Classificatie op basis van partij-affiliatie in 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 bruikbaar zijn voor het bepalen van ideologische positie van andere teksten, aangezien zij zowel een tekst leveren als ook een bekende ideologie. Deze informatie kan vervolgens toegepast worden bij andere teksten die wellicht ideologisch van aard zijn. Bijvoorbeeld, aan de hand van deze informatie kan men teksten uit kranten classificeren op basis van ideologie[1, 2].

In diverse landen zijn al verschillende onderzoeken gedaan naar het classificeren van partij-affiliatie op basis van teksten van politici[3, 1]. Mede omdat elk land een andere politiek stelsel en cultuur heeft, verschillen de resultaten. Daarnaast gebruikt elk onderzoek ook een andere methode voor het classificeren.

Een onderzoek gericht op het Nederlandse parlement ontbreekt hierbij nog.

Dit onderzoek richt zich daarom op een breder scala aan mogelijke methoden en daarnaast ook 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 is deze classificatie afhankelijk van of een partij rechts of links is?

Overzicht van scriptie In sectie 2 zal gerelateerd werk besproken worden, met name vergelijkbare onderzoeken uit andere landen. In sectie 3 zal de methodologie van de verschillende deelvragen behandeld worden. In sectie 4 zullen vervolgens de resultaten weergegeven worden. In sectie 5 zal een evaluatie plaatsvinden van zowel de resultaten als de gehanteerde methodologie. In sectie 6 wordt ten slotte de onderzoeksvraag beantwoord.

2 Gerelateerd werk

Diermeier et al. deden onderzoek naar het classificeren op basis van ideologische positie in de Amerikaanse Senaat[4]. Ze trainden hun classificatie op de speeches van de 25 meest liberale en de 25 meest conservatieve senatoren van het 101e tot en met het 107e congres en testten op de 25 meest liberale en de 25 meest conservatieve senatoren van het 108e congres. Later in het onderzoek vergeleken ze ook de 25 gematigd conservatieve en de 25 gematigd liberale senatoren. Dit onderzoek wist de ideologie van de senatoren te voorspellen met een 94 procent nauwkeurigheid voor de classificatie van de extremen, maar slechts een 52 procent nauwkeurigheid voor de classificatie van de gematigde senatoren.

Als een vervolg op dit onderzoek deden Graeme Hirst et al. een vergelijkbaar onderzoek naar zowel het Canadese Parlement [2]. Hierbij werd zowel gekeken naar de Engelse als Franse teksten. Een document werd hier gezien als de samenvoeging van alle spreekbeurten van een spreker. Afhankelijk van taal en dataset, vinden zij in dit onderzoek nauwkeurigheden van 83.2 procent en hoger. In tegenstelling tot het onderzoek van Diermeier et al., vinden zij minder dat de woorden van de sprekers een uiting zijn van ideologie.

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

2.1 Classificatiemethoden

In het onderzoek van Diermeier et al. werd gebruik gemaakt van support vector machines. Verder maakten ze gebruik van *tf-idf* met een minimale woordfrequentie van 50 en een documentfrequentie van 10, *Part-Of-Speech tagging* en werden alle eigennamen verwijderd.

In het onderzoek van Graeme Hirst et al. maakten ze gebruik van support vector machines[2]. Ze experimenteerden met verschillende vormen van preprocessing, 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.

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.

In het onderzoek van Høyland et al. werd gebruik gemaakt van een multi class support vector machine[5]. 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.

2.2 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). Deze data is 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 een tabel en opgeslagen als csv-bestand.

