Ideologie en classificatie in de Handelingen van de Tweede Kamer

submitted in partial fulfillment for the degree of bachelor of science

Jasper van der Heide
10732721

Bachelor Informatiekunde
Faculteit der Natuurwetenschappen, Wiskunde en Informatica
Universiteit van Amsterdam

2018-06-28

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

Universiteit van Amsterdam

Inhoudsopgave

Samenvatting

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 namen van partijen en 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

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

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

In dit hoofdstuk zullen verschillende onderzoeken behandeld worden die tekstclassificatie hebben toegepast op parlementaire teksten. Eerst zullen de onderzoeken algemeen besproken worden. Vervolgens zal uitgebreider gekeken worden naar de effecten van verschillende classificatiemethoden. In de latere secties zullen aspecten besproken worden die in vergelijkbare onderzoeken genoemd worden als van invloed op de classifcatie.

2.1 Tekstclassificatie van parlementaire teksten

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 categorieën van het 108e Congres. Een document was in dit onderzoek de verzameling van alle speeches van een senator in een congres. Deze classificatie resulteerde uiteindelijk in een nauwkeurigheid van 94% (baseline van 50%). Van de 50 senatoren in de test set, kwamen er 44 al voor in de training set, doordat de training op voorgaande congressen was.

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

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

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

Als een vervolg op deze onderzoeken deden 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 *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 parlementslid 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 leden van het Amerikaanse congres in 2005, op basis van affiliatie (Republikeins of Democratisch)[6]. Een document hierbij was in tegenstelling tot eerdergenoemde 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 unigrams, met minimale woordfrequentie van drie en de top 50 meest voorkomende woorden weggelaten. Voor de wegingen van de features bij Support Vector Machines werd geëxperimenteerd met *boolean*, *tf-norm* en *tf-idf*. Het beste resultaat was afhankelijk van welke kamer Voor het huis van afgevaardigden was het Support Vector Machines met als weging *tf-idf* en voor de senaat Bernouilli Naive Bayes.

Hirst et al. maakten 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 wegingen 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 manieren 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

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.

3 Methodologie

3.1 De data

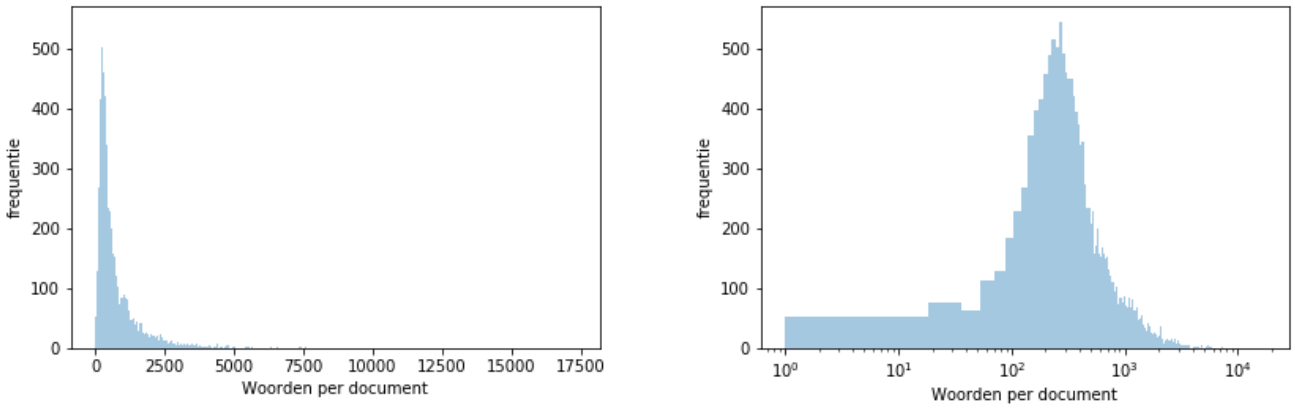
De data die gebruikt worden, zijn de Handelingen van de Tweede Kamer gedurende het missionaire kabinet-Rutte II (5 november 2012 tot 22 maart 2017). Er is gekozen voor dit kabinet, omdat de data hiervoor makkelijk verkrijgbaar was, het kabinet lang zat - waardoor er veel data is - en het recent is waardoor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

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

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

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

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



(a) Normale x-as

(b) Logaritmische x-as

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
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).

