

1 IDEOLOGIE EN CLASSIFICATIE IN DE HANDELINGEN
2 VAN 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 FACULTEIT DER NATUURWETENSCHAPPEN, WISKUNDE EN
9 INFORMATICA
10 UNIVERSITEIT VAN AMSTERDAM
11 2018-06-28

12

	Begeleider	Tweede lezer
Titel, Naam	Dr Maarten Marx	
Affiliatie	UvA, FNWI, IvI	
Email	maartenmarx@uva.nl	.



14	Inhoudsopgave	
15	1 Introductie	3
16	2 Gerelateerd werk	4
17	2.1 Classificatiemethoden	4
18	2.2 Invloed van partijnamen of sprekersnamen	5
19	2.3 Invloed van oppositie of regering	5
20	3 Methodologie	6
21	3.1 De data	6
22	3.2 Methoden	8
23	3.2.1 Deelvraag 1	8
24	3.2.2 Deelvraag 2	10
25	3.2.3 Deelvraag 3	10
26	3.2.4 Deelvraag 4	12
27	4 Evaluatie	12
28	4.1 Resultaten	12
29	4.1.1 Deelvraag 1	12
30	4.1.2 Deelvraag 2	14
31	4.2 Discussie	15
32	4.2.1 Deelvraag 1	15
33	4.2.2 Deelvraag 2	15
34	4.2.3 Deelvraag 4	15
35	5 Conclusies	15
36	A Slides	16

1 Introductie

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

In diverse landen zijn al verschillende onderzoeken gedaan naar het classificeren van partij-affiliatie op basis van teksten van politici[3, 1]. Met deze tekst classificatie proberen onderzoekers ideologie uit te vinden in hoeverre ideologie terug te vinden is in teksten van politici. Mede omdat elk land een ander stelsel, taal en cultuur heeft, verschillen de resultaten. Elk onderzoek gebruikt ook een andere methode voor het classificeren. Daarnaast vindt het onderzoek van Hirst et al. [2] dat deze classificatie minder het gevolg is van ideologie maar meer van regering tegenover oppositie. Deze onderzoeken worden besproken in het gerelateerd werk.

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

Dit onderzoek richt zich daarom op een breder scala aan mogelijke methoden 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 van partij-affiliatie in de Tweede Kamer en wat is het resultaat van dit model?
2. In hoeverre is deze classificatie afhankelijk van partijnamen en namen van Kamerleden?
3. In hoeverre wordt deze classificatie bepaald door partij-status (d.w.z. oppositie of regering)?
4. In hoeverre wordt deze classificatie bepaald door links/rechts verdeling?

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

Overzicht van scriptie Sectie 2 bevat gerelateerd werk, met name vergelijkbare onderzoeken in andere landen. Sectie 3 bevat de methodologie van de verschillende deelvragen. Sectie 4 bevat de resultaten. Sectie 5 bevat de evaluatie van zowel de resultaten als de gehanteerde methodologie. Sectie 6 bevat ten slotte het antwoord op de onderzoeksvraag.

2 Gerelateerd werk

Diermeier et al. deden onderzoek naar het classificeren op basis van ideologische 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 dezelfde categoriën 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 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.

Het onderzoek bevat ook een classificatie van het Europees Parlement. Hierbij voegen ze alle teksten van een parlamentslid bij elkaar en delen die op in stukken van gelijke grootte. Zij vinden voor documentgrootte van 267 woorden een nauwkeurigheid van 44,0% oplopend tot 61,8% voor documentgrootte van 6666.

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 minimale 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 wegenen 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

127 ze support vector machines en naive bayes classificatie. Voor het selecteren
128 van *features* experimenteerden ze met een simpele minimale frequentie en het
129 gebruik van een top aantal woorden op basis van mutual information. Uitein-
130 delijk was het beste model bij hen een met support vector machine, met uni- en
131 bigrams, gekozen op basis van mutual information.

