

1 IDEOLOGIE EN CLASSIFICATIE IN DE HANDELINGEN
2 VAN DE TWEEDE KAMER
3
4 SUBMITTED IN PARTIAL FULFILLMENT FOR THE DEGREE OF
5 BACHELOR OF SCIENCE
6
7 JASPER VAN DER HEIDE
8 10732721
9
10 BACHELOR INFORMATIEKUNDE
11 FACULTEIT DER NATUURWETENSCHAPPEN, WISKUNDE EN
INFORMATICA
UNIVERSITEIT VAN AMSTERDAM
2018-06-28

12

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



| | | |
|----|--|-----------|
| 14 | Inhoudsopgave | |
| 15 | 1 Introductie | 3 |
| 16 | 2 Gerelateerd werk | 4 |
| 17 | 2.1 Tekstclassificatie van parlementaire teksten | 4 |
| 18 | 2.2 Classificatiemethoden | 4 |
| 19 | 2.3 Invloed van partijnamen of sprekersnamen | 5 |
| 20 | 2.4 Invloed van oppositie of regering | 5 |
| 21 | 3 Methodologie | 6 |
| 22 | 3.1 De data | 6 |
| 23 | 3.2 Methoden | 8 |
| 24 | 3.2.1 Deelvraag 1 | 8 |
| 25 | 3.2.2 Deelvraag 2 | 10 |
| 26 | 3.2.3 Deelvraag 3 | 10 |
| 27 | 3.2.4 Deelvraag 4 | 12 |
| 28 | 4 Evaluatie | 13 |
| 29 | 4.1 Resultaten | 13 |
| 30 | 4.1.1 Deelvraag 1 | 13 |
| 31 | 4.1.2 Deelvraag 2 | 15 |
| 32 | 4.1.3 Deelvraag 3 | 16 |
| 33 | 4.2 Discussie | 17 |
| 34 | 4.2.1 Deelvraag 1 | 17 |
| 35 | 4.2.2 Deelvraag 2 | 17 |
| 36 | 4.2.3 Deelvraag 4 | 17 |
| 37 | 5 Conclusies | 17 |
| 38 | A Slides | 18 |

39

Samenvatting

40

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. 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 tekst classificatie proberen onderzoekers ideologie uit te vinden in hoeverre ideologie terug te vinden is in teksten van politici. Mede omdat elk land een ander stelsel, taal en cultuur heeft, verschillen de resultaten. Elk onderzoek gebruikt ook een andere methode voor het classificeren. Daarnaast vindt het onderzoek van Hirst et al. [2] dat deze classificatie minder het gevolg is van ideologie maar meer van regering tegenover oppositie. Deze onderzoeken worden besproken in het gerelateerd werk.

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?

Daarom zal eerst gekeken worden naar classificatiemethoden en resultaten in vergelijkbare onderzoeken. Van deze classificatiemethoden zullen een aantal toegepast worden op teksten van de Tweede Kamer. Vervolgens zal door middel van de overige deelvragen bepaald worden in hoeverre dit een reflectie is van ideologie.

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.

82 2 Gerelateerd werk

83 Toespraken in parlementen worden veel gebruikt in tekstclassificatie, omdat
84 deze veel nette tekst bevatten en vaak gelabeld zijn. Labels zijn bijvoorbeeld
85 naam en partij van de spreker, maar ook daar uit afleidbare labels zoals geslacht,
86 leeftijd en partij-status (oppositie of regering).

87 In dit hoofdstuk zullen verschillende onderzoeken behandeld worden die
88 tekstclassificatie hebben toegepast op parlementaire teksten. Eerst zullen de
89 onderzoeken algemeen besproken worden. Vervolgens zal uitgebreider gekeken
90 worden naar de effecten van verschillende classificatiemethoden. In de latere
91 secties zullen specifieke aspecten van onderzoeken verder besproken worden.

