Methoden voor het voorspellen 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
faculty of science
university of amsterdam

Your Date of defence in the format YYYY-MM-DD

	Internal Supervisor	**Second Supervisor**
Title, Name	Dr Maarten Marx	
Affiliation	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, op basis van deze informatie kan men teksten uit kranten classificeren op basis van ideologie.

In diverse landen zijn al verschillende onderzoeken gedaan naar het classificeren van partij-affiliatie op basis van teksten van politici.[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. Daarnaast focust elk onderzoek tot nu toe op een beperkte aantal methoden, dus geen brede analyse van de mogelijke methoden.

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 partij-affiliatie te voorspellen op basis van spreekbeurten in de Tweede Kamer?"

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

1. Wat is het beste classificatiemodel voor classificatie van sprekers in de Tweede Kamer?
2. In hoeverre vindt de classificatie plaats op basis van namen van de partijen of Kamerleden?
3. In hoeverre vindt de classificatie plaats op basis van partij-status?
4. Is er een verband tussen Linkse en Rechtse partijen?

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

2.1 Deelvraag 1

Diermeier et al. deden onderzoek naar het classificeren op basis van ideologische positie in de Amerikaanse Senaat[2]. Ze trainden hun classicatie 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. Voor classificatie maakten ze gebruik 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. 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 als het Europese Parlement[3]. In dit onderzoek maken zij gebruik van support-vector machines. In tegenstelling tot het onderzoek van Diermeier et al., vinden zij minder dat de woorden van de sprekers een uiting zijn van ideologie. Daarentegen vinden zij wel een grotere invloed van oppositie tegenover regering in de woorden van de sprekers.

Ferreira probeerde interventies van politici te classificeren op basis van geslacht, leeftijdsgroep, partij-affiliatie en oriëntatie[1] In dit onderzoek maakt hij gebruik van twee classificatiemethoden, Logistische regressie en MIRA. Logistische regressie werd aangevuld met *group Lasso* regularisatie. Voor wegingen van woorden werd gebruikt gemaakt van woordfrequentie, TF-IDF, Δ-TF-IDF, 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 alle classificaties kon men aan de hand van logistische regressie en *group Lasso* regularisatie een F1-score van 0.87 of hoger bereiken.

2.2 Deelvraag 3

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).[3] Zo vergeleken 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 het zelfde 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 bevat 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 de Eerste Kamer (10 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: Spreekbeurten per partij gedurende 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 om alle classificatiemethoden te vergeleken, beperkt dit onderzoek zich tot classificatiemethoden die goede resultaten opleverden in gerelateerde werken die besproken zijn in sectie 2.1. De classificatiemethoden zullen opgesplitst worden in feature engineering en machine learning algorithms. De volgende methoden zijn hier in gekozen.

Feature Engineering

1. TF-IDF

Machine Learning Algorithms

1. Support-Vector Machines
2. Logistische Regressie

Voor de classificatiemethoden wordt waar mogelijk gebruik gemaakt van functies van de Python module sklearn, 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. De uitslagen worden beoordeeld op basis van gewogen f1-scores.

3.2.2 Deelvraag 2

3.2.3 Deelvraag 3

Om deze deelvraag te beantwoorden zullen de experimenten die Graeme Hirst et al. uitvoerden voor dezelfde deelvraag gereproduceerd worden op de Tweede Kamer. Bij deze deelvraag zal de beste classifier uit deelvraag 1 gebruikt worden. Hierbij zal naast de spreekbeurten gedurende het missionaire kabinet Rutte II, gebruikt gemaakt worden van de spreekbeurten gedurende het missionaire kabinet Balkenende IV om het effect van oppositie/regering te vergelijken. Er is voor dit parlement gekozen aangezien deze redelijk recent is (dus met redelijk overeenkomstig taalgebruik) en er een andere coalitie is met andere premier. Voor dit experiment was het beter als er een kabinet was zonder de partijen van kabinet Rutte II, aangezien er dan minder overlap is, maar dit is sinds de Tweede Oorlog niet langer dan een jaar gebeurd.

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.

In het derde experiment zullen twee classificaties vergeleken worden. De eerste op Kamerleden die in beide parlementen zaten en een classificatie op Kamerleden die maar in één van de twee parlementen hebben gezeten.

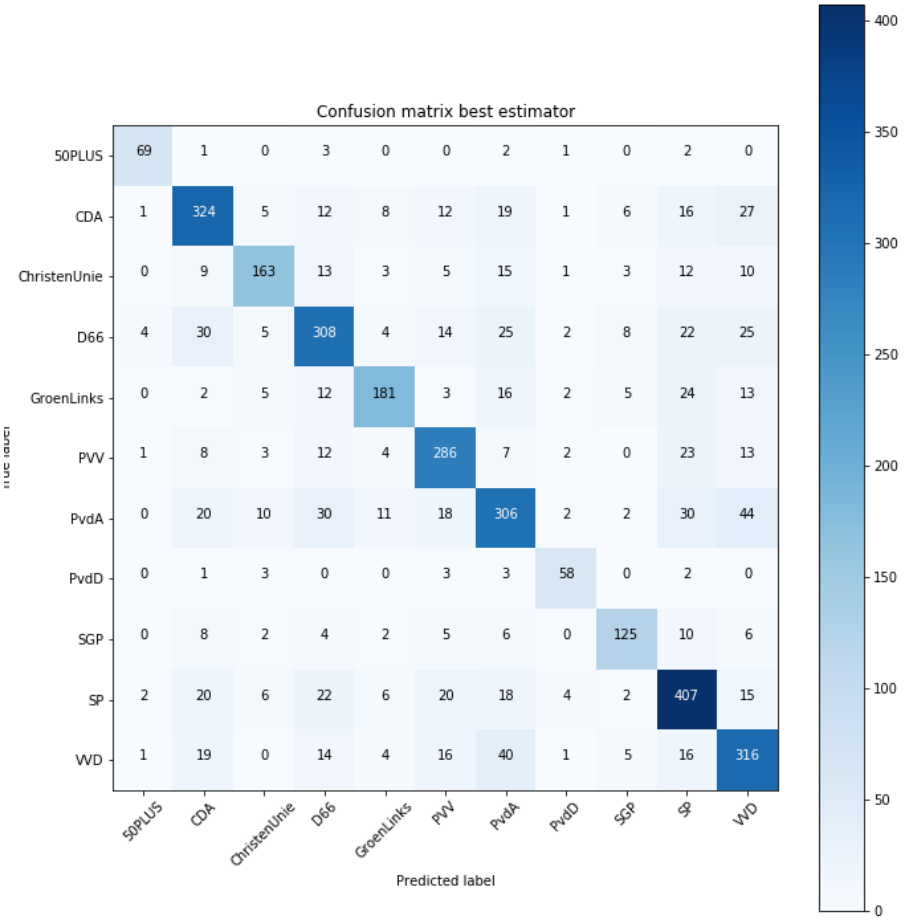
4 Evaluatie

Met een subsectie voor elke deelvraag.

In hoeverre is je vraag beantwoord?

Een mooie graphic/visualisatie is hier heel gewenst.

Hou het kort maar krachtig.



5 Conclusies

Hierin beantwoord je jouw hoofdvraag op basis van het eerder vergaarde bewijs.

Referenties

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

[2] 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.

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

A Slides