3.2 Methoden

3.2.1 DV1: Beste classificatiemethode

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

Pre-processing Voor pre-processing is gebruik gemaakt van tokenisation en lowercasing. Voor tokenisation is de reguliere expressie
$w+$ gebruikt, waardoor allesbehalve letters en cijfers weggehaald wordt. Vervolgens is er gevarieerd tussen wel of geen gebruik maken van stemming. In het geval van stemming is gebruik gemaakt van de Snowball Stemmer via de Python NLTK module.

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

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

Support Vector Machines en Logistische Regressie De meest voorkomende techniek in vergelijkbaar onderzoek is Support Vector Machine (SVM). Een andere techniek die gebruikt wordt is logistische regressie. Beide kennen een eigen implementatie in scikit-learn, maar deze implementaties zijn niet efficiënt met grote datasets. Om deze reden is er in beide gevallen voor gekozen om gebruik te maken van de functie SGDClassifier, die beide technieken leert met *stochastic gradient descent learning*. Voor regularisatie is hier geëxperimenteerd 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, 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 partij horen, en of deze wel of niet als dusdanig zijn geclassificeerd [11] .

	Behorend tot partij	Niet behorend tot partij
Geclassificeerd als partij	*true positive (tp)*	*false positive (fp)*
Niet geclassificeerd als partij	*false negative (fn)*	*true negative (tn)*

$$Precision = \frac{tp}{tp + fp} \tag{1}$$

$$Recall = \frac{tp}{tp + tn} \tag{2}$$

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{3}$$

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4}$$

Accuracy is het percentage van documenten dat correct geclassificeerd is. Precision is het percentage van documenten geclassificeerd als partij, dat ook bij die partij hoort. Recall is het percentage documenten van documenten behorende tot een partij, dat ook als die partij geclassificeerd is. F_1 is het harmonisch gemiddelde van recall en precision. Precision, recall en dus ook F_1 worden per partji 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 partijen met veel documenten belangrijker zijn. Als een classificatie kleine partijen grotendeels fout classificeert, kan deze score alsnog hoog zijn. In het geval van meer dan twee partijen is dit hetzelfde als *accuracy*.

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

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

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

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

3.2.2 DV2: Invloed van namen

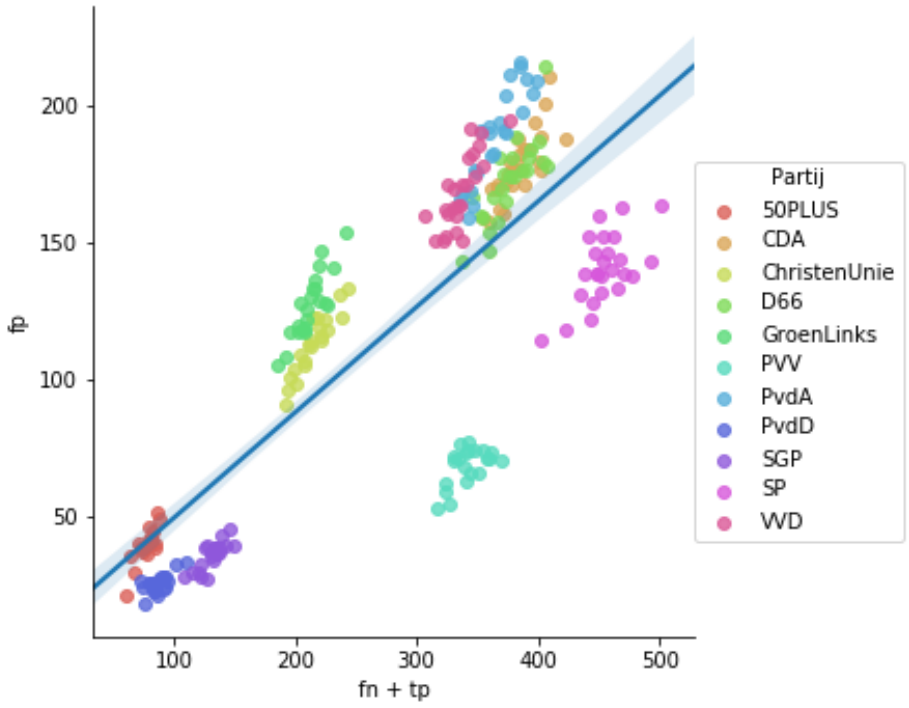
In Diermeier et al. [4] wordt aangenomen dat namen een groot effect hebben op de classificatie en Hirst et al. [2] bevestigen dit voor het Europees Parlement. Aangezien hier bij deelvraag 1 niet voor is gekozen, wordt bij deze deelvraag gekeken hoe groot het effect hiervan is, specifiek gericht op partijnamen 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.