92 2.1 Tekstclassificatie van parlementaire teksten

93 Diermeier et al. deden onderzoek naar het classificeren op basis van ideologische
94 positie in de Amerikaanse Senaat[4]. Ze trainden hun classificatie op de speeches
95 van de 25 meest liberale en de 25 meest conservatieve senatoren van het 101e
96 tot en met het 107e congres en testten op dezelfde categorieën van het 108e
97 congres. Een document was in dit onderzoek de verzameling van alle speeches
98 van een senator in een congres. De baseline nauwkeurigheid is hierdoor 50%.
99 Deze classificatie resulteerde uiteindelijk in een nauwkeurigheid van 94%.

100 Later in het onderzoek vergeleken ze ook de 25 gematigd conservatieve en
101 de 25 gematigd liberale senatoren van dezelfde congressen. De baseline nauw-
102 keurigheid was opnieuw 50% en het resultaat uiteindelijk 52%, dus niet veel
103 hoger dan gokken.

104 Als een vervolg op dit onderzoek deden Graeme Hirst et al. een vergelijk-
105 baar onderzoek naar het Canadese Parlement [2]. Hierbij werd zowel gekeken
106 naar de Engelse als Franse teksten. Een document werd hier gezien als de sa-
107 menvoeging van alle spreekbeurten van een spreker. Afhankelijk van taal en
108 dataset vinden zij in dit onderzoek nauwkeurigheden van 83.2 procent en hoger.

109 Het onderzoek bevat ook een classificatie van het Europees Parlement.
110 Hierbij voegen ze alle teksten van een parlamentslid bij elkaar en delen die op in
111 stukken van gelijke grootte. Zij vinden voor documentgrootte van 267 woorden
112 een nauwkeurigheid van 44,0% oplopend tot 61,8% voor documentgrootte van
113 6666.

114 Het onderzoek van Bhand et al. richtte zich op het classificeren van leden
115 van het Amerikaanse congres in 2005, op basis van affiliatie (Republikeins of
116 Democratisch)[5]. Zij vonden hiervoor uiteindelijk een F_1 score van 0.684647.

117 Ferreira probeerde interventies van politici te classificeren op basis van
118 geslacht, leeftijdsgroep, partij-affiliatie en oriëntatie in het Portugese parlement
119 [3]. In alle classificaties kon men een F_1 score van 0.87 of hoger bereiken.

120 In het onderzoek van Høyland et al. werd een classificatiemodel voor
121 partij-affiliatie op basis van teksten getraind op het vijfde Europese Parlement
122 (1999-2004) en getest op het zesde Europese Parlement[6]. Hier verkregen zij
123 een *macro average* F_1 score van 0.464.

124 2.2 Classificatiemethoden

125 In het onderzoek van Diermeier et al. werd gebruik gemaakt van support vector
126 machines. Verder maakten ze gebruik van *tf-idf* met een minimale woordfre-

127 quentie van 50 en een minimale documentfrequentie van 10, *Part-Of-Speech*
128 *tagging* en werden alle eigennamen verwijderd.

129 In het onderzoek van Graeme Hirst et al. maakten ze gebruik van support
130 vector machines[2]. Ze experimenteerden met verschillende vormen van pre-
131 processing, inclusief stemmen en het verwijderen van woorden op basis van
132 te hoge of te lage frequentie. Deze variaties maakten in hun onderzoek geen
133 grote verschillen en uiteindelijk is gekozen voor het niet stemmen, het weglaten
134 van woorden die in minder dan vijf documenten voorkomen en resultaten van
135 zowel met als zonder de top 500 meest frequente woorden. Daarnaast werd
136 geëxperimenteerd met vier wegingen voor woorden: *boolean*, *tf*, *tf-norm* en *tf-*
137 *idf*, waarvan *tf-idf* het beste resultaat opleverde.

