

1 CLASSIFICATIE OP BASIS VAN PARTIJ-AFFILIATIE IN
2 DE TWEEDE KAMER

3 SUBMITTED IN PARTIAL FULFILLMENT FOR THE DEGREE OF
4 BACHELOR OF SCIENCE

5 JASPER VAN DER HEIDE
6 10732721

7 BACHELOR INFORMATIEKUNDE
8 FACULTEIT DER NATUURWETENSCHAPPEN, WISKUNDE EN
9 INFORMATICA
10 UNIVERSITEIT VAN AMSTERDAM
11 2018-06-28

12

	Begeleider	Tweede lezer
Titel, Naam	Dr Maarten Marx	
Affiliatie	UvA, FNWI, IvI	
Email	maartenmarx@uva.nl	.



UNIVERSITEIT VAN AMSTERDAM

14	Inhoudsopgave	
15	1 Introductie	3
16	2 Gerelateerd werk	3
17	2.1 Classificatiemethoden	4
18	2.2 Invloed van oppositie of regering	5
19	3 Methodologie	5
20	3.1 De data	5
21	3.2 Methoden	6
22	3.2.1 Deelvraag 1	6
23	3.2.2 Deelvraag 2	8
24	3.2.3 Deelvraag 3	8
25	3.2.4 Deelvraag 4	8
26	4 Evaluatie	10
27	4.1 Discussie	10
28	4.1.1 Deelvraag 1	10
29	4.1.2 Deelvraag 4	10
30	5 Conclusies	11
31	A Slides	11

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.

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)[5]. Zij vonden hiervoor uiteindelijk een F_1 score van 0.684647.

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[6]. 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 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 wegen voor woorden: *boolean*, *tf*, *tf-norm* en *tf-idf*, waarvan *tf-idf* het beste resultaat opleverde.

In het onderzoek van Bhand et al. gebruikten ze verschillende n-grams, inclusief verschillende manieren van *smoothing*[5]. Zij gebruikte als weging altijd de aanwezigheid van een woord. Als classificatiemodellen experimenteerden ze support vector machines en naïve bayes classificatie. Voor het selecteren van *features* experimenteerden ze met een simpele minimale frequentie en het gebruik van een top aantal woorden op basis van mutual information. Uiteindelijk was het beste model bij hen een met support vector machine, met uni- en bigrams, gekozen 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 wegen 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[6]. Als beste waarde voor de regularisatieterm,

124 de C-parameter, vonden zij 0.8. Daarnaast gebruikten zij *dependency disambi-*
125 *guated stems* wat bij hen een F_1 score van twee procent hoger opleverden dan
126 normale stemming.

127 2.2 Invloed van oppositie of regering

128 Graeme Hirst et al. vonden in hun onderzoek dat de classificatie van spreker in
129 het Canadese parlement op basis van partij-affiliatie meer zegt over de status
130 van de partij (regering of oppositie).[2] Zo vergeleken zij de top tien karakteris-
131 tieke woorden van de liberalen en conservatieven in het 36e parlement (liberalen
132 in de regering) en het 39e parlement (conservatieven in de regering. Hier vonden
133 zij dat vier van de tien woorden van de liberalen (regering) in het 36e parlement
134 bij het 39e parlement bij de conservatieven (regering) te vinden waren. An-
135 dersom gebeurde hetzelfde met één van de tien woorden van de conservatieven
136 (oppositie) in het 36e parlement naar liberalen (oppositie) in het 39e parlement.

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

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

146 3 Methodologie

147 3.1 De data

148 De data die gebruikt worden, zijn de Handelingen van de Tweede Kamer gedu-
149 rende het missionaire kabinet-Rutte II (5 november 2012 tot 22 maart 2017).
150 Er is gekozen voor dit kabinet, omdat de data hiervoor makkelijk verkrijgbaar
151 was, het kabinet lang zat en dus veel data bevat en het recent is en daardoor
152 makkelijker is om te interpreteren. Deze data is in xml-formaat van de web-
153 site officielebekendmakingen.nl gehaald, samen met corresponderende metadata
154 xml-bestanden. De bestanden van de Handelingen bevatten voornamelijk infor-
155 matie over spreekbeurten tijdens een debat, waaronder naam van een spreker,
156 partij-affiliatie, inhoud van de spreekbeurt en het soort spreekbeurt. Deze ge-
157 gevens zijn samengevoegd tot een tabel en opgeslagen als csv-bestand.