Ook wordt gekeken naar classificatie met alleen partijnamen en namen van Kamerleden. Alle andere woorden worden weggehaald. Namen van Kamerleden en partijen die niet aan elkaar geschreven worden, zoals *Partij van de Arbeid*, worden aan elkaar geschreven zodat het één feature wordt. Doordat alle andere woorden weggehaald zijn, worden de bi- en trigrams combinaties van namen die zinnen uit elkaar kunnen staan, dus die niet meer informatie geven dan unigrams. Daarom wordt er gebruikt van de classificatiemethode uit deelvraag 1, maar dan met alleen unigrams. Hoge scores voor deze classificatie geven aan dat met alleen namen classificatie goed te doen is en dat dit dus een 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 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. 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 50 classificaties met verschillende test en train set. De pearson correlatie is 0.78.

Op basis van dit verband is het verwachte aantal documenten

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

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

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

$$e_{i,j} = D_{i,j} - V_{i,j} \tag{6}$$

met opnieuw $i \neq j$ en i de voorspelde partij en j de echte partij waar een 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 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 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. 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 met de methode uit deelvraag 2. 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

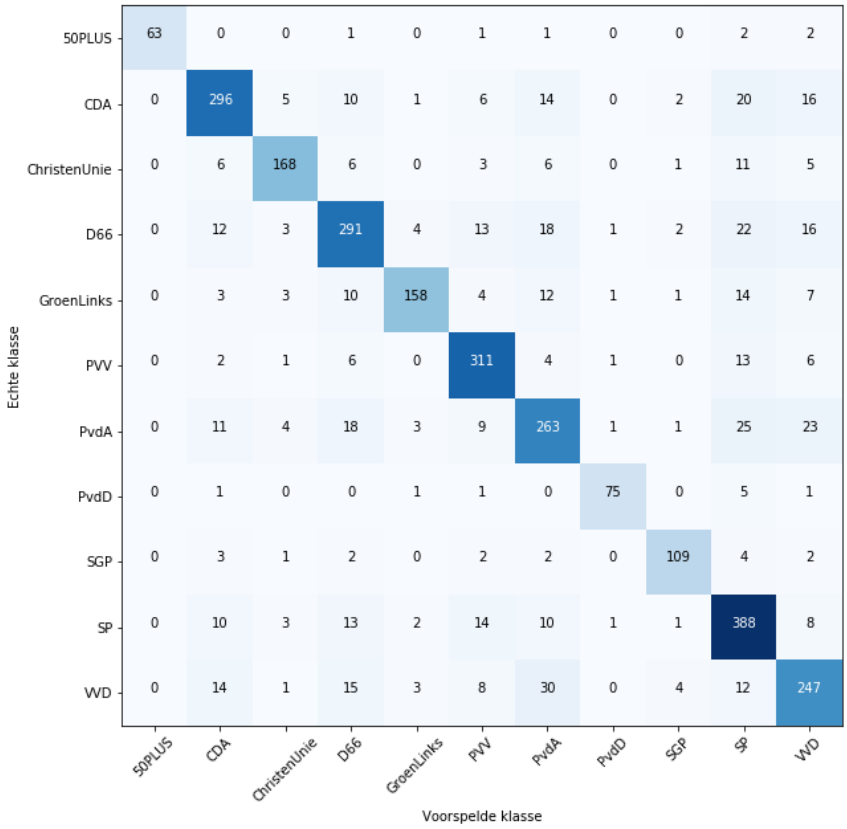
Het beste resultaat werd bereikt met Support Vector Machines gebruikmakend van *stochastic gradient descent learning* en Elasticnet regularisatie. De features waren hierbij gestemd, met unigrams, bigrams en trigrams. Geen features zijn hierin weggelaten door minimale of maximale documentfrequenties.

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 0.9. De one-issuepartijen, 50PLUS, PVV en PvdD, als ook de SGP hebben hoge scores, 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 tabel 4. Met meest karakteristiek worden de n-grams bedoeld die de hoogste coëfficiënt hebben in de classificatie en die dus relatief het meeste invloed hebben. Hierin is te zien dat vrijwel alle n-grams namen van de partijen of Kamerleden bevatten.

Tabel 3: 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 classificatie. Gemiddelde van vijf splitsingen van training en test set. Maximum aantal iteraties is 100.

