Invloed van ideologie beperkt op tekstclassificatie in Tweede Kamer

ingediend voor gedeeltelijke voltooiing van graad van
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

Abstract

In verschillende onderzoeken zijn parlementaire teksten geclassificeerd naar partij-affiliatie. Dit onderzoek is op zoek gegaan naar de beste classificatiemethode voor de Handelingen van de Tweede Kamer. Daarnaast is gekeken naar in hoeverre dit het gevolg is van ideologie. Hiervoor is gekeken naar de invloed van namen, in regering of oppositie zitten, positie op de links-rechts as en woordgebruik van sprekers.

De beste classificatiemethode met een *accuracy* van 0.80 is Support Vector Machines. Dit daalt naar 0.58 als achternamen van Kamerleden en partijnamen weggehaald worden. Het onderzoek vond ook aanwijzingen dat de classificatie afhankelijk is van of een partij in regering of oppositie zit. Aanwijzingen voor afhankelijkheid van positie op links-rechts as ontbreken daarentegen. Als laatste daalt de *accuracy* verder naar 0.27 als Kamerleden verdeeld worden over de training en test set, wat suggereert dat de oorspronkelijke classificatie afhankelijk was van woordgebruik van sprekers. Dit leidt tot de conclusie dat in grote mate de classificatie niet het gevolg is van ideologie.

Contents

1 Introductie

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

In diverse landen zijn al onderzoeken gedaan naar het classificeren naar partij-affiliatie op basis van teksten van politici [1, 3, 2, 4, 5]. Met deze tekst-classificatie naar partij-affiliatie proberen onderzoekers uit te vinden in hoeverre ideologie terug te vinden is in teksten van politici. De resultaten van de tekst-classificaties zijn in de meeste gevallen ruim boven de baseline. Diverse onderzoeken wijzen daarentegen ook naar redenen dat dit niet alleen het gevolg is van ideologie. Zo suggereren de resultaten van Hirst et al. [2] dat de partij-status (oppositie tegenover regering) van invloed is op de classificatie. Daarnaast laat dit onderzoek ook zien dat partijnamen een grote invloed hebben op de classificatie.

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

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

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

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

Op basis van vergelijkbare onderzoeken is de hypothese dat in grote mate de classificatie niet het gevolg is van ideologie, maar in beperkte mate wel.

Voor de eerste deelvraag zullen Support Vector Machine, Logistische Regressie en Naive Bayes met verschillende parameters vergeleken worden aan de hand van *accuracy* en F_1 score. Bij de tweede deelvraag wordt gekeken naar classificatie zonder achternamen van Kamerleden en partijnamen of met alleen achternamen van Kamerleden en partijnamen. De derde vraag bestaat

uit meerdere experimenten, waarin gekeken zal worden naar de hoeveelheid misclassificaties binnen regering of oppositie tegenover tussen regering en oppositie. Daarnaast zal gekeken worden naar overlap in woordgebruik binnen regering en verschil in scores als een partij gewisseld is van partij-status. Bij de vierde vraag zal gekeken worden naar een verband tussen misclassificaties en afstand tussen twee partijen op de links-rechts as. Als laatste zal voor de vijfde vraag de classificatie herhaald worden met Kamerleden verdeeld over training en test set.

Overzicht van scriptie Sectie 2 bevat vergelijkbare onderzoeken in andere parlementen. 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 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 is uitgebreider gekeken worden de effecten van verschillende classificatiemethoden. In de latere secties worden de aspecten besproken 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 *accuracy* 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 de baseline. Als verklaring voor dit verschil ten opzichte van de uitersten zeggen ze dat gematigden een minder duidelijke ideologie hebben.

Yu et al. [5] richtten 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 congreslid en het label de partij. Voor het Huis van Afgevaardigden vonden ze een *accuracy* 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 *accuracy* op van 88.0% (baseline van 55.0%) en andersom 67.6% (baseline van 51.5%). Hun verklaring voor dit verschil was dat het Huis van Afgevaardigden sterker verdeeld is langs partijlijnen.

Vervolgens herhaalden ze de classificaties op het Huis van Afgevaardigden uit 2015, maar testten ditmaal op de Senaat elk jaar tussen 1989 en 2006 afzonderlijk. Hierin was een stijging in *accuracy* van 60.0% (baseline van 55.0%) in 1989 naar 87.0% (baseline van 55.0%) in 2006 te zien, maar met twee duidelijke dalen. Ze presenteren twee mogelijke verklaringen voor de trend; het veranderen van de onderwerpen en het sterker verdeeld 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 vonden 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 voegden ze alle teksten van een parlementslid bij elkaar en deelden die op in documenten van gelijke grootte. Voor documentgrootte van 267 woorden werd een *accuracy* van 44.0% gevonden oplopend tot 61.8% (baseline van 38-39%) voor documentgrootte van 6666.

Bhand et al. [6] richtten zich op het classificeren van leden van het Amerikaanse Congres in 2005, op basis van partij-affiliatie (Republikeins of Democratisch). 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 [3] probeerde interventies van politici te classificeren op basis van geslacht, leeftijdsgroep, partij-affiliatie en oriëntatie in het Portugese parlement. In het geval van classificatie op basis van partij-affiliatie bereikte men een F_1 score van 0.90 (baseline niet vermeld, zes partijen).

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

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*. De beste classificatiemethode was afhankelijk van de dataset. Voor het Huis van Afgevaardig-

den was het Support Vector Machines met als weging *tf-idf* en voor de Senaat Bernouilli Naive Bayes.

Hirst et al. [2] maakten gebruik van Support Vector Machines. 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. [6] gebruikten verschillende n-grams, inclusief verschillende manieren van *smoothing*. Ze testten als weging voor features zowel *boolean* als *tf*, waarbij ze vonden 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*.

Ferreira maakten gebruik van twee classificatiemethoden: Logistische regressie en *margin-infused relaxed algorithm* (MIRA) [3]. Logistische regressie werd aangevuld met *group Lasso* regularisatie, wat het beste resultaat opleverde. 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 een F_1 score van twee procent hoger opleverde dan gebruik van normale stemming.