132 In het onderzoek van Ferreira werd gebruik gemaakt van twee classifi-
133 catiemethoden: Logistische regressie en MIRA[3]. Logistische regressie werd
134 aangevuld met *group Lasso* regularisatie. Voor wegen van woorden werd
135 geëxperimenteerd met *tf*, *tf-idf*, Δ -*tf-idf* en Δ -*BM-25*. Daarnaast wordt er ge-
136 bruik gemaakt van woordclustering, *Concise Semantic Analysis* en stylome-
137 trische eigenschappen. Op *Part-Of-Speech tagging* na hadden stylometrische
138 eigenschappen een duidelijke negatieve invloed op de classificatie.

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

144 2.2 Invloed van partijnamen of sprekersnamen

145 In het onderzoek van Diermeier et al. zijn alle namen weggelaten, omdat deze
146 volgens hen de classificatie te makkelijk zouden maken [4]. Hirst et al. vinden in-
147 derdaad dat partijnamen (en het weglaten daarvan) bij het Europees Parlement
148 een grote invloed hebben op de classificatie [2]. Bij het Europees Parlement zien
149 zij met name het gebruik van de eigen partijnaam door een spreker, terwijl zij in
150 het Canadese parlement vooral zien dat de naam van de andere partij gebruikt
151 wordt door een spreker.

152 2.3 Invloed van oppositie of regering

153 Graeme Hirst et al. vonden in hun onderzoek dat de classificatie van spreker in
154 het Canadese parlement op basis van partij-affiliatie meer zegt over de status
155 van de partij (regering of oppositie).[2] Zo vergeleken zij de top tien karakteris-
156 tieke woorden van de liberalen en conservatieven in het 36e parlement (liberalen
157 in de regering) en het 39e parlement (conservatieven in de regering. Hier vonden
158 zij dat vier van de tien woorden van de liberalen (regering) in het 36e parlement
159 bij het 39e parlement bij de conservatieven (regering) te vinden waren. An-
160 dersom gebeurde hetzelfde met één van de tien woorden van de conservatieven
161 (oppositie) in het 36e parlement naar liberalen (oppositie) in het 39e parlement.

162 In hetzelfde onderzoek trainden ze ook hun classifiers op het ene parle-
163 ment en testten deze op het andere parlement. Hierbij vonden zij in beide
164 gevallen een nauwkeurigheid ver onder de baseline. Daarnaast deden ze ook
165 nog een classificatie op de sprekers die in beide parlementen zaten en een an-
166 dere classificatie op sprekers die niet in beide parlementen zaten. Bij de eerste
167 classificatie vonden ze nauwkeurigheden rond de baseline, terwijl in de tweede
168 situatie nauwkeurigheden gevonden werden ver boven de baseline.

