

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	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	11
27	3.2.4 DV4: Links of rechts	13
28	3.2.5 DV5: Woordgebruik van sprekers	14
29	4 Evaluatie	14
30	4.1 Resultaten	14
31	4.1.1 DV1: Beste classificatiemethode	14
32	4.1.2 DV2: Invloed van namen	17
33	4.1.3 DV3: Oppositie of regering	19
34	4.1.4 DV4: Links of rechts	22
35	4.1.5 DV5: Woordgebruik van sprekers	22
36	4.2 Discussie	22
37	4.2.1 DV1: Beste classificatiemethode	22
38	4.2.2 DV2: Invloed van namen	23
39	4.2.3 DV3: Oppositie of regering	24
40	4.2.4 DV4: Links of rechts	24
41	4.2.5 DV5: Woordgebruik van sprekers	24
42	4.2.6 Algemeen	24
43	5 Conclusies	25
44	A Slides	26

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-affiliatie. Het classificeren op basis van tekst kan inzichten geven over 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 [3, 1]. 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. Maar diverse onderzoeken wijzen ook naar redenen dat dit niet alleen het gevolg is van ideologie. De resultaten van Hirst et al. [2] suggereren dat de partij-status (oppositie tegenover regering) van invloed is op de classificatie. Daarnaast laat dit onderzoek ook zien dat de partijnamen belangrijk zijn in 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 meer 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 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?
5. In hoeverre wordt deze classificatie bepaald door woordgebruik van sprekers?

Voor de eerste deelvraag zullen Support Vector Machine, Logistische Regressie en Naive Bayes vergeleken worden aan de hand van *accuracy* en F_1 score. Bij de tweede deelvraag wordt gekeken naar het effect van het weglaten van partijnamen en namen van Kamerleden. De derde vraag bestaat uit meerdere experimenten, waarin gekeken zal worden naar of de misclassificaties binnen coalitie of oppositie groter zijn dan daartussen, en of er tussen die groepen verschillen zitten in de confusion matrix.

Overzicht van scriptie Sectie 2 bevat gerelateerd werk, met name vergelijkbare onderzoeken in andere landen. Sectie 3 bevat de methodologie van de

90 verschillende deelvragen. Sectie 4 bevat de resultaten. Sectie 5 bevat de eva-
91 luatie van zowel de resultaten als de gehanteerde methodologie. Sectie 6 bevat
92 ten slotte het antwoord op de onderzoeksvraag.

93 2 Gerelateerd werk

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

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

104 2.1 Tekstclassificatie van parlementaire teksten

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

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

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

127 Vervolgens herhaalden ze de classificaties op het huis uit 2015, maar testten
128 ditmaal op de senaat elk jaar tussen 1989 en 2006 afzonderlijk. Hier zien zij een
129 stijging in nauwkeurigheid van 60% (baseline van 55.0%) in 1989 naar 87.0%
130 (baseline van 55.0%) in 2006, maar met twee duidelijke dalen. Ze presenteren
131 twee mogelijke verklaringen voor de trend; het veranderen van de onderwerpen
132 en het meer partisan worden van het congres.

133 Als een vervolg op deze onderzoeken deden Graeme Hirst et al. een verge-
134 lijkbare onderzoek naar het Canadese Parlement [2]. Hierbij werd zowel gekeken

naar de Engelse als Franse teksten. Een document werd hier gezien als de samen-
voeging van alle spreekbeurten van een spreker. Afhankelijk van taal en dataset
vinden zij in dit onderzoek *accuracy* scores van 83.2% en hoger (baseline van
65.5%).

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% (baseline van 38-39%) voor
documentgrootte van 6666.

Het onderzoek van Bhand et al. richtte zich op het classificeren van le-
den van het Amerikaanse congres in 2005, op basis van affiliatie (Republikeins
of Democratisch)[6]. Een document hierbij was in tegenstelling tot eerderge-
noemde onderzoeken een speech. Zij vonden hiervoor uiteindelijk een F_1 score
van 0.68 (baseline niet vermeld).

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

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[7]. Hier verkregen zij
een *macro average* F_1 score van 0.464.

2.2 Classificatiemethoden

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

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

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

Bhand et al. gebruikten verschillende n-grams, inclusief verschillende ma-
nieren van *smoothing*[6]. Ze testten als weging voor features zowel *boolean* als
tf, waarbij ze vonden concludeerden dat *boolean* betere resultaten opleverden.
Voor classificatiemodel experimenteerden ze met SVM en Naive Bayes . Voor
het selecteren van *features* experimenteerden ze met een minimale frequentie en

selectie van woorden op basis van hoogste mutual information. Uiteindelijk was het beste model bij hen een SVM met uni- en bigrams en geselecteerd op basis van mutual information.

In het onderzoek van Ferreira werd gebruik gemaakt van twee classificatiemethoden: Logistische regressie en MIRA[3]. Logistische regressie werd aangevuld met *group Lasso* regularisatie. Voor wegingen van woorden werd geëxperimenteerd met *tf*, *tf-idf*, Δ -*tf-idf* en Δ -*BM-25*. Daarnaast wordt er gebruik gemaakt van woordclustering, *Concise Semantic Analysis* en stylometrische eigenschappen. Op *Part-Of-Speech tagging* na hadden stylometrische eigenschappen een duidelijke negatieve invloed op de classificatie.