Tabel 4: Meest karakteristieke n-grams per partij op basis van beste classificatie gedurende kabinet-Rutte II. N-grams die niet namen van partijen of Kamerleden bevatten, zijn dikgedrukt. Maximum aantal iteraties is 100.

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 4: Meest karakteristieke n-grams per partij op basis van beste classificatie gedurende kabinet-Rutte II. N-grams die niet verwijzen naar partijen of Kamerleden zijn dikgedrukt. Maximum aantal iteraties is 100. *(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

4.1.2 DV2: Invloed van namen

In tabel 4 was al te zien dat de meest karakteristieke n-grams voornamelijk namen van partijen en Kamerleden bevatten. 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 n-grams het meest karakteristiek zijn per partij voor deze classificatie.

Tabel 5: Classificatie scores per partij van beste classificatie zonder namen van Kamerleden of partijnamen. Gemiddelde van vijf splitsingen van training en test set.

	Precision	Recall	F_1 score	Documenten
PvdD	0.75	0.70	0.72	86
SGP	0.71	0.73	0.72	123
PVV	0.63	0.80	0.70	350
50PLUS	0.86	0.49	0.62	76
SP	0.54	0.71	0.61	454
ChristenUnie	0.68	0.46	0.55	214
D66	0.55	0.55	0.55	385
CDA	0.52	0.53	0.52	372
VVD	0.54	0.49	0.52	340
PvdA	0.51	0.48	0.50	366
GroenLinks	0.64	0.38	0.48	213
Totaal	0.59	0.58	0.57	2980

Tabel 6: Meest karakteristieke n-grams 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	PARTIJ fractie	mensenhandel	mijn fractie	zou
ouderen	inwoners	zullen	mijn	kamer hierover te
oudere	PARTIJ	gezinnen	fractie	belastingontwijking
koopkrachtontwikkeling	regering	inderdaad	natuurlijk	in elk geval
plussers	wij	vluchtelingen	het kabinet	persoonsgebonden
50	de regering	kinderen	belangrijk	elk geval
werkenden	hier	hoop	vandaag	in elk
50 plussers	echt	motie	kansen	hierover te informeren
voor gepensioneerden	fractie	onder meer	kabinet	schone energie
overwegende	de	constateer	buitengewoon	hierover te

Tabel 6: Meest relevante n-grams 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	en	dieren	mevrouw de voorzitter	zegt	PARTIJ
nederland	toezeggingen	de bio	mevrouw de	mening dat	speelveld
miljard	circulaire economie	bio	punt	van mening dat	volgens mij
natuurlijk	circulaire	bio industrie	beantwoording	de bevolking	aangegeven
brussel	jongeren	industrie	bewindslieden	bevolking	banen
islam	mijn partij	de bio industrie	voor de beantwoording	mensen	PARTIJ fractie
al	lagere overheden	milieu	de beantwoording	bezuinigingen	regelgeving
miljarden	tevens	natuur	wel	huurders	volgens
de islam	vragen	constaterende	je	bestuurders	aruba
asielzoekers	wij	constaterende dat	de voorzitter ik	van mening	PARTIJ is

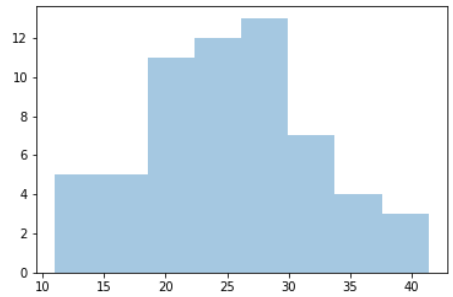
In tabel 7 zijn de scores te zien voor een classificatie met alleen namen van partijen Kamerleden. De scores zijn gedaald ten op zichte van de resultaten van deelvraag 1, maar hoger dan die zonder namen.

Tabel 7: Classificatierapport van beste classificatie met alleen namen van partijen en Kamerleden. Hiervoor is alleen gebruikgemaakt van unigrams. Gemiddelde van vijf splitsingen van training en test set.

