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	
Affliatie	UvA, FNWI, IvI	
Email	maartenmarx@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 hebben als ook een bekende ideologie in de vorm van een partij. Deze informatie kan vervolgens toegepast worden bij andere teksten die wellicht ideologisch van aard zijn. Bijvoorbeeld, aan de hand van deze informatie kan men teksten uit kranten classificeren op basis van ideologie [1, 2].

In diverse landen zijn al verschillende onderzoeken gedaan naar het classificeren van partij-affiliatie op basis van teksten van politici[3, 1]. Met deze 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 een breder scala aan mogelijke methoden en daarnaast specifiek op de Nederlandse politiek. De onderzoeksvraag luidt daarom dus ook: "In hoeverre is classificatie op basis van partij-affiliatie aan de hand van spreekbeurten in de Tweede Kamer het gevolg van ideologie?"

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

1. Wat is het beste classificatiemodel voor classificatie van partij-affiliatie in de Tweede Kamer en wat is het resultaat van dit model?
2. In hoeverre is deze classificatie afhankelijk van partijnamen en namen van Kamerleden?
3. In hoeverre wordt deze classificatie bepaald door partij-status (d.w.z. oppositie of regering)?
4. In hoeverre wordt deze classificatie bepaald door links/rechts verdeling?
5. In hoeverre wordt deze classificatie door taalgebruik eigen aan een spreker?

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 specifieke aspecten van onderzoeken verder besproken worden.

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

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 Graeme Hirst et al. een vergelijkbaar onderzoek naar het Canadese Parlement [2]. Hierbij werd zowel gekeken naar de Engelse als Franse teksten. Een document werd hier gezien als de samenvoeging van alle spreekbeurten van een spreker. Afhankelijk van taal en dataset vinden zij in dit onderzoek nauwkeurigheden van 83.2% 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 alle classificaties kon men een F_1 score van 0.87 of hoger bereiken.

In het onderzoek van Høyland et al. werd een classificatiemodel voor partij-affiliatie op basis van teksten getraind op het vijfde Europese Parlement (1999-2004) en getest op het zesde Europese Parlement[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.

Graeme 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 resulaten opleverden. Voor classificatiemodel experimenteerden ze met SVM en Naive Bayes . Voor het selecteren van *features* experimenteerden ze met een minimale frequentie en selectie van woorden op basis van hoogste mutual information. Uiteindelijk was het beste model bij hen een SVM met uni- en bigrams en geselecteerd op basis van mutual information.

In het onderzoek van Ferreira werd gebruik gemaakt van twee classificatiemethoden: Logistische regressie en MIRA[3]. Logistische regressie werd aangevuld met *group Lasso* regularisatie. Voor wegingen van woorden werd

geëxperimenteerd met *tf*, *tf-idf*, Δ-*tf-idf* en Δ-*BM-25*. Daarnaast wordt er gebruik gemaakt van woordclustering, *Concise Semantic Analysis* en stylometrische eigenschappen. Op *Part-Of-Speech tagging* na hadden stylometrische eigenschappen een duidelijke negatieve invloed op de classificatie.