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

171 3 Methodologie

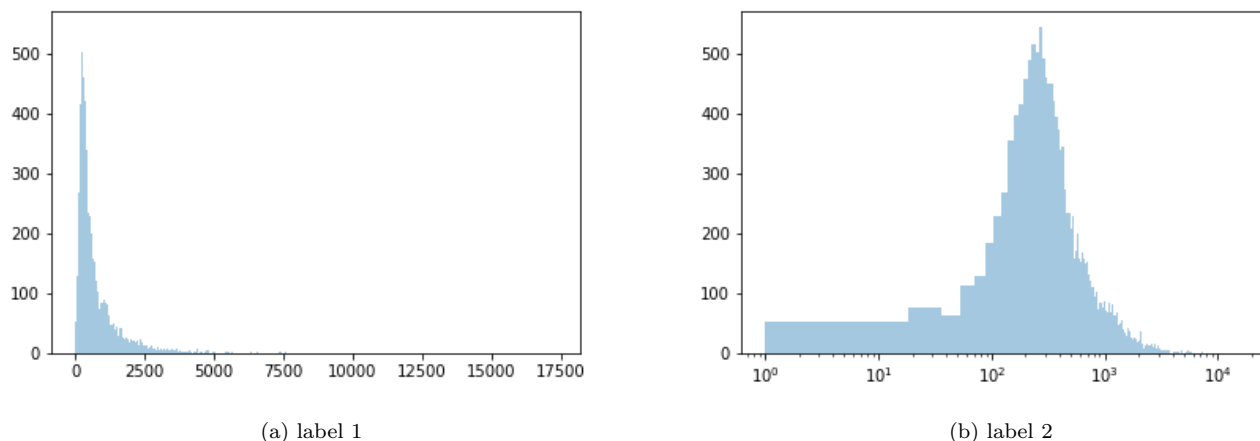
172 3.1 De data

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

183 Deze dataset bestaat uit een aantal soorten spreekbeurten; debat bijdra-
184 gen, interrupties en antwoorden. Debat bijdrage is de eerste onafgebroken
185 spreekbeurt die een spreker geeft achter het spreekgestoelte, aangeduid in de
186 xml-file met het attribuut *nieuw="ja"*. Dit kan een bijdrage in een debat zijn
187 of een vraag tijdens een vragenuur. Interrupties zijn de vragen die andere po-
188 litici stellen vanachter de interruptiemicrofoon aan de spreker. De antwoorden
189 zijn vervolgens de reactie van een spreker achter het spreekgestoelte op een in-
190 terruptie. Aangezien een debat bijdrage geïnterrupteerd kan worden, kan deze
191 inhoudelijk doorlopen in een antwoord van een spreker. Er is in dit onderzoek
192 ervoor gekozen om gebruik te maken van een debat bijdrage samengevoegd tot
193 één document met alle bijbehorende antwoorden van die spreker. Daarnaast zijn
194 er verschillende soorten sprekers; de voorzitter, Tweede Kamerleden, leden van
195 het kabinet en gastsprekers. Daarnaast is alleen gekozen voor sprekers waarvan
196 er een partij-affiliatie vermeld staat, dit is niet het geval voor leden van het
197 kabinet, de voorzitter en gastsprekers (met uitzondering van Nederlandse leden
198 van het Europees Parlement).

199 Deze dataset bevat vervolgens naast de verkozen partijen van de 2012
200 Tweede Kamerverkiezingen, ook afsplitsingen van die partijen (tien in totaal) en
201 bezoeken van vertegenwoordigingen van Nederlandse partijen uit het Europees
202 Parlement (tien in totaal). Omdat van beide categorieën relatief weinig data is
203 en er overlap zit met hun oorspronkelijke partij, zijn deze er uit gehaald.

204 De documenten verschillen vervolgens in grootte. De distributie lijkt op
205 een lognormale verdeling, maar met een Kolmogorov-Smirnov test is hier geen
206 bewijs voor gevonden [7].



Figuur 1: Aantal woorden per document

Om toch de uitschieters er uit te halen, is aangenomen dat het 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. Het gemiddelde is daarna 498 woorden en de mediaan is 386 woorden. Een totaal aantal documenten van 14899 blijven vervolgens over.

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

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 nauwkeurigheid van 0.15 (door altijd grootste partij te kiezen) en baseline F_1 score van 0.11 (door willekeurig te voorspellen gewogen bij aantal spreekbeurten in klasse).

218 3.2 Methoden

219 3.2.1 Deelvraag 1

220 Om deze deelvraag te beantwoorden zullen een aantal classificatiemethoden ver-
221 geleken worden. Aangezien het onmogelijk is om alle classificatiemethoden te
222 vergelijken, beperkt dit onderzoek zich tot classificatiemethoden die gebruikt
223 zijn in vergelijkbare onderzoeken, zoals besproken in 2.1. Er is ervoor gekozen
224 om alleen gebruik te maken van methoden waarvan reeds implementaties be-
225 schikbaar waren in Python. Voor alle methoden wordt gezocht naar de beste
226 parameters; een grid search. Deze grid search wordt gedaan door middel van
227 5-fold cross-validation, waarbij de trainings set steeds 80% is en de test set 20%
228 van de totale dataset.