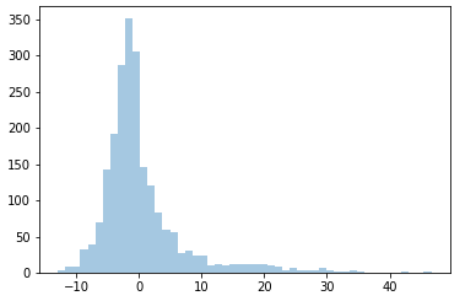
	Precision	Recall	F_1 score
50PLUS	0.82	0.88	0.85
PvdD	0.68	0.78	0.69
GroenLinks	0.71	0.66	0.68
PVV	0.66	0.71	0.67
CDA	0.67	0.65	0.66
ChristenUnie	0.66	0.58	0.62
SP	0.61	0.64	0.62
VVD	0.68	0.57	0.62
SGP	0.69	0.54	0.60
D66	0.56	0.53	0.54
PvdA	0.56	0.51	0.52
Totaal	0.64	0.62	0.62

4.1.3 DV3: Oppositie of regering

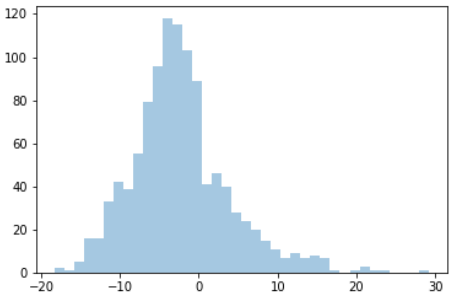
In figuur 4 zijn de distributies van de errors, zoals gedefinieerd in formule 6, te zijn van combinaties tussen regerings- en oppositiepartijen.



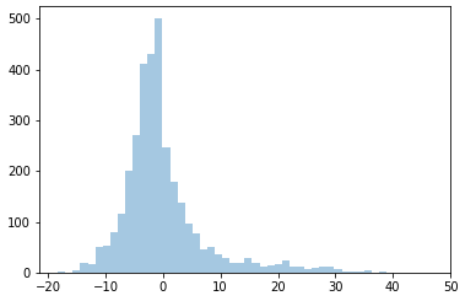
(a) Tussen twee regeringspartijen



(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 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	burger
wij	de fractie van	de premier	fractie van PARTIJ	burgers
fractie	fractie van	ik hoop	de fractie	onze
wij hebben	de fractie	arbeidsmarkt	fractie van	immigratie
KAMERLID	mijn fractie	hoop	de fractie van	niet
dank	moment	de arbeidsmarkt	premier	deze
buitengewoon	beantwoording	hij	politieke	gewoon
via	blij	kunnen	deal	natuurlijk
overleg	verschillende	woningmarkt	ik zou	belastinggeld
aangegeven	termijn	dadelijk	mij	land

Tabel 8: Meest karakteristieke n-grams per partij op basis van classificatie uit deelvraag 2 gedurende kabinet-Balkenende IV. Maximum aantal iteraties is 100 (in plaats van 5 in grid search).*(Vervolg)*

PvdA	PvdD	SGP	SP	VVD
vrouwen	dieren	beantwoording	zegt	PARTIJ
wij	natuur	mijn fractie	mensen	PARTIJ fractie
roc	bio industrie	wel	vandaar	onze fractie
medewerkers	de bio industrie	toch	niet	fractie
vragen	de bio	de bewindslieden	personeel	je
wij willen	veehouderij	een	voorstel	markt
goed	bio	diverse	is	want
belangrijk	de dieren	de voorzitter	nu	timmermans
weten	grondwet	de voorzitter ik	er	echt
in dit	dierenwelzijn	allerlei	leerlingen	voorzitter PARTIJ fractie

In tabel 9 zijn de resultaten van de classificatiescores te zien waarbij de classificatie getraind is op een zittingsperiode, maar getest op een andere.

Tabel 9: 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. Classificatiemethode uit deelvraag 1 zonder namen van partijen en Kamerleden. Maximum aantal iteraties is 100 (in plaats van 5 in grid search). 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
SGP	0.74	0.56	0.49
PvdD	0.73	0.64	0.45
PVV	0.70	0.50	0.60
SP	0.61	0.41	0.53
ChristenUnie*	0.55	0.37	0.22
D66	0.54	0.16	0.28
CDA*	0.53	0.28	0.43
PvdA	0.52	0.29	0.27
VVD*	0.51	0.18	0.10
GroenLinks	0.49	0.31	0.04
Totaal	0.58	0.34	0.35