2.3 Invloed van partijnamen of sprekersnamen

Diermeier et al. [4] 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. Hirst et al. [2] vonden inderdaad dat partijnamen - en het weglaten daarvan - bij het Europees Parlement een grote invloed hebben op de classificatie. Bij het Europees Parlement was te zien dat een spreker de eigennaam gebruikte, terwijl in het Canadese parlement vooral te zien was dat de naam van de andere partij gebruikt wordt door een spreker.

2.4 Invloed van oppositie of regering

Hirst et al. [2] 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). 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 classificaties op het ene parlement en testten deze op het andere parlement. Hierbij vonden zij in beide gevallen een *accuracy* ver onder 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 zijn, zijn de Handelingen van de Tweede Kamer gedurende het missionaire kabinet-Rutte II (5 november 2012 tot 22 maart 2017). Er was 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 bijbehorende metadatabestanden. 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.

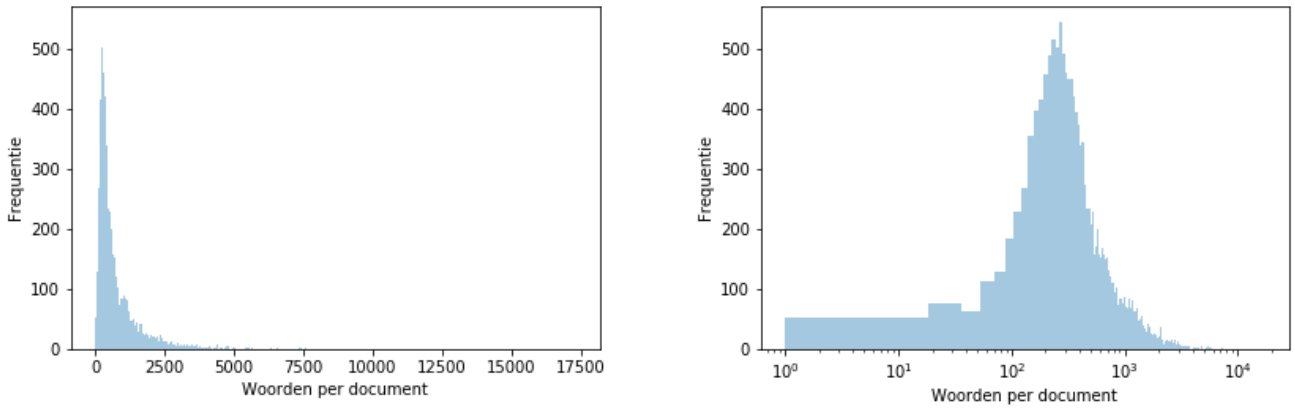
Deze dataset bestaat uit een aantal soorten spreekbeurten; debatbijdragen, interrupties en antwoorden. Een debatbijdrage is de eerste onafgebroken spreekbeurt die een spreker geeft achter een spreekgestoelte, aangeduid in de xml-file met het attribuut *nieuw="ja"*. Dit kan een bijdrage in een debat zijn 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 debatbijdrage geïnterrumpeerd kan worden, kan deze inhoudelijk doorlopen in een antwoord van een spreker. Vergelijkbare onderzoeken voegen vaak alle teksten van een spreker samen tot één document. Dit was alleen niet mogelijk voor dit onderzoek met de hoeveelheid kleine partijen in de Tweede Kamer, die dan niet altijd in een training of test set zijn vertegenwoordigd. Daarom was in dit onderzoek ervoor gekozen om een debatbijdrage samengevoegd met alle bijbehorende antwoorden te beschouwen als één document.

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

Deze dataset bevatte vervolgens naast de verkozen partijen na de Tweede Kamerverkiezingen van 2012 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 was en er overlap zat met hun oorspronkelijke of gelieerde partij, waren 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 verschilden in grootte (aantal woorden). De distributie van documentgrootte lijkt op een lognormale verdeling, maar met een Kolmogorov-Smirnov test was hier geen bewijs voor gevonden [9].



(a) Normale x-as

(b) Logaritmische x-as

Figure 1: Aantal woorden per document

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

Table 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 zijn een aantal classificatiemethoden vergeleken. Aangezien het niet mogelijk was om alle classificatiemethoden te vergelijken, beperkte dit onderzoek zich tot classificatiemethoden die gebruikt zijn in vergelijkbare onderzoeken, zoals besproken in sectie 2.2. Er was ervoor gekozen om alleen gebruik te maken van methoden waarvan reeds implementaties beschikbaar waren in scikit-learn. Voor alle methoden werd gezocht naar de beste parameters, ook wel bekend als een grid search. Deze grid search werd gedaan door 5-fold cross-validation, waarbij de training set steeds 80% was en de test set 20% van de totale dataset. Een totaal aantal van 6480 combinaties van methoden en parameters zijn getest. De verwachting was dat de scores lager zijn dan die gevonden in het gerelateerd werk, omdat de documentgrootte kleiner is en de baseline scores lager zijn.

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

Bag-of-words model Bag-of-words model is de meest gebruikte representatie van data in vergelijkbare onderzoeken. Deze is ook gebruikt in dit onderzoek. Bij het bag-of-words model wordt elk document gerepresenteerd als een vector, waarbij elke kolom een woord is 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 waren de volgende wegingen voor woorden getest: *boolean* (wel of niet aanwezig), *tf* (woordfrequentie), *tf-norm* (woordfrequentie genormaliseerd door documentlengte) en *tf-idf* (woordfrequentie gecompenseerd voor documentfrequentie). Daarnaast werd in dit onderzoek geëxperimenteerd met een minimale of maximale woord- of documentfrequentie. Ook is gekeken naar het effect van combinaties van de volgende 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 van belang zijn, want als bijvoorbeeld het woord *asfalt* er in voorkomt, dan is het voor ideologie mogelijk relevant of er *minder asfalt* of *meer asfalt* staat.