138 In het onderzoek van Bhand et al. gebruikten ze verschillende n-grams,
139 inclusief verschillende manieren van *smoothing*[5]. Zij gebruikte als weging al-
140 tijd de aanwezigheid van een woord. Als classificatiemodellen experimenteerden
141 ze support vector machines en naive bayes classificatie. Voor het selecteren
142 van *features* experimenteerden ze met een simpele minimale frequentie en het
143 gebruik van een top aantal woorden op basis van mutual information. Uitein-
144 delijk was het beste model bij hen een met support vector machine, met uni- en
145 bigrams, gekozen op basis van mutual information.

146 In het onderzoek van Ferreira werd gebruik gemaakt van twee classifi-
147 catiemethoden: Logistische regressie en MIRA[3]. Logistische regressie werd
148 aangevuld met *group Lasso* regularisatie. Voor wegingen van woorden werd
149 geëxperimenteerd met *tf*, *tf-idf*, Δ -*tf-idf* en Δ -*BM-25*. Daarnaast wordt er ge-
150 bruik gemaakt van woordclustering, *Concise Semantic Analysis* en stylome-
151 trische eigenschappen. Op *Part-Of-Speech tagging* na hadden stylometrische
152 eigenschappen een duidelijke negatieve invloed op de classificatie.

153 In het onderzoek van Høyland et al. werd gebruik gemaakt van een multi
154 class support vector machine[6]. Als beste waarde voor de regularisatieterm,
155 de C-parameter, vonden zij 0.8. Daarnaast gebruikten zij *dependency disambi-*
156 *guated stems* wat bij hen een F_1 score van twee procent hoger opleverden dan
157 normale stemming.

158 2.3 Invloed van partijnamen of sprekersnamen

159 In het onderzoek van Diermeier et al. zijn alle namen weggelaten, omdat deze
160 volgens hen de classificatie te makkelijk zouden maken [4]. Hirst et al. vinden in-
161 derdaad dat partijnamen (en het weglaten daarvan) bij het Europees Parlement
162 een grote invloed hebben op de classificatie [2]. Bij het Europees Parlement zien
163 zij met name het gebruik van de eigen partijnaam door een spreker, terwijl zij in
164 het Canadese parlement vooral zien dat de naam van de andere partij gebruikt
165 wordt door een spreker.

166 2.4 Invloed van oppositie of regering

167 Graeme Hirst et al. vonden in hun onderzoek dat de classificatie van spreker in
168 het Canadese parlement op basis van partij-affiliatie meer zegt over de status
169 van de partij (regering of oppositie).[2] Zo vergeleken zij de top tien karakteris-
170 tieke woorden van de liberalen en conservatieven in het 36e parlement (liberalen
171 in de regering) en het 39e parlement (conservatieven in de regering. Hier vonden
172 zij dat vier van de tien woorden van de liberalen (regering) in het 36e parlement

173 bij het 39e parlement bij de conservatieven (regering) te vinden waren. An-
174 dersom gebeurde hetzelfde met één van de tien woorden van de conservatieven
175 (oppositie) in het 36e parlement naar liberalen (oppositie) in het 39e parlement.

