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

Todo list

fix inleiding

Inhoudsopgave

Samenvatting

1 Introduction

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: "Wat is het beste classificatiemodel voor het classificeren van sprekers in de Tweede Kamer op basis van partij-affiliatie?"

In dit onderzoek zal eerst gekeken worden naar welke methoden gangbaar zijn in vergelijkbare onderzoeken, maar ook naar welke methoden nieuw en potentieel zijn. Deze worden vervolgens geëvalueerd en vergeleken, in de hoop hiermee de onderzoeksvraag te kunnen beantwoorden.

Overview of thesis In sectie 2 zullen vergelijkbare onderzoeken uit andere landen besproken worden. In sectie 3 zal vervolgens de wijze waarop de verschillende classificatiemethoden gebruikt zijn als ook geëvalueerd zijn besproken 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 Related Work

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 minimum 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.

2.1 RQ1

2.2 RQ2

3 Methodology

3.1 Description of the data

De data die gebruikt wordt, zijn de Handelingen van de Tweede Kamer gedurende het niet-demissionaire kabinet Rutte 2 (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 en inhoud van de 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

Partij	Aantal spreekbeurten
SP	27034
D66	24600
VVD	22990
CDA	22452
PvdA	22217
PVV	16408
GroenLinks	12954
ChristenUnie	11401
SGP	6316
PvdD	4081
50PLUS	2223

3.2 Wat plotjes en tabelletjes

Zie het IPython Notebook `PandasAndLatex.ipynb` voor de code om vanuit pandas een poltje op te slaan en een dataframe als tabel op te slaan. Het werkt ideaal!

De interrupties van Wilders staan beschreven in Figure **??** en Tabel **??**.

3.3 Methods

Hoe je je vraag gaat beantwoorden.

Dit is de langste sectie van je scriptie.
Als iets erg technisch wordt kan je een deel naar de Appendix verplaatsen.
Probeer er een lopend verhaal van te maken.
Het is heel handig dit ook weer op te delen nav je deelvragen:

3.3.1 RQ1

3.3.2 RQ2

4 Evaluation

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 Conclusions

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

5.1 Acknowledgements

Hier kan je bedanken wie je maar wilt.

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