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

2.3 Invloed van partijnamen of sprekersnamen

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

2.4 Invloed van oppositie of regering

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

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

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

224 3 Methodologie

225 3.1 De data

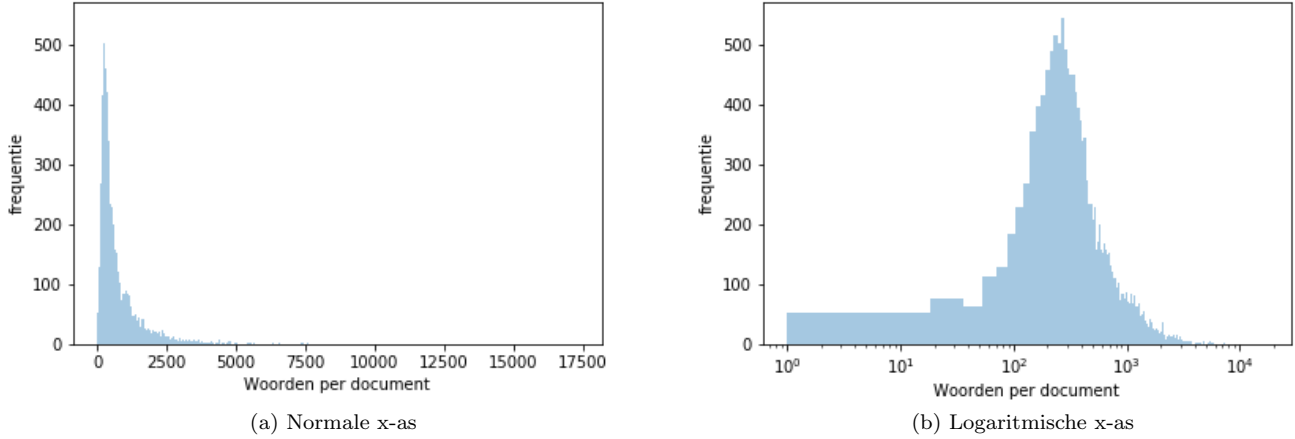
226 De data die gebruikt worden, zijn de Handelingen van de Tweede Kamer gedu-
227 rende het missionaire kabinet-Rutte II (5 november 2012 tot 22 maart 2017). Er
228 is gekozen voor dit kabinet, omdat de data hiervoor makkelijk verkrijgbaar was,
229 het kabinet lang zat - waardoor er veel data is - en het recent is waardoor het
230 makkelijker te interpreteren is. In dit kabinet zaten de PvdA en VVD. Deze data
231 zijn in xml-formaat van de website officiëlebekendmakingen.nl gehaald samen
232 met corresponderende metadata xml-bestanden. De bestanden van de Hande-
233 lingen bevatten voornamelijk informatie over spreekbeurten tijdens een debat,
234 waaronder naam van een spreker, partij-affiliatie, inhoud van de spreekbeurt en
235 het soort spreekbeurt. Deze gegevens zijn samengevoegd tot één tabel.

236 Deze dataset bestaat uit een aantal soorten spreekbeurten voor Kamerle-
237 den; debat bijdragen, interrupties en antwoorden. Debat bijdrage is de eerste
238 onafgebroken spreekbeurt die een spreker geeft achter een spreekgestoelte, aan-
239 geduid in de xml-file met het attribuut *nieuw*="ja". Dit kan een bijdrage in een
240 debat zijn of een vraag tijdens een vragenuur. Interrupties zijn de vragen die
241 andere politici stellen vanachter de interruptiemicrofoon aan een spreker. De
242 antwoorden zijn vervolgens de reactie van een spreker achter het spreekgestoelte
243 op een interruptie. Aangezien een debat bijdrage geïnterrupteerd kan worden,
244 kan deze inhoudelijk doorlopen in een antwoord van een spreker. Er is in dit
245 onderzoek ervoor gekozen om gebruik te maken van een debat bijdrage met alle
246 bijbehorende antwoorden samengevoegd tot één document.

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

252 Deze dataset bevat vervolgens naast de verkozen partijen van de 2012
253 Tweede Kamerverkiezingen, ook afsplitsingen van die partijen (tien in totaal) en
254 bezoeken van vertegenwoordigingen van Nederlandse partijen uit het Europees
255 Parlement (tien in totaal). Omdat van beide categorieën relatief weinig data
256 is en er overlap zit met hun oorspronkelijke partij, zijn deze er uit gehaald.
257 50PLUS is in 2014 [8] uiteengevallen in twee fracties die aanspraak maakten op
258 de partij-affiliatie 50PLUS. Vanaf dit moment zijn deze documenten niet meer
259 meegenomen om onduidelijkheid te voorkomen.

260 De documenten verschillen in grootte. De distributie van documentgrootte
261 lijkt op een lognormale verdeling, maar met een Kolmogorov-Smirnov test is hier
262 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. Het gemiddelde is daarna 498 woorden en de mediaan is 386 woorden. Dit resulteert in een totaal aantal documenten van 14899.

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

274 3.2 Methoden

275 3.2.1 DV1: Beste classificatiemethode

276 Om deze deelvraag te beantwoorden zullen een aantal classificatiemethoden ver-
277 geleken worden. Aangezien het onmogelijk is om alle classificatiemethoden te
278 vergelijken, beperkt dit onderzoek zich tot classificatiemethoden die gebruikt
279 zijn in vergelijkbare onderzoeken, zoals besproken in 2.2. Er is ervoor geko-
280 zen om alleen gebruik te maken van methoden waarvan reeds implementaties
281 beschikbaar waren in scikit-learn. Voor alle methoden wordt gezocht naar de
282 beste parameters, ook wel bekend als een grid search. Deze grid search wordt
283 gedaan door 5-fold cross-validation, waarbij de trainings set steeds 80% is en de
284 test set 20% van de totale dataset. De hypothese is dat de scores lager zijn dan
285 die gevonden in het gerelateerd werk, omdat de documentgrootte kleiner is en
286 de baseline lager.

287 **Pre-processing** Voor pre-processing is gebruik gemaakt van tokenisation en
288 lowercasing. Voor tokenisation is de reguliere expressie
289 *w+* gebruikt, waardoor alles behalve letters en cijfers weggehaald wordt. Ver-
290 volgens is er gevarieerd tussen wel of geen gebruik maken van stemming. In
291 het geval van stemming is gebruik gemaakt van de Snowball Stemmer via de
292 Python NLTK module.