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

2.3 Invloed van partijnamen of sprekersnamen

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

2.4 Invloed van oppositie of regering

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

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

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

3 Methodologie

3.1 De data

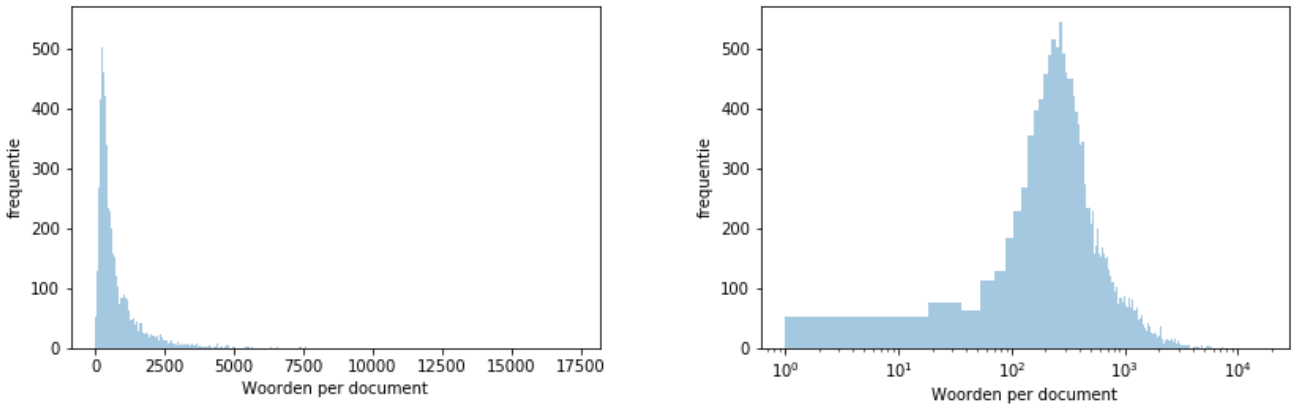
De data die gebruikt worden, zijn de Handelingen van de Tweede Kamer gedurende het missionaire kabinet-Rutte II (5 november 2012 tot 22 maart 2017). Er is gekozen voor dit kabinet, omdat de data hiervoor makkelijk verkrijgbaar was, het kabinet lang zat - waardoor er veel data is - en het recent is waardoor het makkelijker te interpreteren is. 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; debat bijdragen, interrupties en antwoorden. Debat bijdrage is de eerste onafgebroken spreekbeurt die een spreker geeft achter het 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 de 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 samengevoegd tot één document met alle bijbehorende antwoorden van die spreker. Daarnaast zijn er verschillende soorten sprekers; de voorzitter, Tweede Kamerleden, leden van het kabinet en gastsprekers. Daarnaast 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.

De documenten verschillen vervolgens in grootte. De distributie lijkt op een lognormale verdeling, maar met een Kolmogorov-Smirnov test is hier geen bewijs voor gevonden [8].



(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 het wel lognor-

maal verdeeld is en zijn daarmee de documenten buiten het betrouwbaarheidsinterval van 95% eruit gehaald. De documenten met een lengte van minimaal 28 en maximaal 1492 woorden bleven daarmee over. Het gemiddelde is daarna 498 woorden en de mediaan is 386 woorden. Een totaal aantal documenten van 14899 blijven vervolgens over.

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

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

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

3.2 Methoden

3.2.1 Deelvraag 1

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; een grid search. Deze grid search wordt gedaan door middel van 5-fold cross-validation, waarbij de trainings 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 Lasso en Ridge regularisatie, en een combinatie van beide genaamd Elastic-net. De andere parameters zijn gelaten op de standaardwaarden van scikit-learn [9]. 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[9, 6]. Hiervoor zijn de functies van scikit-learn MultinomialNB en BernoulliNB gebruikt.[9, 6]

Beoordelen van kwaliteit De meest gebruikte methoden om kwaliteit van politieke tekstclassificatie te beoordelen zijn *accuracy* en F_1 score, die opgebouwd is uit recall en precision. Deze scores zijn opgebouwd uit vier variabelen.

Deze variabelen geven weer hoeveel documenten wel of niet bij een klasse horen, en of deze wel of niet als dusdanig zijn geclassificeerd [10] .

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

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

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

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

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

3.2.2 Deelvraag 2

In Diermeier et al. [4] wordt aangenomen dat namen een groot effect hebben op de classificatie en Hirst et al. [2] 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