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

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

185 3 Methodologie

186 3.1 De data

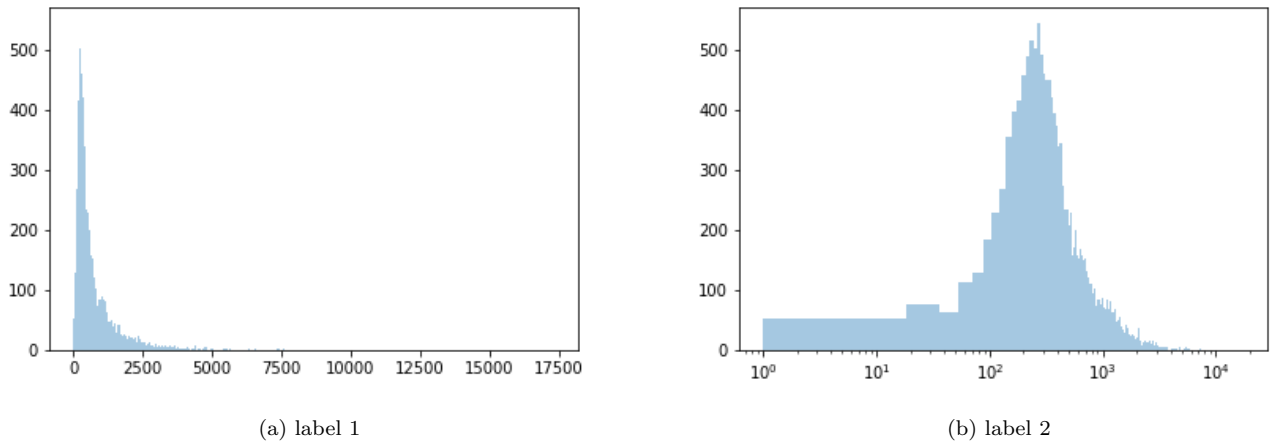
187 De data die gebruikt worden, zijn de Handelingen van de Tweede Kamer gedu-
188 rende het missionaire kabinet-Rutte II (5 november 2012 tot 22 maart 2017).
189 Er is gekozen voor dit kabinet, omdat de data hiervoor makkelijk verkrijgbaar
190 was, het kabinet lang zat, waardoor er veel data is, en het recent is waardoor het
191 makkelijker te interpreteren is. Deze data zijn in xml-formaat van de website
192 officiële bekendmakingen.nl gehaald samen met corresponderende metadata xml-
193 bestanden. De bestanden van de Handelingen bevatten voornamelijk informatie
194 over spreekbeurten tijdens een debat, waaronder naam van een spreker, partij-
195 affiliatie, inhoud van de spreekbeurt en het soort spreekbeurt. Deze gegevens
196 zijn samengevoegd tot één tabel.

197 Deze dataset bestaat uit een aantal soorten spreekbeurten; debat bijdra-
198 gen, interrupties en antwoorden. Debat bijdrage is de eerste onafgebroken
199 spreekbeurt die een spreker geeft achter het spreekgestoelte, aangeduid in de
200 xml-file met het attribuut *nieuw="ja"*. Dit kan een bijdrage in een debat zijn
201 of een vraag tijdens een vragenuur. Interrupties zijn de vragen die andere po-
202 litici stellen vanachter de interruptiemicrofoon aan de spreker. De antwoorden
203 zijn vervolgens de reactie van een spreker achter het spreekgestoelte op een in-
204 terruptie. Aangezien een debat bijdrage geïnterrupteerd kan worden, kan deze
205 inhoudelijk doorlopen in een antwoord van een spreker. Er is in dit onderzoek
206 ervoor gekozen om gebruik te maken van een debat bijdrage samengevoegd tot
207 één document met alle bijbehorende antwoorden van die spreker. Daarnaast zijn
208 er verschillende soorten sprekers; de voorzitter, Tweede Kamerleden, leden van
209 het kabinet en gastsprekers. Daarnaast is alleen gekozen voor sprekers waarvan
210 er een partij-affiliatie vermeld staat, dit is niet het geval voor leden van het
211 kabinet, de voorzitter en gastsprekers (met uitzondering van Nederlandse leden
212 van het Europees Parlement).

213 Deze dataset bevat vervolgens naast de verkozen partijen van de 2012
214 Tweede Kamerverkiezingen, ook afsplitsingen van die partijen (tien in totaal) en
215 bezoeken van vertegenwoordigingen van Nederlandse partijen uit het Europees
216 Parlement (tien in totaal). Omdat van beide categorieën relatief weinig data is
217 en er overlap zit met hun oorspronkelijke partij, zijn deze er uit gehaald.

218 De documenten verschillen vervolgens in grootte. De distributie lijkt op

219 een lognormale verdeling, maar met een Kolmogorov-Smirnov test is hier geen
 220 bewijs voor gevonden [7].