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

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

309 **Support Vector Machines en Logistische Regressie** De meest voorko-
310 mende techniek in vergelijkbaar onderzoek is Support Vector Machine (SVM).
311 Een andere techniek die gebruikt wordt is logistische regressie. Beide kennen een
312 eigen implementatie in scikit-learn, maar deze implementaties zijn niet efficiënt
313 met grote datasets. Om deze reden is er in beide gevallen voor gekozen om
314 gebruik te maken van de functie *SGDClassifier*, die beide technieken leert met
315 *stochastic gradient descent learning*. Voor regularisatie is hier geëxperimenteerd
316 met L1 en L2 regularisatie, en een combinatie van beide genaamd Elasticnet.
317 De andere parameters zijn gelaten op de standaardwaarden van scikit-learn [10].
318 Een belangrijke onaangepaste waarde is die van maximaal aantal iteraties, die

als standaard 5 heeft. 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 gridsearch was het voor dit onderzoek niet mogelijk het maximum iteraties te verhogen.

Naive Bayes Een simpelere techniek die gebruikt wordt voor politieke tekstclassificatie is Naive Bayes. Dit algoritme neemt aan dat elke *feature* onafhankelijk is ten op zichte van de rest. Dit is bij tekstclassificatie vaak niet het geval omdat het gebruik van sommige woorden gepaard kan gaan met het gebruik van andere woorden. Daarnaast is het gebruik van meerdere n-grams in een classificatie schending van de aanname, want als bijvoorbeeld een bigram er in voorkomt dan komen ook beide unigrams er sowieso in voor. Desalniettemin blijkt Naive Bayes effectief te zijn voor tekstclassificatie[10, 6]. Hiervoor zijn de functies van scikit-learn MultinomialNB en BernoulliNB gebruikt.[10, 6]

Beoordelen van kwaliteit De meest gebruikte methoden om kwaliteit van politieke tekstclassificatie te beoordelen zijn *accuracy* en F_1 score, die opgebouwd is uit recall en precision. Deze scores zijn opgebouwd uit vier variabelen. Deze variabelen geven weer hoeveel documenten wel of niet bij een klasse horen, en of deze wel of niet als dusdanig zijn geclassificeerd [11] .

	Behorend tot partij	Niet behorend tot partij
Geclassificeerd als partij	<i>true positive (tp)</i>	<i>false positive (fp)</i>
Niet geclassificeerd 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 geclassificeerd is. *Precision* is het percentage van documenten geclassificeerd als klasse, dat ook bij die klasse hoort. *Recall* is het percentage documenten van documenten behorende tot een klasse, dat ook als dusdanig geclassificeerd is. F_1 is het harmonisch gemiddelde van recall en precision. Precision, recall en dus ook F_1 worden per klasse berekend. Er zijn drie varianten om deze scores voor de hele classificatie te berekenen.

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

Als tweede is er *macro*, daarbij worden alle scores per klasse berekend en wordt daarvan het gemiddelde genomen. Dit leidt er dan weer toe dat resultaten

354 van klassen met weinig documenten net zo belangrijk zijn. Hierdoor kan een
355 classificatie met een laag aantal correct geclassificeerde documenten hoog scoren
356 door vooral kleine klassen goed te classificeren.

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

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

364 3.2.2 DV2: Invloed van namen

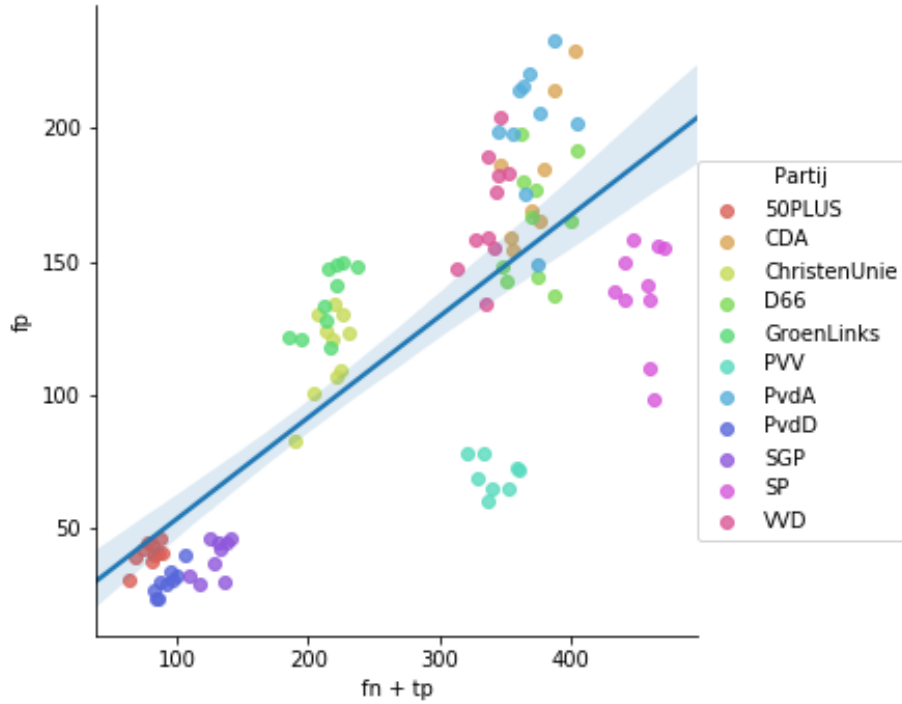
365 In Diermeier et al. [4] wordt aangenomen dat namen een groot effect hebben op
366 de classificatie en Hirst et al. [2] bevestigen dit voor het Europees Parlement.
367 Aangezien hier bij deelvraag 1 niet voor is gekozen, wordt bij deze deelvraag
368 gekeken hoe groot het effect hiervan is, specifiek gericht op partijnamen en ach-
369 ternamen van Kamerleden. Voor deze deelvraag wordt wederom een classificatie
370 gedaan met de classificatiemethode die resulteerde uit deelvraag 1. In deze clas-
371 sificatie worden alle partijnamen vervangen door de tag PARTIJNAAM en alle
372 namen van Kamerleden vervangen door de KAMERLIDNAAM. Deze namen
373 zijn uit de Handelingen gehaald. Voor partijnamen zijn ook lidwoorden toege-
374 voegd, voor achternamen van Kamerleden zijn ook verkortingen meegenomen.
375 Dit laatste omdat bijvoorbeeld *Van Haersma Buma* vaak aangesproken wordt
376 als *Buma*. Voornamen van Kamerleden worden zelden tot nooit gebruikt, dus
377 die zijn er niet uitgehaald. Een nadeel van deze aanpak is dat ook namen van
378 niet-Kamerleden of andere woorden weggehaald kunnen worden als deze het-
379 zelfde zijn als naam van een Kamerlid. Door gebruik van gevoeligheid voor
380 hoofdletters is geprobeerd dit te voorkomen. Een opvallend voorbeeld hiervan
381 is de naam Rutte, die zowel behoort tot het Kamerlid Arno Rutte als de premier
382 Mark Rutte. Steekproefgewijs is gekeken of er nog namen achter zijn gebleven,
383 maar die zijn niet gevonden.