229 **Pre-processing** Voor pre-processing is gebruik gemaakt van tokenisation en
230 lowercasing. Voor tokenisation is de reguliere expressie
231 *w+* gebruikt, die daarmee alleen de letters en cijfers overhoudt. Deze woorden
232 zijn vervolgens allemaal omgezet in kleine letters. Vervolgens is er gevarieerd
233 tussen wel of geen gebruik maken van stemming. In het geval van stemming is
234 gebruik gemaakt van de Snowball Stemmer via de Python NLTK module.

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

241 Voor dit onderzoek zijn de volgende wegeningen voor woorden getest: *boolean*
242 (wel of niet aanwezig), *tf* (woordfrequentie), *tf-norm* (woordfrequentie genor-
243 maliseerd door documentlengte) en *tf-idf*. Daarnaast wordt in dit onderzoek
244 geëxperimenteerd met een minimale of maximale woord- of documentfrequentie.
245 Ook is gekeken naar het effect van combinaties van n-grams; unigrams, bigrams
246 en trigrams. N-grams zijn combinaties van N aantal opeenvolgende woorden.
247 Bij een unigram is elke feature gewoon één woord, terwijl bij een bigram dit
248 twee opvolgende woorden zijn. Dit kan nuttig zijn, want als bijvoorbeeld het
249 woord *asfalt* er in voorkomt, dan maakt het voor ideologie waarschijnlijk meer
250 uit of er *minder asfalt* of *meer asfalt* staat.

251 **Support Vector Machines en Logistische Regressie** De meest voorko-
252 mende techniek in vergelijkbaar onderzoek is Support Vector Machine (SVM).
253 Een andere techniek die gebruikt wordt is logistische regressie. Beide kennen
254 een eigen implementatie in sklearn, maar deze implementaties zijn niet efficiënt met
255 grote datasets. Om deze reden is er in beide gevallen voor gekozen om gebruik
256 te maken van de functie `SGDClassifier`, die beide technieken leert met *stochastic*
257 *gradient descent learning*. Er is hiervoor gevarieerd met de regularisatie, learning
258 rate en maximum aantal iteraties. Voor regularisatie is hier geëxperimenteerd
259 met Lasso en Ridge regularisatie, en een combinatie van beide genaamd Elastic-
260 net. De andere parameters zijn gelaten op de standaardwaarden van scikit-learn
261 [8].

262 **Naive Bayes** Een simpelere techniek die gebruikt wordt voor politieke tekst-
 263 classificatie is Naive Bayes. Dit algoritme neemt aan dat elke *feature* onafhan-
 264 kelijk is ten op zichte van de rest. Dit is bij tekstclassificatie vaak niet het geval
 265 omdat het gebruik van sommige woorden gepaard kan gaan met het gebruik
 266 van andere woorden. Daarnaast is het gebruik van meerdere n-grams in een
 267 classificatie schending van de aanname, want als bijvoorbeeld een bigram er in
 268 voorkomt dan komen ook beide unigrams er sowieso in voor. Desalniettemin
 269 blijkt Naive Bayes effectief te zijn voor tekstclassificatie[8, 5]. Hiervoor zijn de
 270 functies van scikit-learn MultinomialNB en BernoulliNB gebruikt.[8, 5]

271 **Beoordelen van kwaliteit** De meest gebruikte methoden om kwaliteit van
 272 politieke tekstclassificatie te beoordelen zijn nauwkeurigheid en F_1 score, die
 273 opgebouwd is uit recall en precision. Deze scores zijn opgebouwd uit vier varia-
 274 belen te zien in tabel 2. Deze variabelen geven weer hoeveel documenten wel of
 275 niet bij een klasse horen, en of deze wel of niet als dusdanig zijn geclassificeerd.