Support Vector Machine en Logistische Regressie De meest voorkomende techniek in vergelijkbaar onderzoek is Support Vector Machine (SVM). Een an-

dere techniek die gebruikt wordt, is logistische regressie. Beide hebben 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 was 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 was die van maximaal aantal iteraties, waarvoor de scikit-learn standaard 5 is. Volgens scikit-learn convergeert de SGDClassifier rond de $10^6/n$ iteraties waar n het aantal documenten in de training set is. In het geval van deze dataset zou dat 84 iteraties zijn. Vanwege de grootte van de grid search was het voor dit onderzoek niet mogelijk het maximaal aantal iteraties te verhogen tijdens de grid search. De resultaten buiten de grid search zullen gebaseerd zijn op een maximaal aantal iteraties van 100.

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

Beoordelen van kwaliteit De meest gebruikte methoden om kwaliteit van politieke tekstclassificatie te beoordelen zijn *accuracy* en F_1 score, die opgebouwd is uit *recall* en *precision*. Deze scores worden berekend op basis van vier variabelen. Deze variabelen geven weer hoeveel documenten wel of niet bij een partij horen, en of deze wel of niet als dusdanig zijn 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. *Accuracy* wordt voor de hele classificatie gedaan en niet per klasse. *Precision* is het percentage van documenten geclassificeerd als een partij, dat ook bij die partij hoort. *Recall* is het percentage documenten van documenten behorende

tot een partij, dat ook als die partij geclassificeerd is. F_1 is het harmonisch gemiddelde van *recall* en *precision*. *Precision*, *recall* en daarmee F_1 worden per partij berekend. Er zijn drie varianten om deze scores voor de hele classificatie te berekenen.

Allereerst is er *micro*, waarbij alle variabelen bij elkaar opgeteld worden en vervolgens de scores berekend. Dit leidt ertoe dat resultaten van partijen met 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*, waarbij alle scores per partij berekend worden en daarvan het gemiddelde wordt 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, was gekozen voor *gewogen* F_1 score naast *accuracy*.

3.2.2 DV2: Invloed van namen

In Diermeier et al. [4] werd aangenomen dat namen een groot effect hebben op de classificatie. Hirst et al. [2] bevestigden dit voor het Europees Parlement. Aangezien hier bij deelvraag 1 niet voor was gekozen, werd bij deze deelvraag gekeken hoe groot het effect hiervan is, specifiek gericht op partijnamen en achternamen van Kamerleden. Op basis van vergelijkbaar onderzoek is de hypothese dat de achternamen van Kamerleden en partijnamen van invloed zijn.

Voor deze deelvraag werd wederom een classificatie gedaan met de classificatiemethode die resulteerde uit deelvraag 1. In deze classificatie werden alle partijnamen vervangen door *PARTIJNAAM* en alle achternamen van Kamerleden vervangen door *KAMERLIDNAAM*. Deze namen waren uit de Handelingen gehaald. Voor partijnamen waren ook lidwoorden toegevoegd en 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 waren 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 was 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 was gekeken of er nog namen achter zijn gebleven, maar die waren niet gevonden.

Ook werd gekeken naar classificatie met alleen partijnamen en achternamen van Kamerleden. Alle andere woorden worden weggehaald. Namen van Kamerleden en partijen die niet aan elkaar geschreven worden, zoals *Partij van de Arbeid*, zijn aan elkaar geschreven zodat het één feature is. Doordat alle

andere woorden weggehaald zijn, waren de bi- en trigrams combinaties van namen die zinnen uit elkaar kunnen staan, dus die niet meer informatie geven dan unigrams. Daarom werd er gebruikt van de classificatiemethode uit deelvraag 1, maar met alleen unigrams.

Op basis van de hypothese is de verwachting dat voor de classificatie zonder namen de scores een stuk lager zijn dan deelvraag 1 en de scores van de classificatie met alleen namen aanzienlijk hoger zijn dan de baseline scores.

3.2.3 DV3: Oppositie of regering

Om deze deelvraag te beantwoorden zijn drie experimenten uitgevoerd. Twee daarvan zijn gebaseerd op experimenten uit Hirst et al. [2] voor dezelfde vraag. De derde is ontwikkeld voor dit onderzoek. Met deze laatste wordt begonnen. Bij deze deelvraag is de classificatiemethode uit deelvraag 1 zonder achternamen van Kamerleden en partijnamen gebruikt. De hypothese is op basis van de bevindingen van Hirst et al. dat de classificatie inderdaad afhankelijk is van partij-status.

Als er een afhankelijkheid is van partij-status, dan is het te wachten dat het aantal misclassificaties minus verwachte waarde binnen regeringspartijen en binnen oppositiepartijen hoger ligt dan tussen een oppositiepartij en regeringspartij. 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 werd 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) bleek deze correlatie tussen het aantal *false positives* van een partij en het aantal documenten behorend tot die partij.