384 Ook wordt gekeken naar classificatie met alleen partijnamen en namen van
385 Kamerleden. Alle andere woorden worden weggehaald. Namen van Kamerleden
386 en partijen die niet aan elkaar geschreven worden, zoals *Partij van de Arbeid*,
387 worden aan elkaar geschreven zodat het één feature wordt. Doordat alle andere
388 woorden weggehaald zijn, worden de bi- en trigrams combinaties van namen
389 die zinnen uit elkaar kunnen staan, dus die niet meer informatie geven dan
390 unigrams. Daarom wordt er gebruikt van de classificatiemethode uit deelvraag
391 1, maar dan met alleen unigrams. Hoge scores voor deze classificatie geven
392 aan dat met alleen namen classificatie goed te doen is en dat dit dus een grote
393 bijdrage heeft geleverd aan de resultaten uit deelvraag 1.

394 3.2.3 DV3: Oppositie of regering

395 Om deze deelvraag te beantwoorden zal een analyse gedaan worden van de
396 confusion matrix en zullen twee experimenten die Graeme Hirst et al. uitvoerden
397 voor dezelfde vraag gereproduceerd worden op de dataset van de Tweede Kamer.
398 Bij deze deelvraag zal de beste classifier uit deelvraag 1 en 2 gebruikt worden.

399 Als er een confounding bias is op basis van partij-status, dan is te ver-
 400 wachten dat het aantal misclassificaties minus verwachte waarde binnen rege-
 401 ringspartijen en binnen oppositiepartijen hoger ligt dan tussen oppositiepartijen
 402 en regeringspartijen. De verwachte waarde is afhankelijk van het aantal docu-
 403 menten van een partij in de training set [12]. Aangezien de test set uit dezelfde
 404 set als de training is gehaald, is de verwachte waarde ook afhankelijk van het
 405 aantal documenten van een partij in de test set. Uit de voorverkenning (op basis
 406 van resultaten uit deelvraag 1 en 2) blijkt deze correlatie tussen het aantal *false*
 407 *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 50 classificaties met verschillende test en train set. De pearson correlatie is 0.78.

408 Op basis van dit verband is het verwachte aantal documenten

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

409 waar $i \neq j$ met j de voorspelde partij en i de echte partij waar een document
 410 bijhoort.

411 De error is dan het verschil van de verwachte waarde en het daadwerkelijk
 412 aantal documenten

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

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

Als dit een goede benadering is van de error, dan is het te verwachten dat deze normaal verdeeld is [13]. Om te kijken of er een bias is, worden de distributies binnen regeringspartijen, binnen oppositiepartijen en tussen beide groepen met elkaar vergeleken. Om de invloed van variantie door de willekeurige splitsing documenten voor trainen en testen te beperken, wordt de classificatie 50 keer gedaan en worden deze errors bij elkaar in distributie genomen. De nulhypothese is dat er geen verschil is tussen de verdelingen. De alternatieve hypothese is dan dus dat er wel een verschil is tussen de verdelingen. Als de nulhypothese wordt verworpen, kan dus aangenomen worden dat er een verschil is op basis van partij-status.

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

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

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

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

3.2.4 DV4: Links of rechts

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[14] gebaseerd op verkiezingsprogramma's voor de Kamerverkiezing van 2012. In beide gevallen is de nullijn van het politieke spectrum gebruikt om te bepalen of een partij links of rechts is.

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

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

3.2.5 DV5: Woordgebruik van sprekers

De vorige classificaties trainden op documenten en werden getest op andere documenten, maar wel van dezelfde sprekers als uit de training set. Naast de ideologie kan de classificatie daarom ook getraind zijn op het taalgebruik van sprekers. Als een Kamerlid bijvoorbeeld een woord regelmatig in speeches gebruikt, maar niet wordt gebruikt door zijn partijgenoten, wordt dit wel gezien als een belangrijk woord voor de partijclassificatie. Graeme Hirst et al. [2] plaatsten al een soortgelijke kanttekening bij de resultaten van Deiermeier et al.

Om te kijken of dit effect er is, wordt er opnieuw een classificatie gedaan. Ditmaal worden alleen niet de individuele documenten verdeeld over de training en test set, maar worden de Kamerleden, met bijbehorende documenten, verdeeld over de training en test set. Als taalgebruik van een spreker in de training set voorheen invloed had op de classificatie, zal dat nu geen effect meer hebben omdat er geen documenten van die spreker meer voorkomen in de test set. De meest karakteristieke woorden uit de resultaten van deelvraag 2 suggereren dat woordgebruik van Kamerleden invloed heeft (zie tabel 4). De hypothese is daarom ook dat deze nieuwe classificatie lagere scores vindt.