Tabel 2: BLABLA

	Behorend tot klasse	Niet behorend tot klasse
Geclassificeerd als klasse	tp	fp
Niet geclassificeerd als klasse	fn	tn

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

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

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

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

279 Nauwkeurigheid is het percentage van documenten dat correct geclassificeerd
 280 is. Precision is het percentage van documenten geclassificeerd als klasse, dat
 281 ook bij die klasse hoort. Recall is het percentage documenten van documenten
 282 behorende tot een klasse, dat ook als dusdanig geclassificeerd is. F_1 is het
 283 harmonisch gemiddelde van recall en precision. Precision, recall en dus ook F_1
 284 worden in beginsel per klasse berekend. Er zijn drie varianten om deze scores
 285 voor de hele classificatie te berekenen.

286 Allereerst is er *micro*, daarbij worden alle waarden bij elkaar opgeteld en
 287 dan berekend. Dit leidt ertoe dat resultaten van klassen met veel documenten
 288 belangrijker zijn. Als een classificatie kleine klassen grotendeels fout classificeert,
 289 kan deze score alsnog hoog zijn. In het geval van meer dan twee klassen is dit
 290 hetzelfde als nauwkeurigheid.

291 Als tweede is er *macro*, daarbij worden alle scores per klasse berekend en
 292 wordt daarvan het gemiddelde genomen. Dit leidt er dan weer toe dat resultaten
 293 van klassen met weinig documenten net zo belangrijk zijn. Hierdoor kan een
 294 classificatie met een laag aantal correct geclassificeerde documenten hoog scoren
 295 door vooral kleine klassen goed te hebben.

Als laatste is er dan nog *gewogen*, dit berekent net als *macro* de scores per klasse, maar neemt hiervan het gemiddelde gewogen bij het aantal documenten behorend tot een klasse. Deze wijkt weinig af van de *micro* variant, tenzij er uitschieters zijn bij klassen.

Aangezien *micro* al terugkomt in nauwkeurigheid en het nadeel van *macro* te groot is omdat de klassen nogal variëren in grootte, is gekozen voor *gewogen* F_1 scoring naast nauwkeurigheid.

3.2.2 Deelvraag 2

In Diermeier et al. [4] wordt aangenomen dat namen een groot effect hebben op de classificatie en Hirst et al. [2] bevestigt dit voor het Europees Parlement. 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 achternamen 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 namen zijn uit de Handelingen gehaald. Voor partijnamen zijn ook lidwoorden toegevoegd, voor achternamen van Kamerleden zijn ook verkortingen meegenomen. Dit laatste omdat bijvoorbeeld *Van Haersma Buma* vaak aangesproken wordt als *Buma*. Voornamen van Kamerleden worden zelden tot nooit gebruikt, dus die zijn er niet uitgehaald. Een nadeel van deze aanpak is dat ook namen van niet-Kamerleden of andere woorden weggehaald kunnen worden als deze hetzelfde zijn als naam van een Kamerlid. Door gebruik van gevoeligheid voor hoofdletters is geprobeerd dit te voorkomen. Een opvallend voorbeeld hiervan is de naam Rutte, die zowel behoort tot het Kamerlid Arno Rutte als de premier Mark Rutte. Steekproefgewijs is gekeken of er nog namen achter zijn gebleven, maar die zijn niet gevonden.

De nauwkeurigheid en F_1 score worden vervolgens vergeleken met de resultaten uit deelvraag 1. Ook wordt gekeken naar verschillen tussen de meest veelzeggende woorden uit deelvraag 1 en uit deze deelvraag.

3.2.3 Deelvraag 3

Om deze deelvraag te beantwoorden zal een analyse gedaan worden van de confusion matrix en 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 en 2 gebruikt worden.