Deze dataset bestaat uit een aantal soorten spreekbeurten, zoals speeches, interrupties en antwoorden. Daarnaast ook door verschillende soorten sprekers, zoals de voorzitter, Tweede Kamerleden, leden van het kabinet en gastsprekers. Uit deze dataset is gekozen voor de eerste spreekbeurt nadat een spreker achter het spreekgestoelte is gaan staan, aangezien deze vaak langer zijn dan de andere spreekbeurten en naar verwachting meer zeggen over ideologie. In de oorspronkelijke xml-bestanden hadden deze spreekbeurten het attribuut *nieuw="ja"*. 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 daarna naast de verkozen partijen van de 2012 Tweede Kamerverkiezingen, ook afsplitsingen van die partijen (tien in totaal) en bezoeken van vertegenwoordigingen van die partijen uit het Europees Parlement (tien in totaal). Omdat van beide categoriën relatief weinig data is en er overlap zit met hun oorspronkelijke partij, zijn deze er uit gehaald.

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

50PLUS	413
CDA	2216
ChristenUnie	1223
D66	2211
GroenLinks	1193
PVV	1880
PvdA	2269
PvdD	480
SGP	770
SP	2573
VVD	2157

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 goede resultaten hebben opgeleverd in vergelijkbare onderzoeken, zoals besproken in 2.1. Daarnaast is omwille van de tijd ervoor gekozen om alleen gebruik te maken van methoden waarvan reeds implementaties beschikbaar waren in Python. Hieronder worden de verschillende onderdelen besproken.

Pre-processing Voor pre-processing is gebruik gemaakt van tokenisation, lowercasing en stemming. Voor tokenisation is de reguliere expressie
$w+$ gebruikt, die daarmee alleen de letters van het alfabet overhoudt. Deze woorden zijn vervolgens allemaal omgezet in kleine letters. 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 zowel gekeken naar alleen unigrams als features, alsook het toevoegen van bi- en trigrams.

Support Vector Machines Een veel gebruikte techniek is Support Vector Machine (SVM). In meeste onderzoeken wordt niet gespecificeerd welke vorm van SVM gebruikt wordt. Om deze reden zal in dit onderzoek uitgebreid gekeken worden naar welke vorm het beste resultaat geeft. Hierbij wordt gebruik gemaakt van de functie SVC van sklearn en de variant met *stochastic gradient descent learning* SGDClassifier. Bij de eerste functie kan gevarieerd In het geval van de SVC functie wordt ook gevarieerd met de parameter C.

Logistische Regressie

Naive Bayes

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 correct positief (tp), foutief positief (fp), correct negatief (tn) en foutief negatief (fn) geclassificeerde waarden.

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

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

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

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

Deze waarden worden per klasse bepaald en daar wordt vervolgens een gemiddelde van genomen gewogen bij aantal positieve waarden. [6, 7].

Voor de classificatiemethoden wordt waar mogelijk gebruik gemaakt van functies van de Python module scikit-learn[7], aangevuld met zelf geschreven code als dit niet reeds beschikbaar is. Bij al deze classificatiemethoden wordt gevarieerd met meerdere parameters door middel van een gridsearch. Hierbij wordt gebruikt gemaakt van 5-fold cross-validation. Hierbij wordt de data gespleten in vijf delen, waarvan steeds één deel als test wordt gebruikt.

3.2.2 Deelvraag 2

In het onderzoek van Diermeier et al. worden alle eigennamen weggelaten zodat, volgens hen, namen van personen en partijen niet de classificatie domineren. 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 namen 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 resultaten worden vervolgens vergeleken met de resultaten uit deelvraag 1.

3.2.3 Deelvraag 3

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

Als vergelijkingsmateriaal is voor deze experiment een tweede dataset nodig uit een ander kabinet. Hiervoor is het wenselijk dat dit kabinet bestaat uit andere partijen dan kabinet-Rutte II. Er moet voor het derde experiment variatie zijn in de Kamerleden tussen de twee kabinetten, maar ook voldoende Kamerleden die in beide perioden in de kamer zaten. 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 Tweede Kamer tijdens het missionaire kabinet-Balkenende IV (22 februari 2007 tot 20 februari 2010) te gebruiken.

In het eerste experiment zullen de tien meest karakteristieke woorden per partij van het ene parlement vergeleken worden met de tien meest karakteristieke woorden per partij van het andere parlement. 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 worden classifiers getraind op het ene parlement en getest op het andere parlement. 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.