4 Evaluatie

4.1 Resultaten

4.1.1 DV1: Beste classificatiemethode

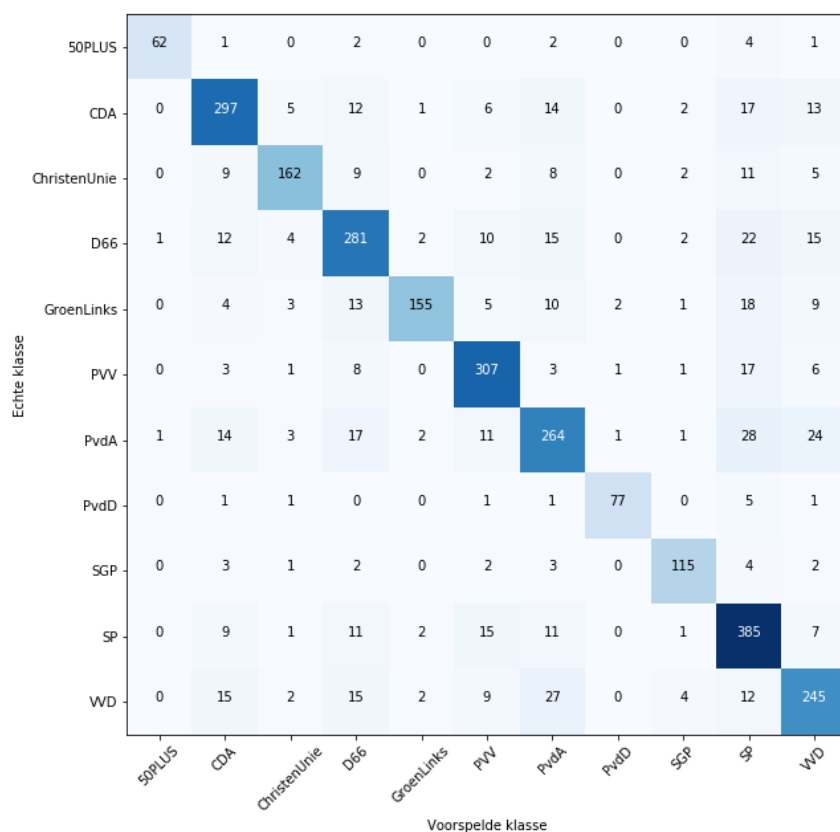
Het beste resultaat werd bereikt met Support Vector Machines gebruikmakend van *stochastic gradient descent learning* en Elasticnet regularisatie. De features waren hierbij gestemd, met unigrams, bigrams en trigrams. Geen features zijn hierin weggelaten door minimale of maximale documentfrequenties.

Tabel 3 laat de scores zien per partij met het aantal documenten in de test set. De F_1 scores per partij liggen tussen de 0.7 en 0.8. De one-issuepartijen, 50PLUS en PvdD, hebben scores daarboven, terwijl de coalitiepartijen, VVD en PvdA, lagere scores hebben. Figuur 3 laat zien waar de fouten in deze classificatie zitten. De meest karakteristieke features per partij zijn te zien in

489 tabel 4. Hierin is te zien dat vrijwel alle woorden verwijzen naar de partij of
 490 een Kamerlid van die partij.

Tabel 3: Classificatierapport van beste classificatiemethode. Gemiddelde van vijf iteraties.

	Precision	Recall	F1_score	Accuracy	Documenten
50PLUS	0.95	0.82	0.89	-	76
CDA	0.80	0.80	0.80	-	372
ChristenUnie	0.87	0.77	0.81	-	212
D66	0.75	0.76	0.76	-	368
GroenLinks	0.92	0.70	0.79	-	222
PVV	0.82	0.88	0.85	-	351
PvdA	0.73	0.72	0.72	-	371
PvdD	0.94	0.84	0.88	-	92
SGP	0.87	0.85	0.86	-	136
SP	0.73	0.86	0.79	-	447
VVD	0.74	0.74	0.74	-	334
Totaal	0.80	0.79	0.79	0.79	2980



Figuur 3: Confusion matrix van beste classificatie.

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

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

Tabel 4: Meest karakteristieke 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	lid ouwehand naar	de sgp	de sp	vvd
islamitische	van de arbeid	het lid ouwehand	sgp fractie	sp fractie	de vvd fractie
klever	de partij van	ouwehand	de sgp fractie	de sp fractie	vvd fractie
nederland	de arbeid	ouwehand naar	leden dijkgraaf	lid van gerven	de vvd is
miljard	partij van de	ouwehand naar mij	de leden dijkgraaf	gerven	vvd is
pvv fractie	arbeid	voor de dieren	leden van der	van gerven	voor de vvd
graus	partij van	thieme	de leden bisschop	smaling	wat de vvd
de pvv fractie	pvda fractie	dieren	leden bisschop	gerven naar	de vvd betreft
graaf	de pvda fractie	de dieren	mevrouw de voorzitter	gerven naar mij	vvd betreft

4.1.2 DV2: Invloed van namen

In tabel 4 was al te zien dat de meest karakteristieke woorden voornamelijk bestaan uit partijnamen en namen van Kamerleden. In tabel 5 zijn de scores te zien van classificatie met partijnamen en namen van Kamerleden vervangen. Deze zijn aanzienlijk lager dan de scores uit deelvraag 1. In tabel 6 is vervolgens te zien welke woorden het meest karakteristiek zijn per partij voor deze classificatie.

Tabel 5: Classificatierapport van beste classificatie zonder namen van Kamerleden of partijnamen.

