentlengte van
foutief geclassificeerde documenten is 321 en voor alle documenten 386.

576 Voor een vervolgonderzoek kan uitgebreider gekeken worden naar dit effect
577 en wat dit betekent voor de resultaten. Het percentage documenten van een
578 vragenuur is tweemaal zo hoog bij foutief geclassificeerde documenten, maar dit
579 lijkt te komen doordat deze documenten vaak kleiner zijn (mediaan is 286).

580 Er is verder nog gekeken naar andere verbanden tussen documenten die
581 verkeerd zijn geclassificeerd. Daarbij is nog te zien dat sprekers met weinig do-
582 cumenten relatief iets meer voorkomen in verkeerd geclassificeerde documenten.



Figuur 7: Aantal misclassificaties gedeeld door totaal aantal documenten per
spreker tegenover totaal aantal documenten van een spreker. Misclassificaties
zijn totaal van 5-fold cross-validation, waardoor documenten vaker mee kunnen
tellen. De pearson correlatie is -0.28 en de p-waarde 1.07×10^{-4} .

Dit versterkt het vermoeden dat de classificatie mede plaatsvindt op basis van woordgebruik van individuele sprekers, zoals besproken wordt in deelvraag 5.

5.2 DV2: Invloed van namen

De resultaten laten zien dat de classificatie sterk afhankelijk is van partijnamen en achternamen van Kamerleden. De hogere scores voor de classificatie met alleen namen dan zonder namen in combinatie met de woorden in tabel 5 suggereert dat dit het belangrijkste was in de classificatie van deelvraag 1. Deze daling was te verwachten op basis van gerelateerd werk.

De n-grams in tabel 8 komen bij veel partijen overeen met hun ideologie, vooral bij de partijen met een sterke focus op één onderwerp; PVV, PvdD en 50PLUS. Daarnaast zijn er ook n-grams die niet veel over ideologie lijken te zeggen, zoals; *volgens mij*, *ik constateer* en *in elk geval*. Vooral de SGP heeft n-grams die niet veel lijken te zeggen over de ideologie, hoewel deze partij desalniettemin de hoogste F_1 score heeft. Met name opvallend hierbij is *mevrouw de voorzitter*, aangezien deze woorden door alle partijen gebruikt worden om via de voorzitter te praten. Voor een vervolgonderzoek kan gekeken naar waarom deze n-grams zo karakteristiek zijn voor partijen.

De classificatiemethode die gebruikt is in deze deelvraag, is gebaseerd op de beste methode voor de dataset uit deelvraag 1. Hierin was gevonden dat een combinatie van uni-, bi- en trigrams het beste resultaat opleverde. In tabel 5 is te zien dat trigrams behoren tot de meest karakteristieke n-grams, hoewel de woorden in trigrams vaak overlappen met uni- en bigrams. In tabel 8 daarentegen zijn er nog maar een paar trigrams, welke grotendeels procedurele zinnen zijn of toevoeging van een lidwoord op een uni- of bigram. Dit verschil suggereert dat trigrams minder belangrijk zijn in de classificatie zonder de namen, dus de classificatiemethode uit deelvraag 1 niet het beste is voor deze classificatie. In vervolgonderzoek kan de opzet van deelvraag 1 toegepast worden op de classificatie zonder de namen, om zo te komen tot een classificatiemethode die het beste resultaat oplevert op de classificatie zonder namen.

Er is ook gekeken naar andere namen in de lijst van 100 meest karakteristieke woorden per partij, zoals van gebieden, bedrijven of bewindspersonen. Bewindspersonen komen hier niet in voor. Er komen een aantal gebieden in voor, zoals *aruba*, *limburg* en *saoedi arabië*. Ook komen er organisaties als *gvo hvo* en *monsanto* in voor. Deze woorden lijken in sommige gevallen een weerspiegeling te zijn voor ideologie, dus voor vervolgonderzoek lijkt het niet nodig te zijn deze te verwijderen.

5.3 DV3: Oppositie of regering

In tabel 4 is het opvallend dat de coalitiepartijen lage scores krijgen. Daarnaast laat figuur 3 zien dat er een hoge overlap zit tussen deze twee partijen.

De statistische toetsresultaten in tabel 9 laten zien dat inderdaad de error groter is binnen oppositie of regering dan tussen een regerings- en oppositiepartij. Dit suggereert dat inderdaad partij-status invloed heeft op de classificatie.

De verwachting was dat de error normaal verdeeld zou zijn. De verdelingen uit figuur 4 hebben globaal wel de vorm van een normaal verdeling. In figuur 2 is het daarnaast opvallend dat partijen zoals SP en PVV ruim onder de

regressielijn zitten, terwijl andere partijen er een stuk boven zitten. Dit geeft aanleiding te vermoeden dat er naast het aantal documenten van een partij nog meer factoren van invloed zijn op het aantal misclassificaties en daarmee de verwachte waarde. En deze verwachte waarde en de daar uit volgende error zijn een belangrijke aanname van deze methode. Voor deze methode is het dus belangrijk uit te vinden of dit een goede benadering is van de verwachte waarde. In deelvraag 4 wordt gekeken of links/rechts positie hier nog invloed heeft. Voor een vervolgonderzoek kan nog verder gekeken worden naar invloeden op verwachte waarde of andere confounding biases.