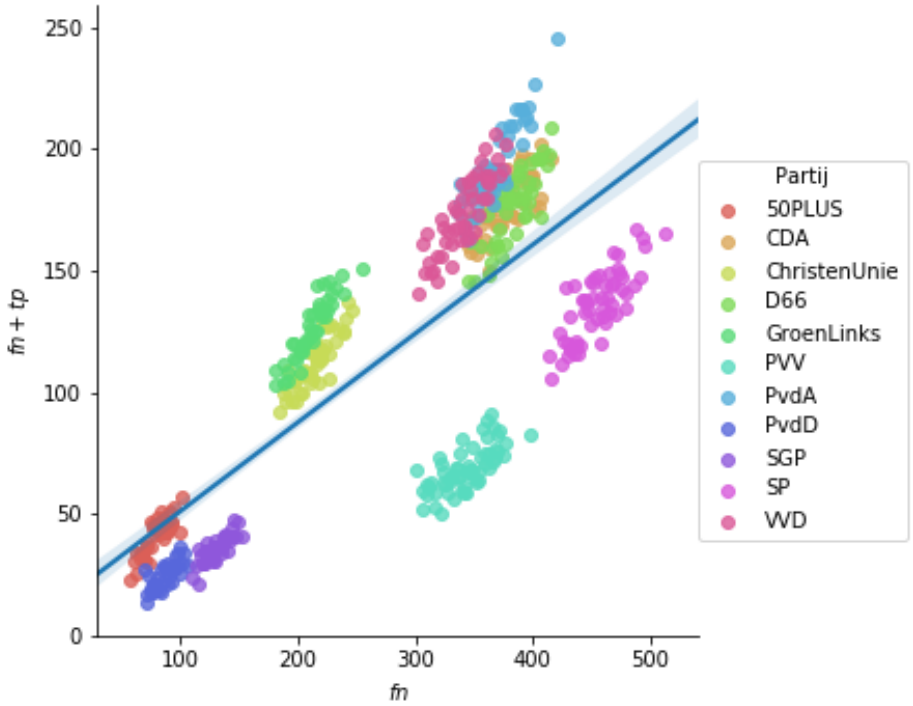


Figure 2: Het aantal *false positives* ten opzichte van het aantal documenten behorend tot die partij (*false negatives* en *true positives*). Dit is op basis van 100 classificaties met verschillende train en test set. De Pearson correlatie is 0.77 en de p-waarde 5.40×10^{-101}.

Op basis van dit verband is het verwachte aantal documenten ($V_{i,j}$) van partij i die foutief geclassificeerd worden als partij j gedefinieerd als

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

waar $i \neq j$. De teller van de breuk is het aantal documenten die bij partij j horen en de noemer het aantal documenten die niet bij partij i horen. Op deze manier is $\sum_{j=0}^{n}(V_{i,j}) = fn_i$ waar n het aantal partijen is minus partij i.

De error ($e_{i,j}$) is dan het verschil van het daadwerkelijk aantal misclassificaties ($D_{i,j}$) en de verwachte waarde ($V_{i,j}$)

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

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

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, werden de distributies binnen regeringspartijen en binnen oppositiepartijen vergeleken met de distributie tussen beide groepen. Om de invloed van variantie door de willekeurige splitsing documenten voor trainen en testen te beperken, werd de classificatie 100 keer gedaan. Als de distributies normaal verdeeld zijn, vond de statistische

test plaats op basis van een eenzijdige t-toets. Als de distributies niet normaal verdeeld zijn, vond dit plaats door een Mann-whitneytoets. Het gekozen significantieniveau (α) is 0.01. De nulhypothese is dat er geen verschil is tussen de verdelingen. De alternatieve hypothese is dan dat de distributie van binnen oppositie of regering groter is dan die tussen een regerings- en oppositiepartij. Op basis van de bevindingen van Hirst et al. was de hypothese dat de nulhypothese verworpen kan worden.

In het eerste experiment gebaseerd op Hirst et al. zijn de meest karakteristieke woorden per partij van de ene zittingsperiode vergeleken met de meest karakteristieke woorden per partij van de andere zittingsperiode, waar een kabinet uit andere partijen bestond. De verwachting is dat als de classificatie niet het gevolg is van partij-status dat de woorden bij een partij blijven en niet gekoppeld zijn aan in oppositie of regering zitten. Aansluitend bij de hypothese is dus de verwachting dat woorden wel wisselen van partij wanneer ze van partij-status gewisseld zijn.

In het tweede experiment gebaseerd op Hirst et al. zijn de classificaties getraind op een zittingsperiode en getest op een andere zittingsperiode, waar wederom een kabinet uit andere partijen bestond. Als de classificatie afhankelijk was van partij-status was de verwachting dat de scores van partijen die gewisseld zijn van partij-status sterker gedaald zijn dan partijen die niet van partij-status zijn veranderd. Op basis van de hypothese is dan ook de verwachting dat bij de partijen die gewisseld zijn partij-status een sterkere daling te zien is.

Als vergelijkingsmateriaal was 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.

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

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

3.2.4 DV4: Links-rechts as

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

Er zijn verschillende methoden om partijen in te delen op een links-rechts as. Er is hier gekozen voor de indeling van het Manifesto Project [14]. Het Manifesto Project geeft scores op een heel aantal politieke posities, waaronder de links-rechts as, op basis van het verkiezingsprogramma van dat jaar. Voor de dataset van kabinet-Rutte II is gebruikt gemaakt van de scores op basis van de verkiezingsprogramma's voor de verkiezingen van 2012.

Table 3: Scores op de links-rechts as per partij van het Manifesto Project voor de verkiezingsprogramma's van 2012.

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

Er wordt vervolgens gekeken door middel van een Pearson correlatie toets

of er een correlatie is tussen de error van twee partijen en de afstand op de links-rechts as van die partij. Het significantieniveau (α) hiervoor is opnieuw 0.01. De nulhypothese is dat er geen negatieve correlatie is tussen de error en de afstand op de links-rechts as. De alternatieve hypothese is dat er wel een negatieve correlatie is tussen de error en de afstand op de links-rechts as.

Als uit deelvraag 3 blijkt dat partij-status invloed heeft op de error, zal bovenstaande methode ook uitgevoerd worden voor de aparte combinaties; binnen oppositie en tussen regeringspartij en oppositiepartij. Binnen regering is dit niet mogelijk aangezien er maar één afstand is (tussen PvdA en VVD).

De voorspelling op basis van de hypothese is dat de nulhypothese verworpen kan worden.

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 classificatie naar partij-affiliatie. Hirst et al. [2] plaatsten al een soortgelijke kanttekening bij de resultaten van Diermeier et al. [4]. De hypothese is dat de classificatie afhankelijk is van woordgebruik van sprekers

Om te kijken of dit effect er is, wordt er opnieuw een classificatie gedaan met de classificatiemethode uit deelvraag 1 zonder achternamen van Kamerleden en partijnamen. 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 verwachting is daarom ook dat deze classificatie lagere scores vindt dan die van deelvraag 2.