Als er een confounding bias is op basis 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. Uit de voorverkenning (op basis van resultaten uit deelvraag 1 en 2) blijkt verder dat er een correlatie is tussen het aantal foutief als partij geclassificeerde documenten en het aantal documenten behorend tot die partij. Aannemend dat dit verband causaal is, is het verwachte aantal documenten

$$V_{i,j} = \left(\sum_{k=0}^n (D_{k,j}) - D_{j,j} \right) * \frac{\sum_{k=0}^n (D_{i,k})}{\sum_{k=0}^n (\sum_{l=0}^n D_{k,l}) - \sum_{k=0}^n (D_{k,j})} \quad (5)$$

339 waar $V_{i,j}$ het verwachte aantal documenten is voor $i \neq j$ met i de voorspelde
 340 partij en j de echte partij waar een document bijhoort, $D_{i,j}$ het aantal docu-
 341 menten als dusdanig geclassificeerd en n het aantal partijen. De linkerterm is
 342 het totaal aantal documenten behorende tot partij i die fout geclassificeerd zijn.
 343 De rechterterm is het percentage van het totaal aantal documenten minus die
 344 van partij i dat tot partij j behoort.

345 De error is dan het verschil van de verwachte waarde en het daadwerkelijk
 346 aantal documenten

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

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

349 Als dit een goede benadering is van de error, dan is het te verwachten
 350 dat deze normaal verdeeld is [9]. Om te kijken of er een confounding bias
 351 is, worden de distributies binnen regeringspartijen, binnen oppositiepartijen en
 352 tussen beide groepen met elkaar vergeleken. Om de invloed van variantie door
 353 de willekeurige splitsing documenten voor trainen en testen te beperken, wordt
 354 de classificatie 50 keer gedaan en worden deze errors bij elkaar in distributie
 355 genomen. De nulhypothese is dat er geen verschil is tussen de verdelingen. De
 356 alternatieve hypothese is dan dus dat er wel een verschil is tussen de verdelingen.
 357 Als de nulhypothese wordt verworpen, kan dus aangenomen worden dat er een
 358 verschil is op basis van partij-status.

359 In het eerste experiment uit Graeme Hirst et al. zullen de tien meest ka-
 360 rakteristieke woorden per partij van de ene zittingsperiode vergeleken worden
 361 met de tien meest karakteristieke woorden per partij van de andere zittingspe-
 362 riode. Als de classificatie op basis van ideologie is in plaats van partij-status, is
 363 het te verwachten dat de woorden bij een partij blijven en niet gekoppeld zijn
 364 aan in oppositie of regering zitten.

365 In het tweede experiment uit Graeme Hirst et al. worden classifiers ge-
 366 traind op de ene zittingsperiode en getest op de andere zittingsperiode. Als de
 367 classificatie op basis van ideologie is in plaats van partij-status, is de verwach-
 368 ting dat er nog steeds aanzienlijke voorspellingen gedaan worden, aangezien de
 369 ideologie naar verwachting redelijk stabiel is binnen tien jaar (hoewel woordge-
 370 bruik varieert). Als de scores aanzienlijk lager zijn, kan dit het gevolg zijn van
 371 het veranderen van partij-status van partijen.

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

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

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 beste 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[10] 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 3: 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