4.1.4 DV4: Links of rechts

4.1.5 DV5: Woordgebruik van sprekers

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

	Precision	Recall	F1 score	Documenten
50PLUS	0.29	0.06	0.09	62
CDA	0.12	0.20	0.14	319
ChristenUnie	0.08	0.14	0.09	74
D66	0.22	0.22	0.22	384
GroenLinks	0.16	0.04	0.05	272
PVV	0.29	0.50	0.37	288
PvdA	0.25	0.19	0.21	422
PvdD	0.46	0.17	0.22	118
SGP	0.17	0.05	0.07	82
SP	0.34	0.33	0.33	620
VVD	0.31	0.26	0.24	381
Totaal	0.31	0.24	0.24	3023

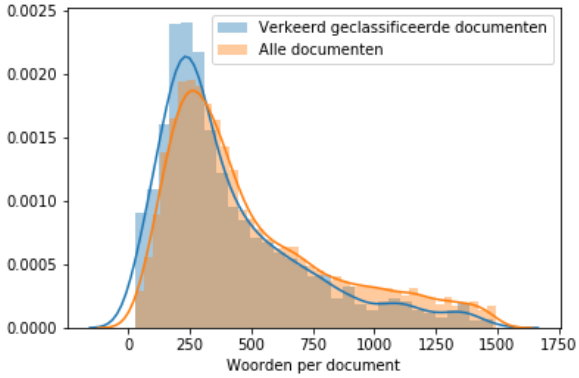
4.2 Discussie

4.2.1 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 4 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 aantal maximum iteraties was bij de beste classificatiemethode beperkt.

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 dat onderzoek. Het effect wat zij vinden tussen documentgrootte van 267 en 6666 is een verschil in nauwkeurigheid van 19,8%. Dit onderzoek vindt inderdaad dat kleinere 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. Mediaan documentlengte van foutief geclassificeerde documenten is 321 en voor alle documenten 386.

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

4.2.2 DV2: Invloed van namen

De resultaten laten zien dat de classificatie sterk afhankelijk is van partijnamen en namen van Kamerleden. Deze daling was te verwachten op basis van gerelateerd werk.

De n-grams in tabel 6 komen bij veel partijen overeen met hun ideologie, vooral bij one-issue partijen 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. Een hypothese is dat deze n-grams eigen zijn aan een individueel Kamerlid.

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 4 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 6 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.

4.2.3 DV3: Oppositie of regering

In tabel 3 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 resultaten van het eerste experiment suggereren een verschil in error binnen oppositie, binnen regering en tussen deze partijen.

De verwachting was dat de error normaal verdeeld zou zijn.

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 11: N-grams die bij minimaal één regeringspartij in beide kabinetten voorkomen in de 100 meest karakteristieke n-grams, maar niet toen deze partijen in oppositie zaten.

	PvdA	VVD
CDA	toezeggingen hun collega KAMERLID in aanpak collega	algemeen algemeen overleg toezegging helder overleg aangegeven voor voor PARTIJ
ChristenUnie	mijn waarop blij collega KAMERLID erg	gaan termijn blij met de volgens volgens mij blij beantwoording
PvdA		volgens volgens mij

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 9 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 en/of verandering van onderwerpen. 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 zittingsperioden 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.

4.2.4 DV4: Links of rechts

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

4.2.5 DV5: Woordgebruik van sprekers

De resultaten uit tabel 10 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 werken dit effect niet vinden. De meest karakteristieke n-grams van deze classificatie wijken daarnaast grotendeels niet af van die uit tabel 6.

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. Stel dat documenten van een spreker in de test set geclassificeerd moeten worden, dan kan het zijn dat deze meer karakteristieke vertoont met een andere partij, aangezien er geen woordvoerder van die partij en dat onderwerp in de training set zit, maar mogelijk wel van een andere partij. Een vervolgonderzoek kan kijken of dit een verklaring is.

4.2.6 Algemeen

Het vergelijken van deze resultaten met vergelijkbaar werk is problematisch, aangezien de keuzes en eigenschappen van hun onderzoek het niet een één-op-één vergelijking maken. Voorbeelden hiervan zijn de documentgrootte, baselines, 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? In Bertie Kaal, Isa Maks, and Annemarie van Elfrinkhof, editors, *From Text to Political Positions*, chapter 5, pages 93–115. John Benjamins Publishing Company, Amsterdan, 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] Conal Sathi Maneesh Bhand, Dan Robinson. Text classifiers for political ideologies, 2009.

[7] Bjørn Høyland, Jean-François Godbout, Emanuele Lapponi, and Erik Velldal. 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–. [Online; accessed ¡today¿].

[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 Grupta. 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.

A Slides