4 Resultaten

4.1 DV1: Beste classificatiemethode

In figuur 3 zijn de uitslagen van de grid search te zien. Hierin is te zien dat SVM en logistische regressie beide hoge scores behalen, maar dat logistische regressie ook veel lage scores haalt tussen 0 en 0.1. Naive Bayes zit tussen de 0.25 en 0.6.

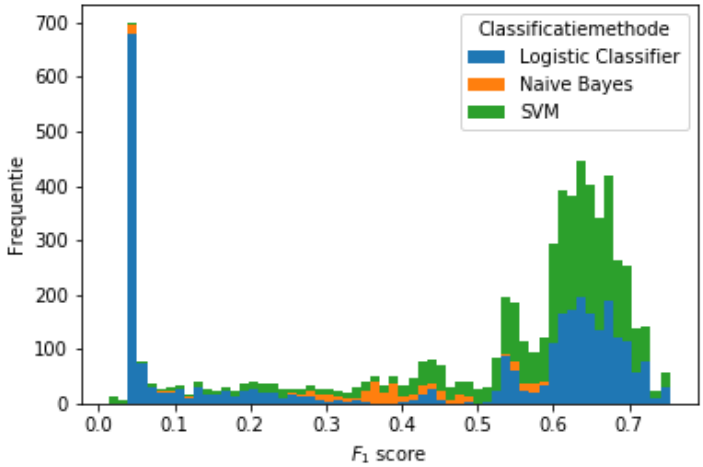


Figure 3: Histogram van de grid search met de F_1 scores van de classificatiemethoden

Het beste resultaat werd bereikt met Support Vector Machines gebruikmakend van *stochastic gradient descent learning* en L2 regularisatie. In de grid search behaalde deze methode een F_1 score en *accuracy* van 0.75. Voor beide scores was dit het hoogste van de grid search. De woorden waren hierbij gestemd. De features waren zowel unigrams, bigrams als trigrams. Geen features zijn weggelaten door minimale of maximale documentfrequenties. De waarden van deze features waren *tf-idf* scores. Het maximum aantal iteraties was 5 voor de grid search, maar de rest van resultaten zijn op basis van 100 iteraties.

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

Table 4: Classificatie scores per partij van de beste classificatiemethode (SVM). Gemiddelde van vijf splitsingen van training en test set.

	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

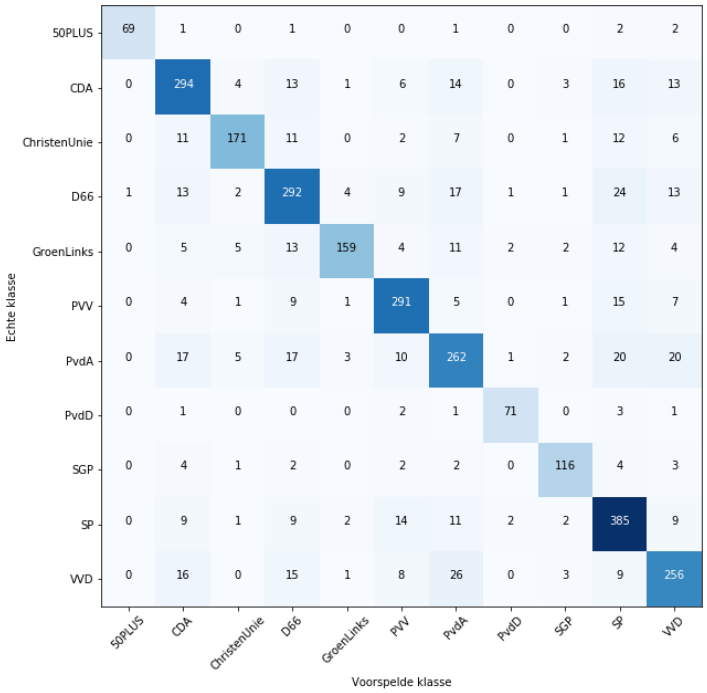


Figure 4: Confusion matrix van de beste classificatiemethode (SVM). Gemiddelde van vijf splitsingen van training en test set.

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

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

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

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

4.2 DV2: Invloed van namen

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

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

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

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

Table 7: Classificatie scores per partij van beste classificatiemethode (SVM) uit deelvraag 1 zonder achternamen van Kamerleden en partijnamen met het relatieve verschil in F_1 score ten opzichte van tabel 4. Gemiddelde van vijf splitsingen van training en test set.

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

Table 8: Meest karakteristieke n-grams per partij op basis van de classificatiemethode (SVM) uit deelvraag 1 zonder achternamen van Kamerleden en partijnamen gedurende kabinet-Rutte II.

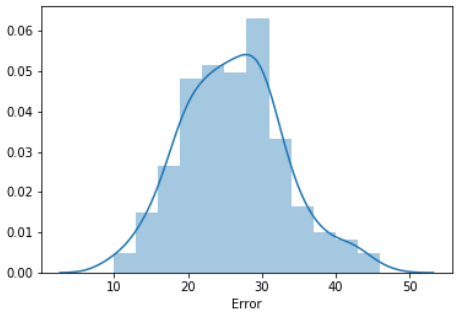
50PLUS	CDA	ChristenUnie	D66	GroenLinks
ouderen	PARTIJ fractie	dementie	mijn fractie	belastingontwijking
gepensioneerden	inwoners	gezinnen	mijn	zou
plussers	regering	zullen	natuurlijk	kamer hierover te
50 plussers	PARTIJ	vluchtelingen	fractie	persoonsgebonden
oudere	de regering	ik hoop	het kabinet	in elk geval
koopkrachtontwikkeling	diverse	inderdaad	buitengewoon	elk geval
50	hier	motie	belangrijk	vluchtelingen
werkenden	echt	hoop	vandaag	in elk
PARTIJ	een aantal	begeleiding	kabinet	hierover te informeren
overwegende dat	fractie	horeca	daarom	budget