	Precision	Recall	F1_score	Accuracy	Documenten
50PLUS	0.98	0.84	0.90	-	80
CDA	0.81	0.77	0.79	-	380
ChristenUnie	0.87	0.76	0.81	-	219
D66	0.77	0.75	0.76	-	389
GroenLinks	0.91	0.72	0.80	-	220
PVV	0.82	0.89	0.85	-	331
PvdA	0.70	0.71	0.71	-	361
PvdD	0.92	0.86	0.88	-	84
SGP	0.86	0.86	0.86	-	129
SP	0.74	0.85	0.79	-	458
VVD	0.73	0.76	0.74	-	330
Totaal	0.79	0.79	0.79	0.79	2980

Tabel 6: 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
50plus	cda	de christenunie	d66	groenlinks
lid krol	het cda	christenunie	leden van veldhoven	lid van tongeren
lid krol naar	cda fractie	lid dik	mijn fractie	lid voortman naar
het lid krol	de cda	het lid dik	lid van veldhoven	het lid voortman
van 50plus	de cda fractie	lid dik faber	van veldhoven	lid voortman
krol naar mij	het lid omtzigt	leden voordewind	veldhoven	van tongeren
krol naar	lid omtzigt	de leden voordewind	lid van meenen	tongeren
gepensioneerd	lid omtzigt naar	dik faber	d66 is	voortman
krol	omtzigt naar mij	faber	leden schouw	leden van tongeren
ouderen	omtzigt naar	dik	de leden schouw	de leden voortman

Tabel 6: 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
pvv	de pvda	lid ouwehand	sgp	sp	de vvd
de pvv	pvda	lid ouwehand naar	de sgp	de sp	vvd
islamitische	van de arbeid	het lid ouwehand	sgp fractie	sp fractie	de vvd fractie
miljard	de arbeid	ouwehand	de sgp fractie	de sp fractie	vvd fractie
klever	de partij van	ouwehand naar mij	leden van der	van gerven	de vvd is
nederland	partij van de	ouwehand naar	leden dijkgraaf	gerven	vvd is
graaf	partij van	dieren	de leden dijkgraaf	lid van gerven	voor de vvd
graus	arbeid	voor de dieren	mevrouw de voorzitter	smaling	wat de vvd
natuurlijk	pvda fractie	de dieren	mevrouw de	gerven naar mij	de vvd betr
madlener naar mij	de pvda fractie	thieme	de leden bisschop	gerven naar	vvd betreft

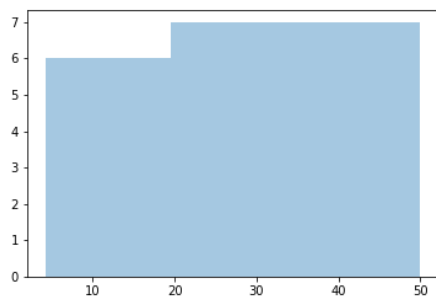
498 In tabel 7 zijn de scores te zien voor een classificatie met alleen namen van
499 partijen Kamerleden. De scores zijn gedaald ten op zichte van de resultaten van
500 deelvraag 1, maar hoger dan die zonder namen.

Tabel 7: Classificatierapport van beste classificatie met alleen namen van partijen en Kamerleden. Hiervoor is alleen gebruikgemaakt van unigrams.

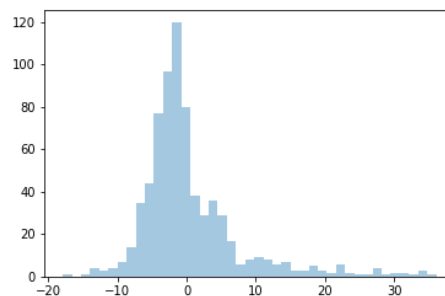
	Precision	Recall	F1_score	Accuracy	Documenten
50PLUS	0.82	0.84	0.83	-	81.0
CDA	0.66	0.64	0.65	-	367.0
ChristenUnie	0.63	0.59	0.61	-	217.0
D66	0.53	0.54	0.52	-	390.0
GroenLinks	0.69	0.66	0.67	-	211.0
PVV	0.65	0.71	0.67	-	345.0
PvdA	0.63	0.50	0.56	-	351.0
PvdD	0.76	0.74	0.75	-	81.0
SGP	0.66	0.53	0.58	-	128.0
SP	0.60	0.64	0.61	-	470.0
VVD	0.64	0.57	0.60	-	339.0
Totaal	0.63	0.61	0.62	0.61	2980.0

501 4.1.3 DV3: Oppositie of regering

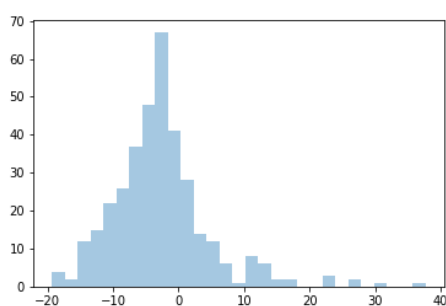
502 In figuur 4 zijn de distributies van de errors te zijn van combinaties tussen
503 regerings- en oppositiepartijen.



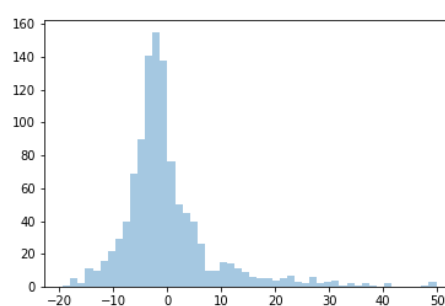
(a) Tussen twee regeringspartijen



(b) Tussen twee oppositiepartijen



(c) Tussen een regeringspartij en een oppositiepartij



(d) Totaal

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

Tabel 8: Meest karakteristieke woorden 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	burger
wij	de fractie van	de premier	premier	onze
fractie	de fractie	ik hoop	fractie van PARTIJ	door
dank	fractie van	hij	de fractie	burgers
wij hebben	van PARTIJ	arbeidsmarkt	de fractie van	natuurlijk
aangegeven	blij	hoop	ik	politie
hebben	beantwoording	de arbeidsmarkt	fractie van	immigratie
KAMERLID	volgens mij	schone energie	mij	land
zorgvuldige	moment	patiënt	politieke	niet
onze	over	plannen	en	gewoon

Tabel 8: Meest karakteristieke woorden 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	dierenwelzijn	beantwoording	mensen	PARTIJ fractie
belangrijk	bio industrie	wel	is	onze fractie
punt	de bio industrie	de bewindslieden	niet	fractie
roc	de bio	bewindslieden	nu	dat
medewerkers	natuur	toch	voorstel	je
ik	bio	voor zijn beantwoording	militaire	ondernemers
van de arbeid	de dieren	natuurlijk	de staatssecretaris	markt
antwoorden	dierproeven	de voorzitter	personeel	awbz
ben	veehouderij	enkele	er	want

Tabel 9: Classificatierapport van beste classificatie met de Kamerleden verdeeld over training en test set.