158 Deze dataset bestaat uit een aantal soorten spreekbeurten, zoals speeches,
159 interrupties en antwoorden. Daarnaast ook door verschillende soorten sprekers,
160 zoals de voorzitter, Tweede Kamerleden, leden van het kabinet en gastsprekers.
161 Uit deze dataset is gekozen voor de eerste spreekbeurt nadat een spreker achter
162 het spreekgestoelte is gaan staan, aangezien deze vaak langer zijn dan de andere
163 spreekbeurten en naar verwachting meer zeggen over ideologie. In de oorspron-
164 kelijke xml-bestanden hadden deze spreekbeurten het attribuut *nieuw="ja"*.
165 Daarnaast is alleen gekozen voor sprekers waarvan er een partij-affiliatie ver-
166 meld staat, dit is niet het geval voor leden van het kabinet, de voorzitter en

167 gastsprekers (met uitzondering van Nederlandse leden van het Europees Parle-
168 ment).

169 Deze dataset bevat daarna naast de verkozen partijen van de 2012 Tweede
170 Kamerverkiezingen, ook afsplitsingen van die partijen (tien in totaal) en bezoeken van
171 vertegenwoordigingen van die partijen uit het Europees Parlement (tien in totaal). Omdat van beide
172 categorieën relatief weinig data is en er overlap zit met hun oorspronkelijke partij, zijn deze er uit
173 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

174 3.2 Methoden

175 3.2.1 Deelvraag 1

176 Om deze deelvraag te beantwoorden zullen een aantal classificatiemethoden ver-
177 geleken worden. Aangezien het onmogelijk is om alle classificatiemethoden te
178 vergelijken, beperkt dit onderzoek zich tot classificatiemethoden die goede re-
179 sultaten hebben opgeleverd in vergelijkbare onderzoeken, zoals besproken in
180 2.1. Daarnaast is omwille van de tijd ervoor gekozen om alleen gebruik te ma-
181 ken van methoden waarvan reeds implementaties beschikbaar waren in Python.
182 Hieronder worden de verschillende onderdelen besproken.

183 **Pre-processing** Voor pre-processing is gebruik gemaakt van tokenisation, lo-
184 wercasing en stemming. Voor tokenisation is de reguliere expressie
185 $w+$ gebruikt, die daarmee alleen de letters van het alfabet overhoudt. Deze
186 woorden zijn vervolgens allemaal omgezet in kleine letters. Vervolgens is er
187 gevarieerd tussen wel of geen gebruik maken van stemming. In het geval van
188 stemming is gebruik gemaakt van de Snowball Stemmer via de Python NLTK
189 module.

190 **Bag-of-words model** Bag-of-words model is de meest gebruikte representa-
191 tie van data in vergelijkbare onderzoeken. Bij het bag-of-words model wordt
192 elk document gerepresenteerd door een vector, waarbij elke kolom een woord
193 voorstelt met een bijbehorende waarde. Voornaamste beperking van dit model
194 is dat het geen rekening houdt met de volgorde van woorden, wat een groot
195 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 n-grams en zijn combinaties van unigrams, bigrams en trigrams vergeleken.

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 worden met de kernels, bij de tweede functie is deze lineair. Voor de SVC functie zijn een aantal waarden voor de C-parameter gebruikt, waaronder de standaardwaarde 1 van sklearn en 0.8 gevonden door Høyland et al.[6, 7]. Omdat de classificaties vrij lang duren met zoveel variaties, is er voor gekozen voor een maximaal aantal iteraties van vijf.

Logistische Regressie

Naïve 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 het aantal correct positief (*tp*), foutief positief (*fp*), correct negatief (*tn*) en foutief negatief (*fn*) geclassificeerde waarden.

$$Precision = \frac{tp}{tp + fp} \quad (1)$$

$$Recall = \frac{tp}{tp + tn} \quad (2)$$

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \quad (3)$$

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (4)$$

Deze waarden worden per klasse bepaald en daar wordt vervolgens een gemiddelde van genomen gewogen bij aantal positieve waarden. [8, 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.