Figuur 1: Aantal woorden per document

221 Om toch de uitschieters er uit te halen, is aangenomen dat het wel lognor-
 222 maal verdeeld is en zijn daarmee de documenten buiten het betrouwbaarheids-
 223 interval van 95% eruit gehaald. De documenten met een lengte van minimaal
 224 28 en maximaal 1492 woorden bleven daarmee over. Het gemiddelde is daarna
 225 498 woorden en de mediaan is 386 woorden. Een totaal aantal documenten van
 226 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 |

227 Deze 14899 documenten zijn verdeeld over 2984 debatten, waarbij elke
 228 vraag tijdens het vragenuur als één debat gezien wordt. Op basis van de aan-
 229 tallen is er voor classificatie een baseline nauwkeurigheid van 0.15 (door altijd

grootste partij te kiezen) en baseline F_1 score van 0.11 (door willekeurig te voorspellen gewogen bij aantal spreekbeurten 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 Python. 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.

Pre-processing Voor pre-processing is gebruik gemaakt van tokenisation en lowercasing. Voor tokenisation is de reguliere expressie $w+$ gebruikt, die daarmee alleen de letters en cijfers 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 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 sklearn, 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*. Er is hiervoor gevarieerd met de regularisatie, learning rate en maximum aantal iteraties. Voor regularisatie is hier geëxperimenteerd

273 met Lasso en Ridge regularisatie, en een combinatie van beide genaamd Elastic-
 274 net. De andere parameters zijn gelaten op de standaardwaarden van scikit-learn
 275 [8].

276 **Naive Bayes** Een simpelere techniek die gebruikt wordt voor politieke tekst-
 277 classificatie is Naive Bayes. Dit algoritme neemt aan dat elke *feature* onafhan-
 278 kelijk is ten op zichte van de rest. Dit is bij tekstclassificatie vaak niet het geval
 279 omdat het gebruik van sommige woorden gepaard kan gaan met het gebruik
 280 van andere woorden. Daarnaast is het gebruik van meerdere n-grams in een
 281 classificatie schending van de aanname, want als bijvoorbeeld een bigram er in
 282 voorkomt dan komen ook beide unigrams er sowieso in voor. Desalniettemin
 283 blijkt Naive Bayes effectief te zijn voor tekstclassificatie[8, 5]. Hiervoor zijn de
 284 functies van scikit-learn MultinomialNB en BernoulliNB gebruikt.[8, 5]

285 **Beoordelen van kwaliteit** De meest gebruikte methoden om kwaliteit van
 286 politieke tekstclassificatie te beoordelen zijn nauwkeurigheid en F_1 score, die
 287 opgebouwd is uit recall en precision. Deze scores zijn opgebouwd uit vier varia-
 288 belen te zien in tabel 2. Deze variabelen geven weer hoeveel documenten wel of
 289 niet bij een klasse horen, en of deze wel of niet als dusdanig zijn geclassificeerd.

Tabel 2: BLABLA

| | Behorend tot klasse | Niet behorend tot klasse |
|---------------------------------|---------------------|--------------------------|
| Geclassificeerd als klasse | tp | fp |
| Niet geclassificeerd als klasse | fn | tn |

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

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

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

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

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

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

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 hebben.

Als laatste is er dan nog *gewogen*, dit 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 nauwkeurigheid en het nadeel van *macro* te groot is omdat de klassen nogal variëren in grootte, is gekozen voor *gewogen F_1* scoring naast nauwkeurigheid.

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] bevestigt 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.