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

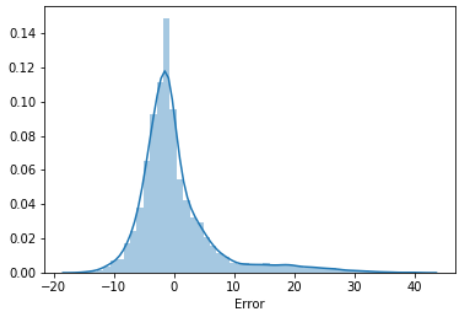
PVV	PvdA	PvdD	SGP	SP	VVD
islamitische	en	dieren	mevrouw de voorzitter	zegt	PARTIJ
nederland	jongeren	natuur	mevrouw de	mening dat	volgens mij
miljard	daarbij	constaterende	beantwoording	van mening dat	PARTIJ fractie
natuurlijk	tevens	constaterende dat	voor de beantwoording	bezuinigingen	aruba
islam	vragen	dierenwelzijn	bewindslieden	mensen	regelgeving
de islam	wij	bio industrie	de beantwoording	huurders	aangegeven
al	beter	industrie	wel	voorstellen	speelveld
dit kabinet	kinderen	de bio	punt	segregatie	volgens
brussel	samen	dierproeven	nadrukkelijk	van mening	essentieel
asielzoekers	toezeggingen	de bio industrie	je	bestuurders	en

4.3 DV3: Oppositie of regering

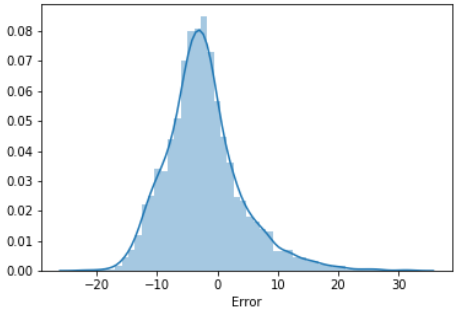
In figuur 5 zijn de distributies van de errors, zoals gedefinieerd in formule 6 te zien van combinaties van regerings- en oppositiepartijen. Bijgevoegd zijn het aantal combinaties (N), de mediaan en de interkwartielafstand (IKA).



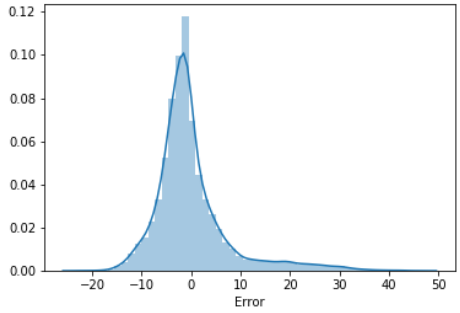
(a) Tussen twee regeringspartijen (N=200, Mediaan=25.83, IKA=9.18)



(b) Tussen twee oppositiepartijen (N=8100, Mediaan=-1.09, IKA=4.98)



(c) Tussen een regeringspartij en een oppositiepartij (N=4000, Mediaan=-2.92, IKA=6.82)



(d) Totaal (N=12100, Mediaan=-1.40, IKA=5.76)

Figure 5: Genormaliseerde distributie van de error uit formule 6 voor de verschillende combinaties.

Voor alle distributies was de nulhypothese verworpen worden dat deze normaal verdeeld zijn ($p < 0.01$). In tabel 9 is vervolgens te zien dat er een significant verschil is tussen de distributies binnen regering en oppositie tegenover de distributie tussen regering en oppositiepartij. Tussen regeringspartijen zijn er gemiddeld 26.11 misclassificaties meer dan verwacht en tussen oppositiepartijen gemiddeld 0.43.

Table 9: Uitslagen van eenzijdige Mann-whitneytoets tussen de distributie tussen een regeringspartij en oppositiepartij en twee distributies. α is 0.01.

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

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

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

CDA	ChristenUnie	D66	GroenLinks	PVV
PARTIJ fractie	fractie van PARTIJ	premier	PARTIJfractie	burgers
wij	de fractie van	de premier	fractie van PARTIJ	onze
fractie	de fractie	hij	de fractie	burger
wij hebben	fractie van	ik hoop	de fractie van	gewoon
aangegeven	moment	arbeidsmarkt	fractie van	natuurlijk
PARTIJ fractie heeft	mijn fractie	plannen	premier	de burgers
dank	verschillende	hoop	mij	door
overleg	beantwoording	de arbeidsmarkt	ik	politie
KAMERLID	PARTIJfractie	dadelijk	politieke	land
buitengewoon	blij	ministerie	en	niet

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

PvdA	PvdD	SGP	SP	VVD
vrouwen	dieren	mijn fractie	mensen	PARTIJ
wij	bio industrie	wel	zegt	PARTIJ fractie
belangrijk	dierenwelzijn	beantwoording	leerlingen	onze fractie
kinderen	bio	voorzitter ik wil	is	fractie
goed	de bio industrie	toch	niet	ondernemers
vragen	de bio	diverse	vandaar	je
antwoorden	natuur	de bewindslieden	verdrag	praten
medewerkers	dierproeven	allerlei	personeel	markt
ben	veehouderij	natuurlijk	problemen	dat
iedereen	industrie	bewindslieden	waarom	voorzitter PARTIJ fractie

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

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

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

4.4 DV4: Links-rechts as

In tabel 6 is de error te zien ten opzichte van de afstand op de links-rechts as.