231 3.2.2 Deelvraag 2

232 In het onderzoek van Diermeier et al. worden alle eigennamen weggelaten zo-
233 dat, volgens hen, namen van personen en partijen niet de classificatie domineren.
234 Aangezien hier bij deelvraag 1 niet voor is gekozen, wordt bij deze deelvraag ge-
235 keken hoe groot het effect hiervan is, specifiek gericht op partijnamen en namen
236 van kamerleden. Voor deze deelvraag wordt wederom een classificatie gedaan
237 met de classificatiemethode die resulteerde uit deelvraag 1. In deze classificatie
238 worden alle partijnamen vervangen door de tag PARTIJNAAM en alle namen
239 van Kamerleden vervangen door de KAMERLIDNAAM. Deze resultaten wor-
240 den vervolgens vergeleken met de resultaten uit deelvraag 1.

241 3.2.3 Deelvraag 3

242 Om deze deelvraag te beantwoorden zullen de twee experimenten die Graeme
243 Hirst et al. uitvoerden voor dezelfde vraag gereproduceerd worden op de dataset
244 van de Tweede Kamer. Bij deze deelvraag zal de beste classifier uit deelvraag
245 1 gebruikt worden.

246 Als vergelijkingsmateriaal is voor deze experiment een tweede dataset no-
247 dig uit een ander kabinet. Hiervoor is het wenselijk dat dit kabinet bestaat
248 uit andere partijen dan kabinet-Rutte II. Er moet voor het derde experiment
249 variatie zijn in de Kamerleden tussen de twee kabinetten, maar ook voldoende
250 Kamerleden die in beide perioden in de kamer zaten. Daarnaast is het ook
251 wenselijk als het niet te ver terug is, zodat onderwerpen en taalgebruik enig-
252 zins overeenkomstig zijn. Omdat kabinet-Rutte I een minderheidskabinet was
253 met een bijzondere partij-status voor de PVV, is ervoor gekozen om de Tweede
254 Kamer tijdens het missionaire kabinet-Balkenende IV (22 februari 2007 tot 20
255 februari 2010) te gebruiken.

256 In het eerste experiment zullen de tien meest karakteristieke woorden per
257 partij van het ene parlement vergeleken worden met de tien meest karakteristieke
258 woorden per partij van het andere parlement. Als de classificatie op basis van
259 ideologie is in plaats van partij-status, is het te verwachten dat de woorden bij
260 een partij blijven en niet gekoppeld zijn aan in oppositie of regering zitten.

261 In het tweede experiment worden classifiers getraind op het ene parlement
262 en getest op het andere parlement. Als de classificatie op basis van ideologie
263 is in plaats van partij-status, is de verwachting dat er nog steeds aanzienlijke
264 voorspellingen gedaan worden, aangezien de ideologie naar verwachting redelijk
265 stabiel is binnen tien jaar (hoewel woordgebruik varieert). Als de scores aan-
266 zienlijk lager zijn, kan dit het gevolg zijn van het veranderen van partij-status
267 van partijen.