	Precision	Recall	F1_score	Accuracy	Documenten
50PLUS	0.340	0.180	0.123	0.013293	108.3
CDA	0.128	0.151	0.124	NaN	343.6
ChristenUnie	0.064	0.046	0.046	NaN	145.4
D66	0.253	0.204	0.198	NaN	481.2
GroenLinks	0.108	0.034	0.049	NaN	184.6
PVV	0.389	0.430	0.386	NaN	407.4
PvdA	0.190	0.199	0.179	NaN	367.4
PvdD	0.400	0.095	0.115	NaN	112.5
SGP	0.235	0.048	0.080	0.036673	244.4
SP	0.224	0.385	0.248	NaN	402.5
VVD	0.235	0.249	0.206	NaN	348.9
Totaal	0.300	0.206	0.205	0.205660	3146.2

Tabel 10: Classificatierapport van beste classificatie met de Kamerleden verdeeld over training en test set.

	Precision	Recall	F1_score	Accuracy	Documenten
50PLUS	0.340	0.180	0.123	0.013293	108.3
CDA	0.128	0.151	0.124	NaN	343.6
ChristenUnie	0.064	0.046	0.046	NaN	145.4
D66	0.253	0.204	0.198	NaN	481.2
GroenLinks	0.108	0.034	0.049	NaN	184.6
PVV	0.389	0.430	0.386	NaN	407.4
PvdA	0.190	0.199	0.179	NaN	367.4
PvdD	0.400	0.095	0.115	NaN	112.5
SGP	0.235	0.048	0.080	0.036673	244.4
SP	0.224	0.385	0.248	NaN	402.5
VVD	0.235	0.249	0.206	NaN	348.9
Totaal	0.300	0.206	0.205	0.205660	3146.2

504 4.1.4 DV4: Links of rechts

505 4.1.5 DV5: Woordgebruik van sprekers

506 In tabel 11 staan de scores van classificatie waarbij de Kamerleden verdeeld zijn
507 over de training en test set. De scores zijn hierbij amper hoger dan de baseline.

Tabel 11: Classificatierapport van beste classificatie met de Kamerleden verdeeld over training en test set.

	Precision	Recall	F1_score	Accuracy	Documenten
50PLUS	0.340	0.180	0.123	0.013293	108.3
CDA	0.128	0.151	0.124	NaN	343.6
ChristenUnie	0.064	0.046	0.046	NaN	145.4
D66	0.253	0.204	0.198	NaN	481.2
GroenLinks	0.108	0.034	0.049	NaN	184.6
PVV	0.389	0.430	0.386	NaN	407.4
PvdA	0.190	0.199	0.179	NaN	367.4
PvdD	0.400	0.095	0.115	NaN	112.5
SGP	0.235	0.048	0.080	0.036673	244.4
SP	0.224	0.385	0.248	NaN	402.5
VVD	0.235	0.249	0.206	NaN	348.9
Totaal	0.300	0.206	0.205	0.205660	3146.2

508 4.2 Discussie

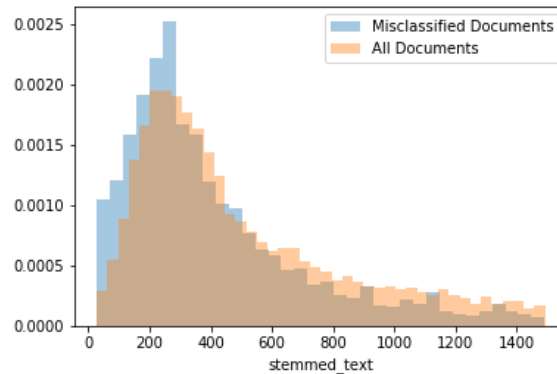
509 4.2.1 DV1: Beste classificatiemethode

510 Het onderzoek behaalt resultaten in lijn der verwachting op basis van gere-
511 lateerd en daarnaast ruim boven de baseline scores. De lage scores voor de
512 coalitiepartijen steunen de hypothese van een afhankelijkheid van partij-status,

513 zoals besproken wordt in deelvraag 3 Het bijna alleen voorkomen van partijna-
 514 men en Kamerleden in de meest karakteristieke woorden per partij in tabel 4
 515 steunt daarnaast het vermoeden dat deze classificatie sterk afhankelijk is van
 516 die namen, zoals besproken wordt in deelvraag 2.

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

526 Het onderzoek van Hirst et al. vond dat resultaten afhankelijk kunnen
 527 zijn van documentgrootte. Alle documenten in dit onderzoek zijn kleiner dan
 528 de grootste documentgrootte uit het onderzoek van Hirst et al. en ook de mi-
 529 nimumfrequentie lager ligt dan de kleinste documentgrootte uit dat onderzoek.
 530 Het effect wat zij vinden tussen documentgrootte van 267 en 6666 is een ver-
 531 schil in nauwkeurigheid van 19,8%. Dit onderzoek vindt inderdaad dat kleinere
 532 documenten vaker foutief geclassificeerd worden.



Figuur 5: Genormaliseerde distributie van documentlengtes van foutief geclassificeerde documenten en alle documenten. Totaal van 5-fold cross-validation, waardoor documenten vaker voor kunnen komen.