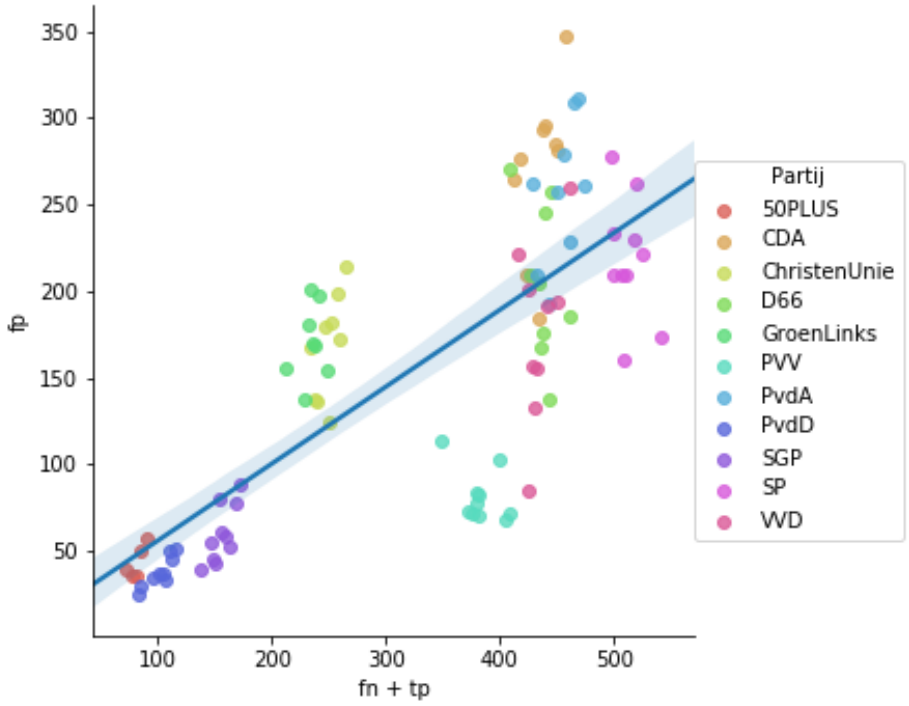
gedaan met de classificatiemethode die resulteerde uit deelvraag 1. In deze classificatie worden alle partijnamen vervangen door de tag PARTIJNAAM en alle namen van Kamerleden vervangen door de KAMERLIDNAAM. Deze namen zijn uit de Handelingen gehaald. Voor partijnamen zijn ook lidwoorden toegevoegd, 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 Deelvraag 3

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

Als er een confounding bias is op basis van partij-status, dan is te verwachten dat het aantal misclassificaties minus verwachte waarde binnen regeringspartijen en binnen oppositiepartijen hoger ligt dan tussen oppositiepartijen en regeringspartijen. Uit de voorverkenning (op basis van resultaten uit deelvraag 1 en 2) blijkt verder dat er een correlatie is tussen het aantal *false positives* van een partij en het aantal documenten behorend tot die partij.



Figuur 2: Het aantal foutief als bepaalde partij geclassificeerde documenten ten opzichte van het aantal documenten behorend tot die partij. 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 i de voorspelde partij en j 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 [11]. Om te kijken of er een confounding bias is, worden de distributies binnen regeringspartijen, binnen oppositiepartijen en tussen beide groepen met elkaar vergeleken. Om de invloed van variantie door de willekeurige splitsing documenten voor trainen en testen te beperken, wordt de classificatie 50 keer gedaan en worden deze errors bij elkaar in distributie genomen. De nulhypothese is dat er geen verschil is tussen de verdelingen. De alternatieve hypothese is dan dus dat er wel een verschil is tussen de verdelingen. Als de nulhypothese wordt verworpen, kan dus aangenomen worden dat er een verschil is op basis van partij-status.

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

In het tweede experiment uit Graeme Hirst et al. worden classifiers getraind op de ene zittingsperiode en getest op de andere zittingsperiode. Als de classificatie op basis van ideologie is in plaats van partij-status, is de verwachting dat er nog steeds aanzienlijke voorspellingen gedaan worden, aangezien de ideologie naar verwachting redelijk stabiel is binnen tien jaar (hoewel woordgebruik varieert). Als de scores aanzienlijk lager zijn, kan dit het gevolg zijn van het veranderen van partij-status van partijen.

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.

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 Deelvraag 4

Voor deze deelvraag vergelijken we de resultaten van de eerdere classificatie per partij met een binaire classificatie op basis van rechts en links. Hiervoor wordt wederom de dataset van kabinet-Rutte 2 gebruikt, met het beste model wat resulteerde uit deelvraag 1.

Voor deze vraag moet vastgesteld worden welke partijen links en rechts zijn. Omdat dit lastig te bepalen is en er meerdere indelingen zijn, wordt hier gebruik gemaakt van twee verschillende indelingen. De indeling op basis van het Kieskompas van Andre Krouwel voor de Kamerverkiezing 2012 en de indeling volgens het Manifesto Project[12] 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 Deelvraag 5

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

Om te kijken of dit effect er is, wordt er opnieuw een classificatie gedaan, maar worden de Kamerleden met al hun documenten verdeeld over de training en test set, in plaats van de individuele documenten. De meest karakteristieke woorden uit de resultaten van deelvraag 2 suggereren dat woordgebruik van Kamerleden invloed heeft (zie tabel reftab:MostImportantWords). De hypothese is daarom ook dat deze nieuwe classificatie lagere scores vindt.