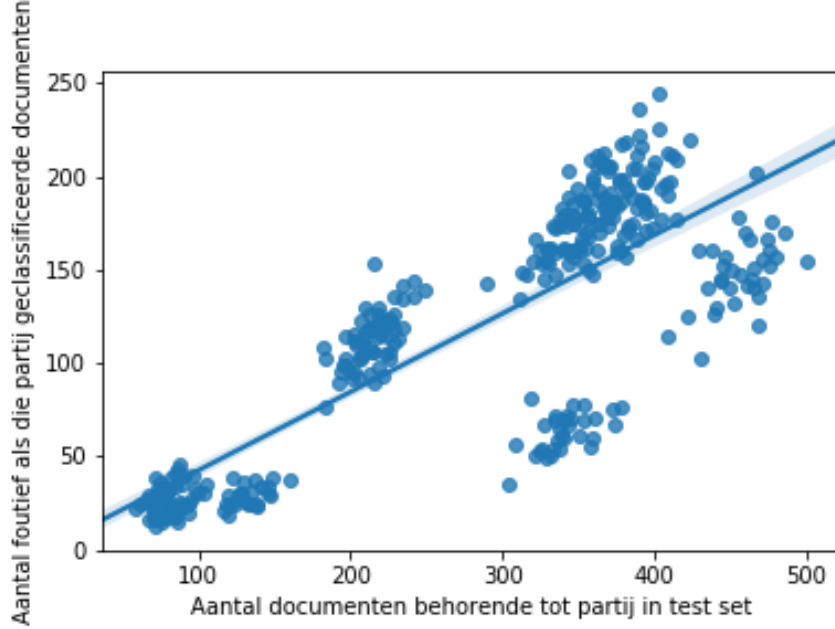
De nauwkeurigheid en F_1 score worden vervolgens vergeleken met de resultaten uit deelvraag 1. Ook wordt gekeken naar verschillen tussen de meest veelzeggende woorden uit deelvraag 1 en uit deze deelvraag.

3.2.3 Deelvraag 3

Om deze deelvraag te beantwoorden zal een analyse gedaan worden van de confusion matrix en 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 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 foutief als partij
geclassificeerde documenten (fp) 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. De pearson correlatie is 0.78.

Aannemend dat dit verband causaal is, is het verwachte aantal documenten

$$V_{i,j} = \left(\sum_{k=0}^n (D_{k,j}) - D_{j,j} \right) * \frac{\sum_{k=0}^n (D_{i,k})}{\sum_{k=0}^n (\sum_{l=0}^n D_{k,l}) - \sum_{k=0}^n (D_{k,j})} \quad (5)$$

waar $i \neq j$ met i de voorspelde partij en j de echte partij waar een document
bijhoort, $D_{i,j}$ het aantal documenten als dusdanig geclassificeerd en n het aantal
partijen. De linkerterm is het totaal aantal documenten behorend tot partij
 i die fout geclassificeerd zijn. De rechterterm is het percentage van het totaal
aantal documenten minus die van partij i dat tot partij j behoort.

De error is dan het verschil van de verwachte waarde en het daadwerkelijk
aantal documenten

$$e_{i,j} = D_{i,j} - V_{i,j} \quad (6)$$

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 [9]. 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

368 de classificatie 50 keer gedaan en worden deze errors bij elkaar in distributie
369 genomen. De nulhypothese is dat er geen verschil is tussen de verdelingen. De
370 alternatieve hypothese is dan dus dat er wel een verschil is tussen de verdelingen.
371 Als de nulhypothese wordt verworpen, kan dus aangenomen worden dat er een
372 verschil is op basis van partij-status.

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

379 In het tweede experiment uit Graeme Hirst et al. worden classifiers ge-
380 traint op de ene zittingsperiode en getest op de andere zittingsperiode. Als de
381 classificatie op basis van ideologie is in plaats van partij-status, is de verwach-
382 ting dat er nog steeds aanzienlijke voorspellingen gedaan worden, aangezien de
383 ideologie naar verwachting redelijk stabiel is binnen tien jaar (hoewel woordge-
384 bruik varieert). Als de scores aanzienlijk lager zijn, kan dit het gevolg zijn van
385 het veranderen van partij-status van partijen.

386 Als vergelijkingsmateriaal is voor deze experimenten een tweede dataset
387 nodig uit een ander kabinet. Hiervoor is het wenselijk dat dit kabinet bestaat
388 uit andere partijen dan kabinet-Rutte II. Daarnaast is het ook wenselijk als het
389 niet te ver terug is, zodat onderwerpen en taalgebruik enigszins overeenkomstig
390 zijn. Omdat kabinet-Rutte I een minderheidskabinet was met een bijzondere
391 partij-status voor de PVV, is ervoor gekozen om de Handelingen van de Tweede
392 Kamer tijdens het missionaire kabinet-Balkenende IV (22 februari 2007 tot 20
393 februari 2010) te gebruiken.