533 Voor een vervolgonderzoek kan uitgebreider gekeken worden naar dit effect
 534 en wat dit betekent voor de resultaten.

535 4.2.2 DV2: Invloed van namen

536 De resultaten laten zien dat de classificatie sterk afhankelijk is van partijna-
 537 men en namen van Kamerleden. Deze daling was te verwachten op basis van
 538 gerelateerd werk.

539 De woorden in tabel 6 komen bij veel partijen overeen met hun ideologie,
 540 vooral bij PVV, PvdD en 50PLUS. Daarnaast zijn er ook woorden die niet veel
 541 over ideologie zeggen, zoals; *volgens mij*, *ik constateer* en *in elk geval*. Vooral de

542 SGP heeft woorden die niet veel lijken te zeggen over de ideologie, hoewel deze
543 partij desalniettemin de hoogste f_1 score heeft. Met name opvallend hierbij
544 is *mevrouw de voorzitter*, aangezien deze woorden door alle partijen gebruikt
545 worden om via de voorzitter te praten. Voor een vervolgonderzoek kan gekeken
546 naar waarom deze woorden zo karakteristiek zijn voor partijen. Een hypothese
547 is dat deze woorden eigen zijn aan een individueel Kamerlid.

548 De classificatiemethode die gebruikt is in deze deelvraag, is gebaseerd op
549 de beste methode voor de dataset uit deelvraag 1. Hierin was gevonden dat een
550 combinatie van uni-, bi- en trigrams het beste resultaat opleverde. In tabel 4 is
551 te zien dat trigrams behoren tot de meest karakteristieke woorden, hoewel de
552 woorden in trigrams vaak overlappen met uni- en bigrams. In tabel 6 daaren-
553 tegen zijn er nog maar een paar trigrams, welke grotendeels procedurele zinnen
554 zijn of toevoeging van een lidwoord op een uni- of bigram. Dit verschil sugge-
555 reert dat trigrams minder belangrijk zijn in de classificatie zonder de namen,
556 dus de classificatiemethode uit deelvraag 1 niet het beste is voor deze classifica-
557 tie. In vervolgonderzoek kan de opzet van deelvraag 1 toegepast worden op de
558 classificatie zonder de namen, om zo te komen tot een classificatiemethode die
559 het beste resultaat oplevert op de classificatie zonder namen.

560 4.2.3 DV3: Oppositie of regering

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

563 4.2.4 DV4: Links of rechts

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

566 4.2.5 DV5: Woordgebruik van sprekers

567 De resultaten uit tabel 11 zijn laag, amper hoger dan de baseline. Dit suggereert
568 inderdaad dat eerdere classificaties in grote mate toch afhankelijk waren van het
569 woordgebruik van sprekers. Dit is opmerkelijk aangezien vergelijkbare werken
570 dit effect niet vinden.

571 Een alternatieve verklaring is dat de classificatie nu mede op basis van
572 woordvoerderschap is. Per onderwerp heeft een partij vaak maar één woord-
573 voerder, met uitzonderingen van wijzigingen in de fractie. Het is aannemelijk
574 dat het taalgebruik afhankelijk is van woordvoerderschap, aangezien er andere
575 termen gebruikt worden bij bijvoorbeeld een debat over zorg dan bij een debat
576 over onderwijs. Stel dat documenten van een spreker in de test set geclassifi-
577 ceerd moeten worden, dan kan het zijn dat deze meer karakteristieke vertoont
578 met een andere partij, aangezien er geen woordvoerder van die partij en dat
579 onderwerp in de training set zit, maar mogelijk wel van een andere partij. Een
580 vervolgonderzoek kan kijken of dit een verklaring is.

581 4.2.6 Algemeen

582 Het vergelijken van deze resultaten met vergelijkbaar werk is problematisch,
583 aangezien de keuzes en eigenschappen van hun onderzoek het niet een één-op-
584 één vergelijking maken. Voorbeelden hiervan zijn de documentgrootte, baseli-

nes, behouden of weglaten van namen, een spreker als document zien en het trainen en testen op dezelfde spreker. Hoewel de resultaten dus lager zijn dan die uit vergelijkbaar werk, moet hiermee rekening gehouden worden. 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.

5 Conclusies

Dit onderzoek vindt een *accuracy* van XXX en een F_1 score van XXX 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 namen van partijen en Kamerleden, daalt de *accuracy* naar XXX en de F_1 score. 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? *.
- [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] Conal Sathi Maneesh Bhand, Dan Robinson. Text classifiers for political ideologies, 2009.
- [7] Bjørn Høyland, Jean-François Godbout, Emanuele Lapponi, and Erik Vell-dal. 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.

- 625 [8] Laura Klompenhouwer. Extra ledenvergadering 50plus om splitsing. *NRC*
626 *Handelsblad*, June 2014.
- 627 [9] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source
628 scientific tools for Python, 2001–. [Online; accessed `today`].
- 629 [10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Gri-
630 sel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas,
631 A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay.
632 Scikit-learn: Machine learning in Python. *Journal of Machine Learning*
633 *Research*, 12:2825–2830, 2011.
- 634 [11] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Intro-*
635 *duction to Information Retrieval*. Cambridge University Press, New York,
636 NY, USA, 2008.
- 637 [12] Mahendra Sahare and Hitesh Gupta. A review of multi-class classifica-
638 tion for imbalanced data. *International Journal of Advanced Computer*
639 *Research*, 2(3), 2012.
- 640 [13] *NIST/SEMATECH e-Handbook of Statistical Methods*. NIST/SEMA-
641 TECH, April 2012.
- 642 [14] Andrea Volkens, Pola Lehmann, Theres Matthieß, Nicolas Merz, Sven Re-
643 gel, and Bernhard Weßels. The manifesto data collection. manifesto project
644 (mrg/cmp/marpor). version 2017b, 2017.

645 A Slides