3.2.4 Deelvraag 4

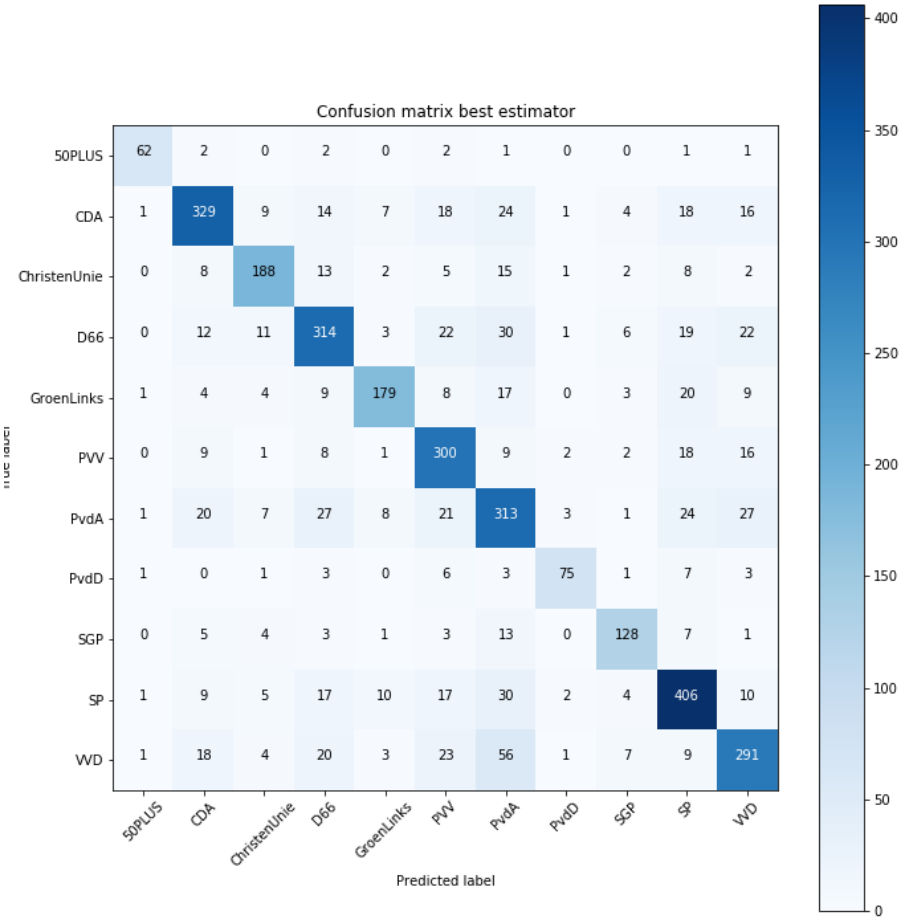
Voor deze deelvraag vergelijken we de resultaten van de eerdere classificatie per partij met een binaire classificatie op basis van rechts en links. Hiervoor wordt wederom de dataset van kabinet-Rutte 2 gebruikt, met het model wat resulteerde uit deelvraag 1.

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 gebaseerd op verkiezingsprogramma's voor de Kamerverkiezing van 2012[8]. 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

4 Evaluatie



4.1 Discussie

4.1.1 Deelvraag 1

Dit onderzoek heeft zich beperkt tot methoden genoemd in eerdere onderzoeken én waarvan de implementatie beschikbaar is in Python. Een aantal methoden die in gerelateerde literatuur leidden tot goede classificaties zijn daarom niet getest. Ook nieuwe methoden die nog niet gebruikt zijn in een gepubliceerd artikel voor politieke tekst classificatie zijn daarom niet getest. Omdat niet alle opties getest zijn, kan geen uitsluitsel gegeven worden dat dit daadwerkelijk het classificatiemodel is. Voor vervolgonderzoek kan daarom gekeken worden om meer van deze methoden mee te nemen.

4.1.2 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] 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.

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

[7] 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.

[8] 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.

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

A Slides