268 3.2.4 Deelvraag 4

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

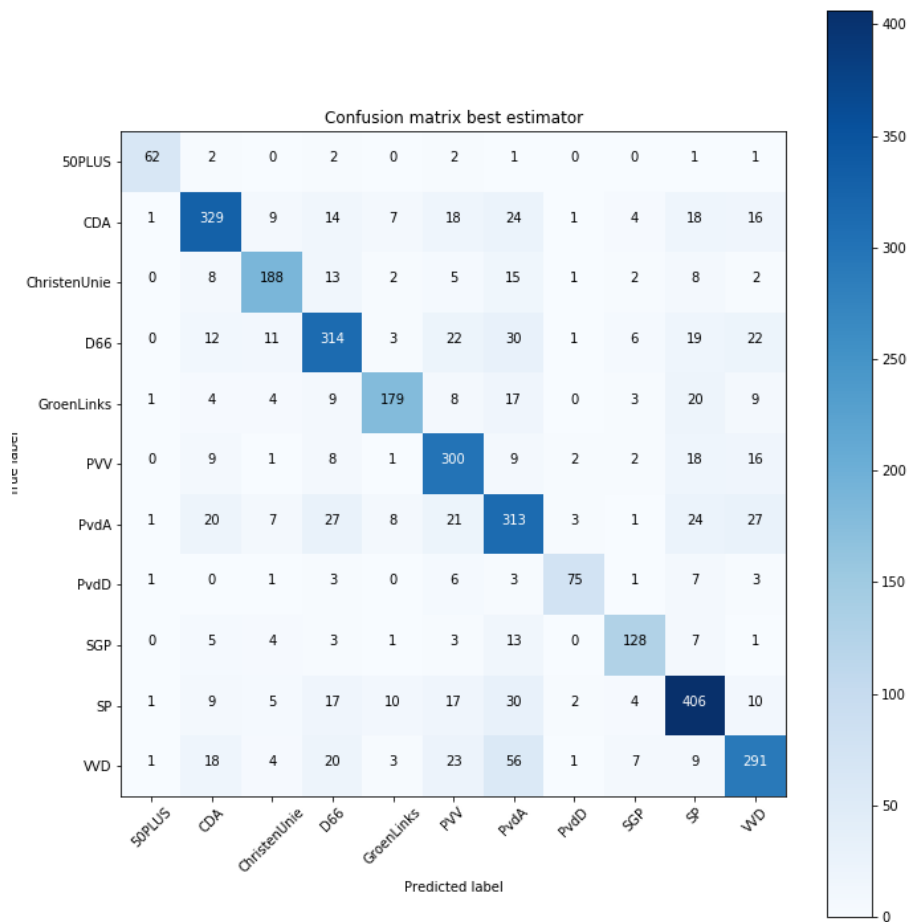
273 Voor deze vraag moet vastgesteld worden welke partijen links en rechts
274 zijn. Omdat dit lastig te bepalen is en er meerdere indelingen zijn, wordt hier
275 gebruik gemaakt van twee verschillende indelingen. De indeling op basis van het
276 Kieskompas van Andre Krouwel voor de Kamerverkiezing 2012 en de indeling

277 volgens het Manifesto Project[9] gebaseerd op verkiezingsprogramma's voor de
 278 Kamerverkiezing van 2012. In beide gevallen is de nullijn van het politieke
 279 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



281

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.

294

295 5 Conclusies

296 Referenties

- 297 [1] Felix Bießmann. Automating political bias prediction. *CoRR*,
298 abs/1608.02195, 2016.
- 299 [2] Graeme Hirst, Yaroslav Riabinin, Jory Graham, and Magali Boizot-roche.
300 Text to ideology or text to party status? *.
- 301 [3] Vasco Ferreira. Using textual transcripts of parliamentary interventions for
302 profiling portuguese politicians. 2016.
- 303 [4] Daniel Diermeier, Jean-François Godbout, Bei Yu, and Stefan Kaufmann.
304 Language and ideology in congress. *British Journal of Political Science*,
305 42(1):31–55, 2012.
- 306 [5] Conal Sathi Maneesh Bhand, Dan Robinson. Text classifiers for political
307 ideologies, 2009.
- 308 [6] Bjørn Høyland, Jean-François Godbout, Emanuele Lapponi, and Erik Vell-
309 dal. Predicting party affiliations from european parliament debates. In
310 *Proceedings of the ACL 2014 Workshop on Language Technologies and*
311 *Computational Social Science*, pages 56–60. Association for Computatio-
312 nal Linguistics, 2014.
- 313 [7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Gri-
314 sel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas,
315 A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay.
316 Scikit-learn: Machine learning in Python. *Journal of Machine Learning*
317 *Research*, 12:2825–2830, 2011.
- 318 [8] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Intro-*
319 *duction to Information Retrieval*. Cambridge University Press, New York,
320 NY, USA, 2008.
- 321 [9] Andrea Volkens, Pola Lehmann, Theres Matthieß, Nicolas Merz, Sven Re-
322 gel, and Bernhard Weßels. The manifesto data collection. manifesto project
323 (mrg/cmp/marpor). version 2017b, 2017.
- 324 [10] Bei Yu, Stefan Kaufmann, and Daniel Diermeier. Classifying party affilia-
325 tion from political speech. *Journal of Information Technology & Politics*,
326 5(1):33–48, 2008.

327 A Slides