De overlap van 100 meest karakteristieke n-grams tussen regeringspartijen die niet voorkomen bij oppositiepartijen gedurende kabinet-Rutte II beperkt zich tot de woorden *en* en *blij*, als ook *toezegging* voor VVD en *toezeggingen* voor PvdA.

Tabel 14: N-grams die bij minimaal één regeringspartij in beide kabinetten voorkomen in de 100 meest karakteristieke n-grams, maar niet voor één van de twee partijen tijdens het andere kabinet.

		Kabinet-Rutte II	
		PvdA	VVD
Kabinet-Balkenende IV	CDA	<i>toezeggingen</i> <i>hun</i> <i>collega KAMERLID</i> <i>in</i> <i>aanpak</i> <i>collega</i>	<i>algemeen</i> <i>algemeen overleg</i> <i>toezegging</i> <i>helder</i> <i>overleg</i> <i>aangegeven</i> <i>voor</i> <i>voor PARTIJ</i>
	ChristenUnie	<i>mijn</i> <i>waarop</i> <i>blij</i> <i>collega KAMERLID</i> <i>erg</i>	<i>gaan</i> <i>termijn</i> <i>blij met de</i> <i>volgens</i> <i>volgens mij</i> <i>blij</i> <i>beantwoording</i>
	PvdA		<i>volgens</i> <i>volgens mij</i>

Hoewel er een aantal overeenkomsten zijn qua meest karakteristieke n-grams tussen regeringspartijen van de twee kabinetten, lijkt dit beperkt. De meeste overeenkomsten lijken daarnaast niet heel inhoudelijk gerelateerd aan partij-status. Deze resultaten suggereren daarom ook maar een beperkte invloed van partij-status op de classificatie. Voor een vervolgonderzoek kan uitgebreider gekeken worden naar de overlappende meest karakteristieke n-grams en wat deze zeggen over een regeringspartij.

De scores in tabel 11 laten een duidelijke daling zien ten opzichte van een classificatie van alleen kabinet-Rutte II. Deze algemene daling kan verklaard worden door verschuiving in ideologie, verschil in woordgebruik, verandering van onderwerpen en/of verandering in aantal documenten per partij. De daling

is het grootst bij VVD, maar valt mee bij de twee andere partijen die gewisseld zijn van partij-status, ChristenUnie en CDA. Daarnaast is de daling ook heel sterk bij oppositiepartijen GroenLinks en D66, alsook de regeringspartij in beide kabinetten, PvdA. Dat de daling niet consequent groter is bij partijen die gewisseld zijn van partij-status, suggereert dat de invloed van partij-status beperkt is op de classificatie.

Dat de experimenten uit Hirst et al. in hun onderzoek wel invloed vinden, maar in dit onderzoek niet kan komen doordat hun onderzoek zich richt op binaire classificatie, terwijl dit onderzoek meerdere partijen heeft. Zo kan het ontbreken van gemeenschappelijke n-grams komen doordat regeringspartijen zich ook van elkaar moeten onderscheiden in dit onderzoek, waarvoor n-grams die relevant zijn voor partij-status weinig effect hebben, terwijl in het onderzoek van Hirst et al. de regeringspartij alleen onderscheiden hoeft te worden van de oppositiepartij. Daarnaast verklaren zij dat een daling tussen twee zittingsperiodes met een wisseling van partij-status het gevolg is van deze wisseling, terwijl in dit onderzoek gekeken kan worden naar dit effect voor partijen die wel en niet gewisseld zijn.

5.4 DV4: Links/rechts

De resultaten

Alle correlaties zijn statistisch significant, maar de pearson correlatie suggereert dat het verband zwak is.

Er zijn verschillende visies op links en rechts, en de indeling van partijen op die links. Daarnaast zijn er nog meerdere dimensies waarlangs partijen vergeleken kunnen worden. Bijvoorbeeld op basis van conservatief tegenover progressief. Een vervolgonderzoek kan uitgebreider kijken naar welke dimensies relevant zijn voor partijen in de Tweede Kamer en in hoeverre deze invloed hebben op de classificatie.

5.5 DV5: Woordgebruik van sprekers

De resultaten uit tabel 13 zijn laag, amper hoger dan de baseline. Dit suggereert inderdaad dat eerdere classificaties in grote mate toch afhankelijk waren van het woordgebruik van sprekers. Dit is opmerkelijk aangezien vergelijkbare onderzoeken dit effect niet vinden. De meest karakteristieke n-grams van deze classificatie wijken daarnaast grotendeels niet af van die uit tabel 8.