394 De partij 50PLUS bestond nog niet gedurende kabinet-Balkenende IV,
395 dus documenten van deze partij zijn weggelaten. Verder heeft dezelfde verwer-
396 king van data plaatsgevonden, zoals beschreven in 3.1. Alleen de minimum- en
397 maximumlengte is overgenomen van de dataset van kabinet-Rutte II.

398 **3.2.4 Deelvraag 4**

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

403 Voor deze vraag moet vastgesteld worden welke partijen links en rechts
404 zijn. Omdat dit lastig te bepalen is en er meerdere indelingen zijn, wordt hier
405 gebruik gemaakt van twee verschillende indelingen. De indeling op basis van het
406 Kieskompas van Andre Krouwel voor de Kamerverkiezing 2012 en de indeling
407 volgens het Manifesto Project[10] gebaseerd op verkiezingsprogramma's voor
408 de Kamerverkiezing van 2012. In beide gevallen is de nullijn van het politieke
409 spectrum gebruikt om te bepalen of een partij links of rechts is.

Tabel 3: 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 Resultaten

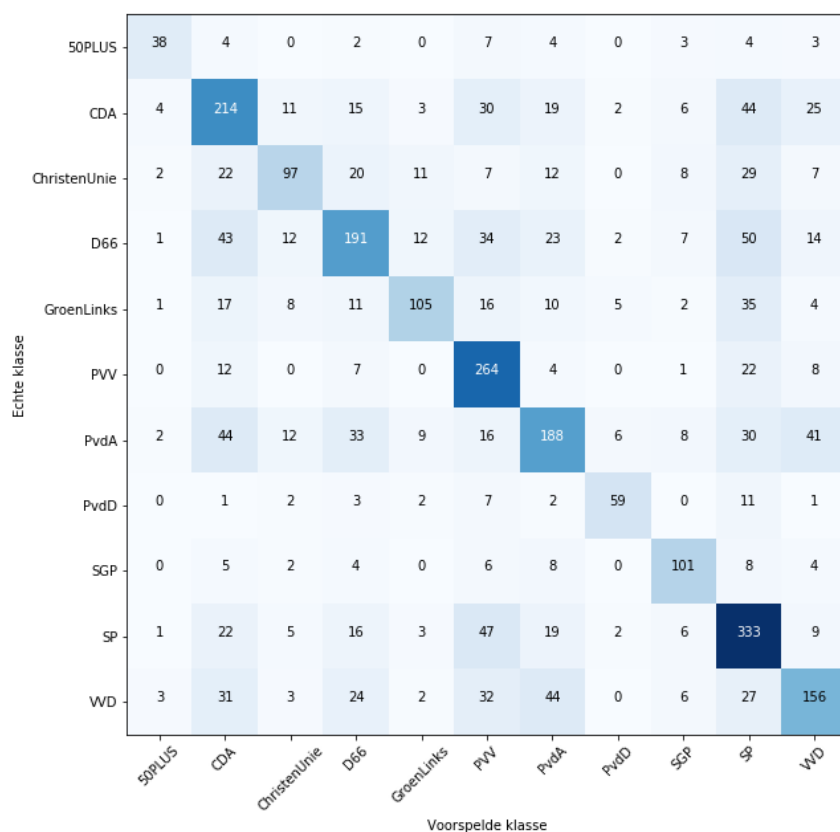
4.1.1 Deelvraag 1

Het beste resultaat werd bereikt met SVM gebruikmakend van *stochastic gradient descent learning* en Ridge regularisatie.