4 Evaluatie

4.1 Resultaten

4.1.1 Deelvraag 1

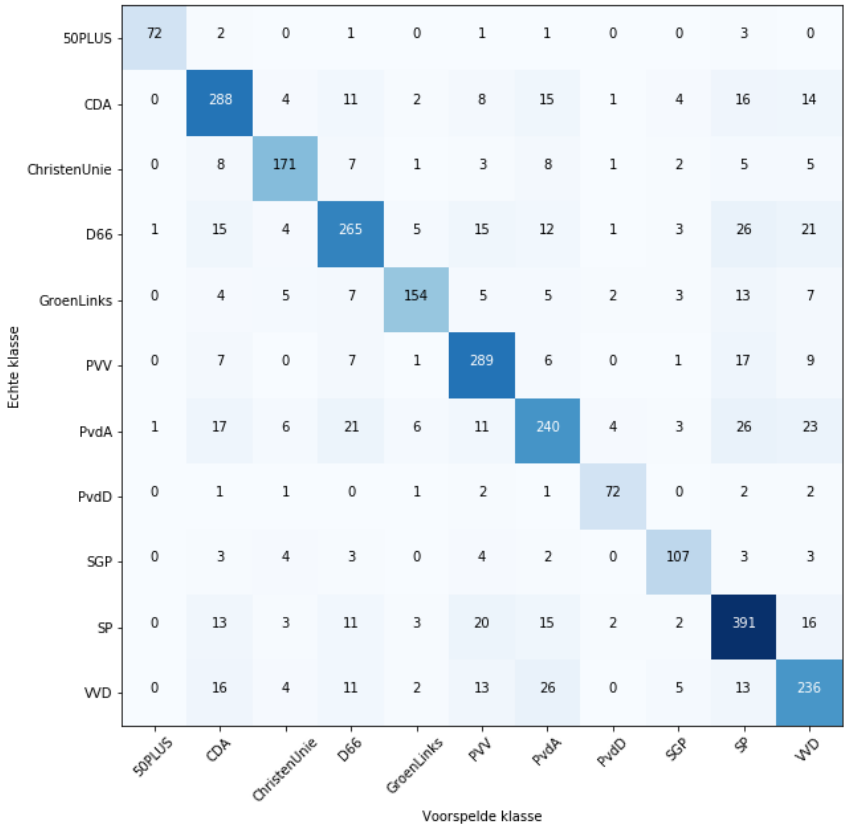
Het beste resultaat werd bereikt met Support Vector Machines gebruikmakend van *stochastic gradient descent learning* en Elasticnet regularisatie. De features waren hierbij gestemd, met unigrams, bigrams en trigrams. Geen features zijn hierin weggelaten door minimale of maximale documentfrequenties. Het verschil in scores is zeer klein, zoals te zien in figuur XXX. In bijlage YYY staan uitgebreidere figuren over het effect van de classificatiemethoden. De scores zijn ruim hoger dan de baseline scores. De scores liggen binnen de scores gevonden in gerelateerd werk, ondanks dat de baseline scores aanzienlijk lager zijn en de documentgrootte kleiner is.

Tabel 3 laat de scores zien per partij met het aantal documenten in de test set. De F_1 scores per partij liggen tussen de 0.7 en 0.8. De one-issuepartijen, 50PLUS en PvdD, hebben scores daarboven, terwijl de coalitiepartijen, VVD

en PvdA, lagere scores hebben. Figuur 3 laat zien waar de fouten in deze classificatie zitten. De meest karakteristieke features per partij zijn te zien in tabel 4. Hierin is te zien dat vrijwel alle woorden verwijzen naar de partij of een Kamerlid van die partij.

Tabel 3: Classificatierapport van beste classificatie.

Partij	Precision	Recall	F1_score	Documenten
50PLUS	0.930	0.872	0.900	82.8
CDA	0.764	0.784	0.774	367.4
ChristenUnie	0.834	0.796	0.812	215.6
D66	0.762	0.714	0.736	370.8
GroenLinks	0.862	0.736	0.794	210.0
PVV	0.778	0.846	0.810	342.6
PvdA	0.726	0.666	0.694	361.6
PvdD	0.836	0.850	0.842	85.2
SGP	0.796	0.810	0.800	132.4
SP	0.756	0.812	0.782	480.6
VVD	0.698	0.718	0.704	331.0
avg / total	0.770	0.768	0.768	2980.0



Figuur 3: Confusion matrix van beste classificatie.