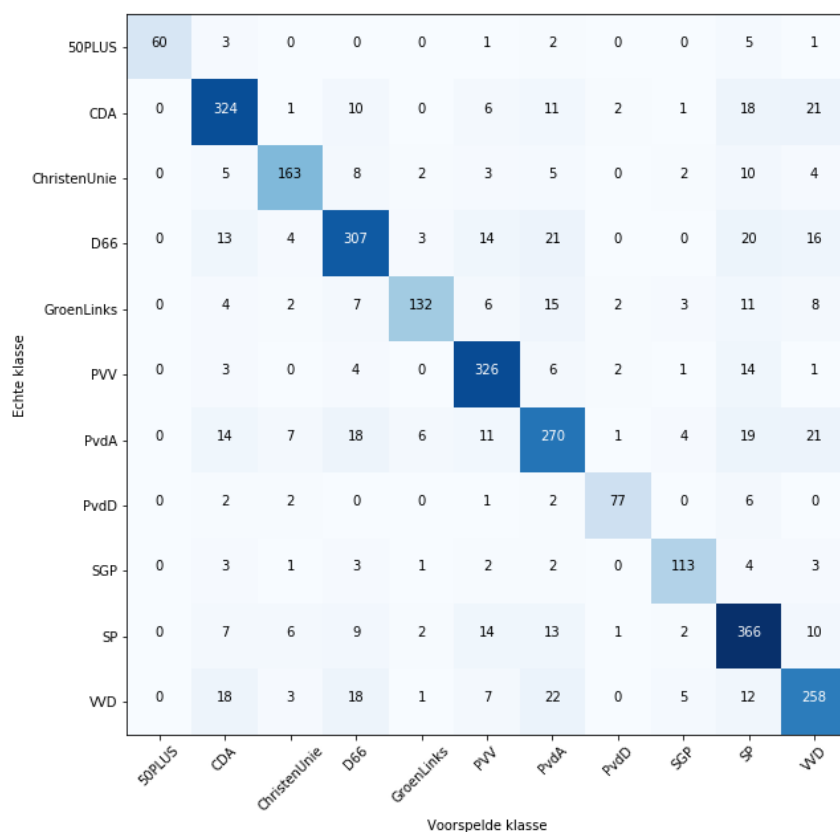
4 Evaluatie

4.1 Resultaten

4.1.1 Deelvraag 1

Het beste resultaat werd bereikt met SVM gebruikmakend van *stochastic gradient descent learning* en Ridge regularisatie.

Figuur 2 laat zien waar de fouten in deze classificatie zitten. De meest karakteristieke features per partij zijn te zien in figuur 4. Beide resultaten zijn op basis van één classificatie.



Figuur 2: Confusion matrix van beste classificatie.

Tabel 4: Meest relevante woorden per partij op basis van beste classificatie gedurende kabinet-Rutte II.

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

Tabel 4: Meest relevante woorden per partij op basis van beste classificatie gedurende kabinet-Rutte II. (*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
islamitische	van de arbeid	lid ouwehand naar	leden dijkgraaf	sp fractie	de vvd is
klever	de arbeid	thieme	de leden dijkgraaf	de sp fractie	vvd is
madlener naar mij	pvda fractie	ouwehand naar mij	leden van der	smaling	vvd fractie
madlener naar	de partij van	ouwehand naar	sgp fractie	het lid smaling	de vvd fractie
graaf	de pvda fractie	de dieren	de sgp fractie	lid smaling	voor de vvd
graus	arbeid	dieren	de leden bisschop	lid van gerven	wat de vvd
pvv fractie	partij van de	voor de dieren	leden bisschop	lid smaling naar	voorzitter de vvd
de pvv fractie	partij van	ouwehand	mevrouw de voorzitter	lid leijten	de vvd betreft

4.1.2 Deelvraag 2

In figuur 4 was al te zien dat de meest karakteristieke woorden voornamelijk bestaan uit partijnamen en namen van Kamerleden.

In figuur 5 is vervolgens te zien welke woorden het meest karakteristiek zijn per partij, als partijnamen namen van Kamerleden vervangen zijn door een generieke placeholder.

Tabel 5: Meest relevante woorden per partij op basis van classificatie uit deelvraag 1 zonder partijnamen of namen van Kamerleden gedurende kabinet-Rutte II.

50PLUS	CDA	ChristenUnie	D66	GroenLinks
gepensioneerden	inwoners	mensenhandel	mijn fractie	zou
ouderen	PARTIJNAAM fractie	inderdaad	mijn	in elk geval
koopkrachtontwikkeling	regering	gezinnen	natuurlijk	elk geval
oudere	PARTIJNAAM	begeleiding	buitengewoon	in elk
plussers	hier	rechtsstaat	het kabinet	belastingontwijking
50	de regering	het ingezetenen criterium	vandaag	kamer hierover te
50 plussers	echt	voedselverspilling	fractie	voorzitter motie de
werkenden	bepaalde	ingezetenen criterium	daarom	voorzitter motie
mevrouw de voorzitter	zouden	zullen	belangrijk	vergroening
mevrouw de	microplastics	de eurozone	kabinet	sekswerkers

Tabel 5: Meest relevante woorden per partij op basis van classificatie uit deelvraag 1 zonder partijnamen of namen van Kamerleden gedurende kabinet-Rutte II. (*Vervolg*)