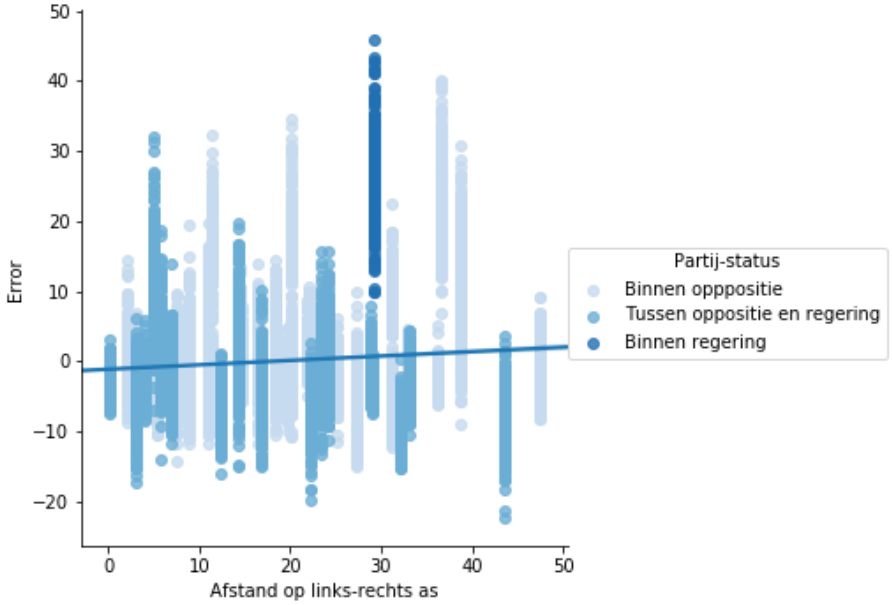


Figure 6: Error ten opzichte van de afstand op de links-rechts as van twee partijen. Gebaseerd op 100 classificaties met verschillende test en train set. De Pearson correlatie is 0.09 en de p-waarde 2.39×10^{-20}.

De Pearson correlatie van 0.09 is daarmee met een p-waarde van 2.39×10^{-20} significant op het significantieniveau van 0.01, maar wel positief. Uit deelvraag 3 bleek dat de error binnen oppositie of regering significant afweek van de error tussen regering en oppositie. Dit effect lijkt ook zichtbaar in figuur 6. Daarom wordt er in tabel 12 ook gekeken naar de correlatie tussen afstand op de links-rechts as en error binnen oppositie en tussen regerings- en oppositiepartij. Beide correlaties zijn statistische significant op het significantieniveau van 0.01, maar opvallend genoeg in tegengestelde richting.

Table 12: Pearson correlatie tussen error en afstand op de links-rechts as voor combinaties van partij-status.

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

4.5 DV5: Woordgebruik van sprekers

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

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

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

5 Discussie

5.1 DV1: Beste classificatiemethode

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

Dit onderzoek heeft zich beperkt tot methoden genoemd in vergelijkbare onderzoeken en waarvan de implementatie beschikbaar is in scikit-learn. Een aantal methoden die in gerelateerde literatuur leidden tot goede classificaties zijn daarom niet getest. Ook nieuwe methoden die nog niet gebruikt zijn in een vergelijkbaar onderzoek voor politieke tekst classificatie zijn daarom niet getest. Daarnaast richtte zich dit ook maar op een beperkt aantal parameterwaarden. Voor vervolgonderzoek kan daarom dit onderdeel uitgebreid worden. Het effect van het beperkte maximum iteraties was bij de beste classificatiemethode 0.05.

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

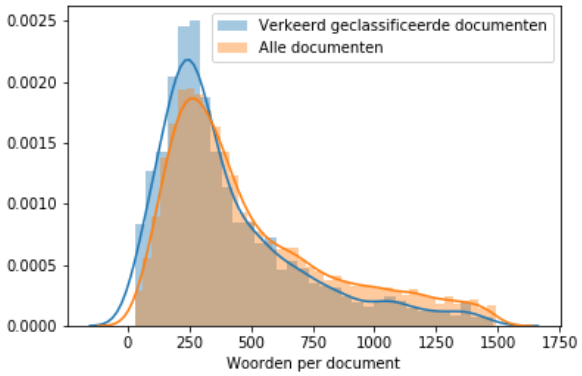


Figure 7: Genormaliseerde distributie van documentlengtes van foutief geclassificeerde documenten en alle documenten. Totaal van 5-fold cross-validation, waardoor documenten vaker voor kunnen komen. Mediaan documentlengte van foutief geclassificeerde documenten is 321 en voor alle documenten 386.

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).

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

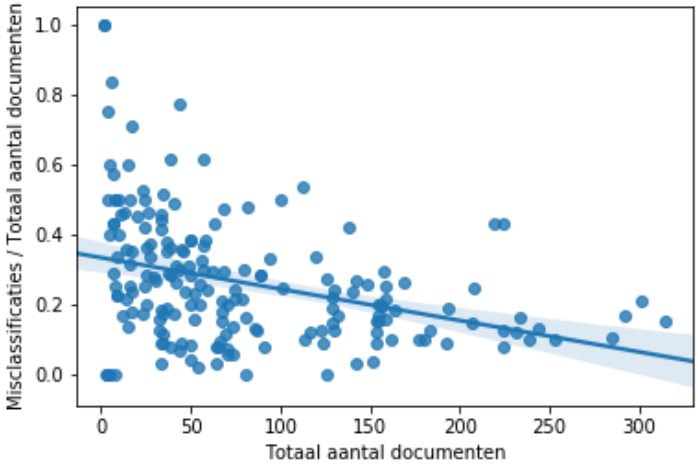


Figure 8: Aantal misclassificaties gedeeld door totaal aantal documenten per spreker tegenover totaal aantal documenten van een spreker. Misclassificaties zijn totaal van 5-fold cross-validation, waardoor documenten vaker mee kunnen tellen. De pearson correlatie is -0.28 en de p-waarde 1.07×10^{-4}.

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

5.2 DV2: Invloed van namen

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

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

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

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

5.3 DV3: Oppositie of regering

In tabel 4 is het opvallend dat de coalitiepartijen lage scores krijgen. Daarnaast laat figuur 4 zien dat er een hoge overlap zit tussen deze twee partijen. De resultaten van het eerste experiment, ontwikkeld voor dit onderzoek, vinden een afhankelijkheid van partij-status. De twee andere experimenten versterken deze bevindingen niet, zoals wel het geval was bij Hirst et al. [2]. Hieronder worden de resultaten nader besproken.