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

50PLUS	CDA	ChristenUnie	D66	GroenLinks
50plus	cda	de christenunie	d66	groenlink
het lid krol	het cda	christenunie	led van veldhov	lid van tonger
lid krol	cda fractie	lid dik faber	lid van veldhov	lid voortman nar
lid krol nar	de cda fractie	lid dik	lid van men	lid voortman
krol nar mij	de cda	het lid dik	d66 is	het lid voortman
krol nar	lid omtzigt nar	de led dik	d66 wil	led van tonger
van 50plus	lid omtzigt	led dik	led van men	tonger nar mij
krol	het lid omtzigt	led dik faber	van veldhov	tonger nar
gepensioneerd	cda is	de led voordewind	veldhov	van tonger nar
fractie van 50plus	het cda is	led voordewind	led schouw	van tonger

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

PVV	PvdA	PvdD	SGP	SP	VVD
pvv	de pvda	lid ouwehand nar	sgp	sp	de vvd
de pvv	pvda	lid ouwehand	de sgp	de sp	vvd
islamitisch	pvda fractie	het lid ouwehand	sgp fractie	lid van gerv	de vvd is
lid graus	de pvda fractie	ouwehand nar mij	de sgp fractie	gerv nar mij	vvd is
het lid graus	van de arbeid	ouwehand nar	led van der	gerv nar	de vvd fractie
lid graus nar	de partij van	ouwehand	led dijkgraf	van gerv nar	vvd fractie
graus nar	de arbeid	vor de dier	de led dijkgraf	sp fractie	vor de vvd
graus nar mij	partij van de	de dier	sgp is	de sp fractie	wat de vvd
miljard	partij van	dier	de sgp is	leijt nar	vvd betreft
graf	de arbeid is	thiem	de led bisschop	leijt nar mij	de vvd betreft

4.1.2 Deelvraag 2

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

Tabel 5: Classificatierapport van beste classificatie.

Partij	Precision	Recall	F1_score	Documenten
50PLUS	0.572	0.454	0.494	75.6
CDA	0.518	0.418	0.442	458.6
ChristenUnie	0.550	0.318	0.394	239.0
D66	0.512	0.510	0.502	437.8
GroenLinks	0.624	0.216	0.314	239.4
PVV	0.548	0.776	0.640	381.2
PvdA	0.558	0.466	0.496	457.4
PvdD	0.530	0.632	0.578	96.4
SGP	0.620	0.666	0.602	157.0
SP	0.542	0.570	0.554	509.4
VVD	0.484	0.596	0.530	425.2
avg / total	0.540	0.516	0.504	3477.0

Tabel 6: Meest relevante woorden per partij op basis van classificatie uit deelvraag 1 zonder partijnamen of namen van Kamerleden gedurende kabinet-Rutte II.

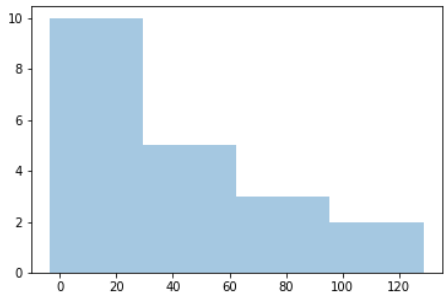
50PLUS	CDA	ChristenUnie	D66	GroenLinks
50 plusser	inwoner	gezinn	mijn fractie	schon energie
plusser	de nederland spoorweg	inderdad	hervorm	banenplan
gepensioneerd	nederland spoorweg	koerd	buitengewon	schon
koopkrachtontwikkel	spoorweg	rookvrij	daarom	in elk geval
exact	PARTIJNAAM fractie	ik constater	natur	eerlijk del
ouderenwerklos	onz inwoner	gezinn met	unido	elk geval
werkend	middeninkomen	wijsheid	kans	huishoud zorg
de 50 plusser	echt	rout	belangrijk dat	zou
50	hier	rechtsstat	vandag	kamer hierover te
vor gepensioneerd	uiteind	elkar	ding	werkgeleg

Tabel 6: Meest relevante woorden per partij op basis van classificatie uit deelvraag 1 zonder partijnamen of namen van Kamerleden gedurende kabinet-Rutte II. *(Vervolg)*