Een alternatieve verklaring is dat de classificatie nu mede op basis van woordvoerderschap is. Per onderwerp heeft een partij vaak maar één woordvoerder, met uitzonderingen van wijzigingen in de fractie. Het is aannemelijk dat het taalgebruik afhankelijk is van woordvoerderschap, aangezien er andere termen gebruikt worden bij bijvoorbeeld een debat over zorg dan bij een debat over onderwijs. Als een woordvoerder op een bepaald onderwerp van een partij in de test set voorkomt, is er een grote kans dat geen enkele spreker van die partij eerder over dat onderwerp heeft gepraat, want de woordvoerder gaat nou eenmaal daarover. Daardoor heeft deze spreker veel n-grams die ook voorkomen bij andere woordvoerders over dat onderwerp, maar van andere partij. Als deze n-grams ook belangrijk zijn voor de classificatie kan het zijn dat de woordvoerder geassocieerd wordt bij een partij van een andere woordvoerder. Een vervolgonderzoek kan kijken of dit een verklaring is.

Vergelijkbare onderzoeken vermijden dit mogelijke probleem door alle spreekbeurten van een spreker samen te voegen tot één document. Zoals al eerder vermeld is dit onpraktisch voor de kleinere partijen. Voor een vervolgonderzoek kan desalniettemin gekeken worden naar deze methode om te kijken of dat wel een weerspiegeling is van ideologische verschillen.

5.6 Algemeen

Het vergelijken van deze resultaten met vergelijkbaar werk is ingewikkeld, gezien de keuzes en eigenschappen van die onderzoeken het niet een één-op-één vergelijking maken. Voorbeelden hiervan zijn de taal, het parlement, de documentgrootte, baselines, behouden of weglaten van namen, een spreker als document zien en het trainen en testen op dezelfde spreker. Hoewel de resultaten in sommige gevallen lager zijn dan die uit vergelijkbaar werk, is het belangrijk hier rekening mee te houden. Een vervolgonderzoek zou daarom dit onderzoek kunnen reproduceren op een ander parlement om daarmee te kunnen vergelijken.

Dit onderzoek richtte zich hoofdzakelijk op de Handelingen gedurende kabinet-Rutte II. Om te kijken in hoeverre het mogelijk is om deze conclusie door te trekken naar de algemene Handelingen van de Tweede Kamer, kan er in vervolgonderzoek gekeken worden naar meerdere zittingsperioden. Ook kan gekeken worden naar veranderingen als een kabinet demissionair is.

Dit onderzoek heeft een aantal beperkingen die in dit hoofdstuk besproken zijn. Het uitvoeren van deze aanbevelingen kan de validiteit en betrouwbaarheid van dit onderzoek vergroten. Ook is dit onderzoek moeilijk te vergelijken met andere onderzoeken om diverse redenen, maar vooral ook omdat het toegepast is op een ander parlement. Desalniettemin geeft dit onderzoek reden om te twijfelen aan de bruikbaarheid van tekstclassificatie van de Handelingen van de Tweede Kamer voor een relatie tussen woordgebruik en ideologie. Daarnaast levert dit onderzoek ook kritieken op een aantal vergelijkbare onderzoeken.

6 Conclusies

Dit onderzoek vindt een *accuracy* en F_1 score van 0.80 voor het classificeren van spreekbeurten in de Tweede Kamer naar partij-affiliatie. De baseline scores zijn respectievelijk 0.11 en 0.15. Als rekening wordt gehouden met partijnamen en achternamen Kamerleden daalt de *accuracy* naar 0.58 en de F_1 score naar 0.57. Dit onderzoek vindt aanwijzingen dat deze classificatie afhankelijk is van de partij-status (oppositie of regering). Als rekening wordt gehouden met woordgebruik van individuele Kamerleden, daalt de nauwkeurigheid verder naar.... Hoewel dit onderzoek hoge scores vindt voor classificatie, lijken deze in grote mate afhankelijk te zijn van andere factoren dan ideologie.

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? In Bertie Kaal, Isa Maks, and Annemarie van Elfrinkhof, editors, *From Text to Political Positions*, chapter 5, pages 93–115. John Benjamins Publishing Company, Amsterdam, 2014.
- [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.
- [5] Bei Yu, Stefan Kaufmann, and Daniel Diermeier. Classifying party affiliation from political speech. *Journal of Information Technology & Politics*, 5(1):33–48, 2008.
- [6] Maneesh Bhand, Dan Robinson, and Conal Sathi. Text classifiers for political ideologies, 2009.
- [7] Bjørn Høyland, Jean-François Godbout, Emanuele Lapponi, and Erik Vellidal. Predicting party affiliations from european parliament debates. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 56–60. Association for Computational Linguistics, 2014.
- [8] Laura Klompenhouwer. Extra ledenvergadering 50plus om splitsing. *NRC Handelsblad*, June 2014.
- [9] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001.
- [10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [11] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.
- [12] Mahendra Sahare and Hitesh Gupta. A review of multi-class classification for imbalanced data. *International Journal of Advanced Computer Research*, 2(3), 2012.
- [13] *NIST/SEMATECH e-Handbook of Statistical Methods*. NIST/SEMATECH, April 2012.
- [14] Andrea Volkens, Pola Lehmann, Theres Matthieß, Nicolas Merz, Sven Regel, and Bernhard Weßels. The manifesto data collection. manifesto project (mrg/cmp/marpor). version 2017b, 2017.