PVV	PvdA	PvdD	SGP	SP	VVD
islamitische	open standaarden	dieren	mevrouw de voorzitter	huurders	PARTIJNAAM
natuurlijk	keurmerk	bio industrie	mevrouw de	bezuinigingen	regelgeving
al	jongeren	de bio	punt	zegt	aruba
nederland	daarbij	bio	beantwoording	mening dat	aangegeven
de islam	mijn partij	dierenwelzijn	voor de beantwoording	voorstellen	speelveld
miljarden	beter	industrie	wel	bevolking	essentieel
islam	toezeggingen	natuur	de beantwoording	de bevolking	PARTIJNAAM
miljard	collega	de bio industrie	bewindslieden	armoede	volgens mij
belastingbetaler	hulpverleners	aarde	nadrukkelijk	mensen	banen
verzorgingshuizen	vragen	milieu	allereerst	van mening dat	verantwoorde

4.2 Discussie

4.2.1 Deelvraag 1

Dit onderzoek heeft zich beperkt tot methoden genoemd in vergelijkbare onderzoeken en waarvan de implementatie beschikbaar is in Python. 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. Omdat dus niet alle opties getest zijn, kan geen uitsluitel gegeven worden dat dit daadwerkelijk het classificatiemodel is. Voor vervolgonderzoek kan daarom gekeken worden naar meer verschillende methoden.

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 minimumfrequentie lager ligt dan de kleinste documentgrootte uit het onderzoek van Hirst et al. Het effect wat zij vinden tussen documentgrootte van 267 en 6666 is een verschil in nauwkeurigheid van 19,8%. Voor een vervolgonderzoek kan gekeken worden naar of dit effect er is en wat dit betekent voor de resultaten.

4.2.2 Deelvraag 2

4.2.3 Deelvraag 4

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

5 Conclusies

Referenties

- [1] Felix Bießmann. Automating political bias prediction. *CoRR*, abs/1608.02195, 2016.

- 435 [2] Graeme Hirst, Yaroslav Riabinin, Jory Graham, and Magali Boizot-roche.
436 Text to ideology or text to party status? *.
- 437 [3] Vasco Ferreira. Using textual transcripts of parliamentary interventions for
438 profiling portuguese politicians. 2016.
- 439 [4] Daniel Diermeier, Jean-François Godbout, Bei Yu, and Stefan Kaufmann.
440 Language and ideology in congress. *British Journal of Political Science*,
441 42(1):31–55, 2012.
- 442 [5] Conal Sathi Maneesh Bhand, Dan Robinson. Text classifiers for political
443 ideologies, 2009.
- 444 [6] Bjørn Høyland, Jean-François Godbout, Emanuele Lapponi, and Erik Vell-
445 dal. Predicting party affiliations from european parliament debates. In
446 *Proceedings of the ACL 2014 Workshop on Language Technologies and*
447 *Computational Social Science*, pages 56–60. Association for Computatio-
448 nal Linguistics, 2014.
- 449 [7] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source
450 scientific tools for Python, 2001–. [Online; accessed ;today].
- 451 [8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Gri-
452 sel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas,
453 A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay.
454 Scikit-learn: Machine learning in Python. *Journal of Machine Learning*
455 *Research*, 12:2825–2830, 2011.
- 456 [9] *NIST/SEMATECH e-Handbook of Statistical Methods*. NIST/SEMA-
457 TECH, April 2012.
- 458 [10] Andrea Volkens, Pola Lehmann, Theres Matthieß, Nicolas Merz, Sven Re-
459 gel, and Bernhard Weßels. The manifesto data collection. manifesto project
460 (mrg/cmp/marpor). version 2017b, 2017.
- 461 [11] Bei Yu, Stefan Kaufmann, and Daniel Diermeier. Classifying party affilia-
462 tion from political speech. *Journal of Information Technology & Politics*,
463 5(1):33–48, 2008.

464 A Slides