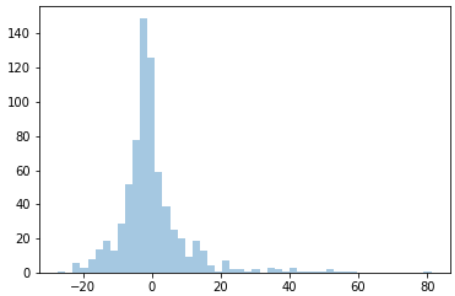
PVV	PvdA	PvdD	SGP	SP	VVD
islamitisch	mijn partij	dier	dank zer	huurder	volgen mij
islam	leerkracht	de bio	mevrouw de voorzitter	segregatie	liberal
miljard	tevred	bio industrie	mevrouw de	herindel	speelveld
de islam	circulair	bio	eenverdiener	armoed	verzekerar
asielzoeker	open standaard	aan de bio	allerlei	de bevolk	aruba
brussel	gezamen	de bio industrie	punt	jazeker	ondernemer
nederland	ieder kind	milieu	nadruk	zegt	regelgev
grenz	duurzam energie	dierenwelzijn	woord	bureaucratie	aangegev
immigratie	en	de natur	vanuit	tenderned	PARTIJNAAM is
al	lager over	klimaatverander	oog	ouderbijdrag	essentieel

4.1.3 Deelvraag 3

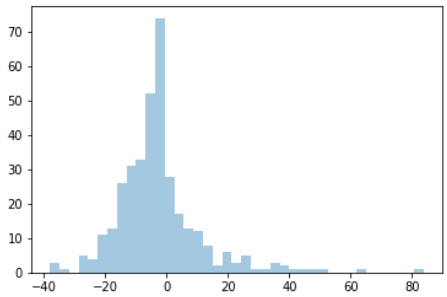
In figuur 4 zijn de distributies van de errors te zijn van combinaties tussen regerings- en oppositiepartijen.



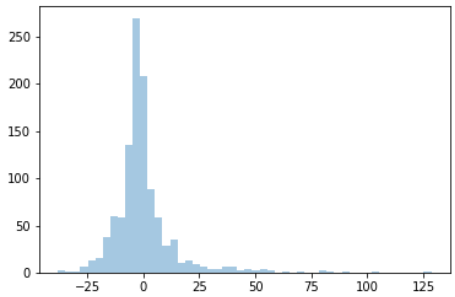
(a) Tussen twee regeringspartijen



(b) Tussen twee oppositiepartijen



(c) Tussen een regeringspartij en een oppositiepartij



(d) Totaal

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

4.2 Discussie

4.2.1 Deelvraag 1

Het onderzoek behaalt resultaten in lijn der verwachting op basis van gerelateerd en daarnaast ruim boven de baseline scores.

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. Een belangrijke hierbij is het maximaal iteraties, wat ver onder het aantal iteraties benodigd voor convergentie ligt. Voor vervolgonderzoek kan daarom dit onderdeel uitgebreid worden.

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%. Voor een vervolgonderzoek kan gekeken worden

naar of dit effect er is en wat dit betekent voor de resultaten.

4.2.2 Deelvraag 2

De resultaten laten zien dat de classificatie afhankelijk is van partijnamen en namen van Kamerleden.

De woorden in tabel 6 komen bij veel partijen overeen met hun ideologie, vooral bij PVV, PvdD en 50PLUS. Daarnaast zijn er ook woorden die niet veel over ideologie zeggen, zoals; *volgens mij*, *ik constateer* en *in elk geval*. Vooral de SGP heeft woorden die niet veel lijken te zeggen over de ideologie. 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 woorden zo karakteristiek zijn voor partijen. Een hypothese is dat deze woorden 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 woorden, 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 Deelvraag 3

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

4.2.4 Deelvraag 4

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

5 Conclusies

Referenties

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

[2] Graeme Hirst, Yaroslav Riabinin, Jory Graham, and Magali Boizot-roche. Text to ideology or text to party status? *.

[3] Vasco Ferreira. Using textual transcripts of parliamentary interventions for profiling portuguese politicians. 2016.

[4] Daniel Diermeier, Jean-François Godbout, Bei Yu, and Stefan Kaufmann. Language and ideology in congress. *British Journal of Political Science*, 42(1):31–55, 2012.

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

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

[7] Bjørn Høyland, Jean-François Godbout, Emanuele Lapponi, and Erik 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] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. [Online; accessed ¡today¿].

[9] 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.

[10] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[11] *NIST/SEMATECH e-Handbook of Statistical Methods*. NIST/SEMATECH, April 2012.

[12] 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