De statistische toetsresultaten in tabel 9 laten zien dat inderdaad de error groter is binnen oppositie of regering dan tussen een regerings- en oppositiepar-

tij. Met name regeringspartijen lijken lastiger uit elkaar te halen. Dit suggereert dat inderdaad partij-status invloed heeft op de classificatie.

De verwachting was dat de error normaal verdeeld zou zijn. De verdelingen uit figuur 5 hebben globaal wel de vorm van een normaal verdeling. In figuur 2 is het daarnaast opvallend dat partijen zoals SP en PVV ruim onder de regressielijn zitten, terwijl andere partijen er een stuk boven zitten. Dit geeft aanleiding te vermoeden dat er naast het aantal documenten van een partij nog meer factoren van invloed zijn op het aantal misclassificaties en daarmee de verwachte waarde. En deze verwachte waarde en de daar uit volgende error zijn een belangrijke aanname van deze methode. Voor deze methode is het dus belangrijk uit te vinden of dit een goede benadering is van de verwachte waarde. In deelvraag 4 wordt gekeken of links/rechts positie hier nog invloed heeft. Voor een vervolgonderzoek kan nog verder gekeken worden naar invloeden op verwachte waarde of andere confounding biases.

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

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

		Kabinet-Rutte II	
		PvdA	VVD
Kabinet-Balkenende IV	CDA	*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 11 laten een duidelijke daling zien ten opzichte van een classificatie van alleen kabinet-Rutte II. Deze algemene daling kan verklaard worden door verschuiving in ideologie, verschil in woordgebruik, verandering van onderwerpen en/of verandering in aantal documenten per partij. De daling is het grootst bij VVD, maar valt mee bij de twee andere partijen die gewisseld zijn van partij-status, ChristenUnie en CDA. Daarnaast is de daling ook heel sterk bij oppositiepartijen GroenLinks en D66, alsook de regeringspartij in beide kabinetten, PvdA. Dat de daling niet consequent groter is bij partijen die gewisseld zijn van partij-status, suggereert dat de invloed van partij-status beperkt is op de classificatie.

Dat de experimenten uit Hirst et al. in hun onderzoek wel invloed vinden, maar in dit onderzoek niet kan komen doordat hun onderzoek zich richt op binaire classificatie, terwijl dit onderzoek meerdere partijen heeft. Zo kan het ontbreken van gemeenschappelijke n-grams komen doordat regeringspartijen zich ook van elkaar moeten onderscheiden in dit onderzoek, waarvoor n-grams die relevant zijn voor partij-status weinig effect hebben, terwijl in het onderzoek van Hirst et al. de regeringspartij alleen onderscheiden hoeft te worden van de oppositiepartij. Daarnaast verklaarden zij dat de daling tussen twee zittingsperioden het gevolg was van de wisseling van partij-status. In dit onderzoek kon daarentegen gekeken worden naar effecten op partijen niet die niet van partij-status zijn gewisseld. Hierin was te zien dat de daling ook aanwezig was bij partijen die niet gewisseld zijn van partij-status.

5.4 DV4: Links-rechts as

De correlatie was tegen de verwachting in positief, waardoor de nulhypothese niet verworpen kan worden. Een deel van deze positieve correlatie lijkt te wijten aan de error tussen de twee regeringspartijen. Daarnaast is het opvallend dat tussen oppositiepartijen de correlatie ook positief is, maar tussen oppositie en regeringspartij juist, zoals eigenlijk verwacht, negatief. Een verklaring hiervoor is niet gevonden.

Alle correlaties zijn statistisch significant, maar de Pearson correlatie en daarmee effectgrootte is klein. Daarnaast is het ook opvallend dat de twee combinaties van partij-statussen een andere correlatierichting hebben. Dit suggereert dat de statistische significantie het gevolg is van de grote steekproef en maar een klein effect [15].

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

5.5 DV5: Woordgebruik van sprekers

De resultaten uit tabel 13 zijn laag, amper hoger dan de baseline. Dit suggereert inderdaad dat eerdere classificaties in grote mate toch afhankelijk waren van het woordgebruik van sprekers. Dit is opmerkelijk aangezien vergelijkbare

onderzoeken dit effect niet vinden. De meest karakteristieke n-grams van deze classificatie wijken daarnaast grotendeels niet af van die uit tabel 8.

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

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

5.6 Algemeen

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

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

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

6 Conclusies

Dit onderzoek vindt een *accuracy* en F_1 score van 0.80 voor het classificeren van spreekbeurten in de Tweede Kamer naar partij-affiliatie. De beste classifi-

catiemethode maakt gebruik van Support-Vector Machines. De baseline scores zijn respectievelijk 0.11 en 0.15. Als rekening wordt gehouden met partijnamen en achternamen Kamerleden daalt de *accuracy* naar 0.58 en de F_1 score naar 0.57. Dit onderzoek vindt aanwijzingen dat deze classificatie afhankelijk is van de partij-status (oppositie of regering). Daarnaast vindt dit onderzoek geen aanwijzingen dat de classificatie afhankelijk is van positie op de links-rechts as. Als rekening wordt gehouden met woordgebruik van individuele Kamerleden, dalen de *accuracy* en F_1 verder naar 0.27. Daarmee lijkt de classificatie naar partij-affiliatie in grote mate niet het gevolg van ideologie. Deze conclusie trekt daarmee de bruikbaarheid van tekstclassificatie voor het vinden van een relatie tussen woordgebruik en ideologie in twijfel. Op een aantal punten wijken de bevindingen van dit onderzoek af van vergelijkbare onderzoeken. Voor een vervolgonderzoek kan dit onderzoek uitgebreid worden met een aantal aanbevelingen.

References

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

[6] Maneesh Bhand, Dan Robinson, and Conal Sathi. Text classifiers for political ideologies, 2009.

[7] Bjørn Høyland, Jean-François Godbout, Emanuele Lapponi, and Erik 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.

[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.

[15] Joseph F. Hair, Jr., Rolph E. Anderson, Ronald L. Tatham, and William C. Black. *Multivariate Data Analysis (6th Ed.)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2006.