Figuur 4 laat de classificatiescores zien per partij met hoe vaak een partij voorkwam in de test set. Figuur 3 laat zien waar de fouten in deze classificatie zitten. De meest karakteristieke features per partij zijn te zien in figuur 5. Alle resultaten zijn op basis van één classificatie, dus zonder cross validation.

Tabel 4: Classificatierapport van beste classificatie.

| Partij | Precision | Recall | F1_score | Documenten |
|--------------|-----------|--------|----------|------------|
| 50PLUS | 0.73 | 0.58 | 0.65 | 65 |
| CDA | 0.52 | 0.57 | 0.54 | 373 |
| ChristenUnie | 0.64 | 0.45 | 0.53 | 215 |
| D66 | 0.59 | 0.49 | 0.53 | 389 |
| GroenLinks | 0.71 | 0.49 | 0.58 | 214 |
| PVV | 0.57 | 0.83 | 0.67 | 318 |
| PvdA | 0.56 | 0.48 | 0.52 | 389 |
| PvdD | 0.78 | 0.67 | 0.72 | 88 |
| SGP | 0.68 | 0.73 | 0.71 | 138 |
| SP | 0.56 | 0.72 | 0.63 | 463 |
| VVD | 0.57 | 0.48 | 0.52 | 328 |



Figuur 3: Confusion matrix van beste classificatie.

Tabel 5: Meest relevante woorden per partij op basis van beste classificatie gedurende kabinet-Rutte II.

| 50PLUS | CDA | ChristenUnie | D66 | GroenLinks |
|----------------|------------------|---------------------|---------------------|--------------------|
| 50plus | cda | de christenunie | d66 | groenlinks |
| lid krol naar | het cda | christenunie | leden van veldhoven | lid van tongeren |
| lid krol | cda fractie | de leden voordewind | lid van veldhoven | leden voortman |
| het lid krol | de cda fractie | leden voordewind | mijn fractie | de leden voortman |
| krol naar mij | de cda | lid dik | lid van meenen | lid voortman naar |
| krol naar | lid omtzigt naar | lid dik faber | van veldhoven | het lid voortman |
| van 50plus | het lid omtzigt | het lid dik | veldhoven | lid voortman |
| krol | lid omtzigt | de leden segers | d66 wil | leden van tongeren |
| 50plus is | geurts | leden segers | d66 is | voortman |
| lid klein naar | omtzigt | de leden dik | de leden schouw | tongeren |

Tabel 5: Meest relevante woorden per partij op basis van beste classificatie gedurende kabinet-Rutte II. (*Vervolg*)

| PVV | PvdA | PvdD | SGP | SP | VVD |
|-----------------|-----------------|-------------------|-----------------------|------------------|----------------|
| pvv | de pvda | het lid ouwehand | sgp | sp | de vvd |
| de pvv | pvda | lid ouwehand | de sgp | de sp | vvd |
| islamitische | pvda fractie | lid ouwehand naar | leden dijkgraaf | sp fractie | de vvd is |
| klever | de pvda fractie | ouwehand naar mij | de leden dijkgraaf | de sp fractie | vvd is |
| graaf | de arbeid | ouwehand naar | leden van der | smaling | de vvd fractie |
| natuurlijk | van de arbeid | ouwehand | sgp fractie | lid van gerven | vvd fractie |
| agema naar mij | de partij van | voor de dieren | de sgp fractie | lid smaling | voor de vvd |
| agema naar | arbeid | thieme | mevrouw de voorzitter | het lid smaling | wat de vvd |
| klever naar mij | partij van de | de dieren | mevrouw de | lid smaling naar | de vvd wil |
| klever naar | partij van | dieren | de leden bisschop | lid leijten naar | vvd wil |

4.1.2 Deelvraag 2

In figuur 5 was al te zien dat de meest karakteristieke woorden voornamelijk bestaan uit partijnamen en namen van Kamerleden.

In figuur 6 is vervolgens te zien welke woorden het meest karakteristiek zijn per partij, als partijnamen namen van Kamerleden vervangen zijn door een generieke placeholder.

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

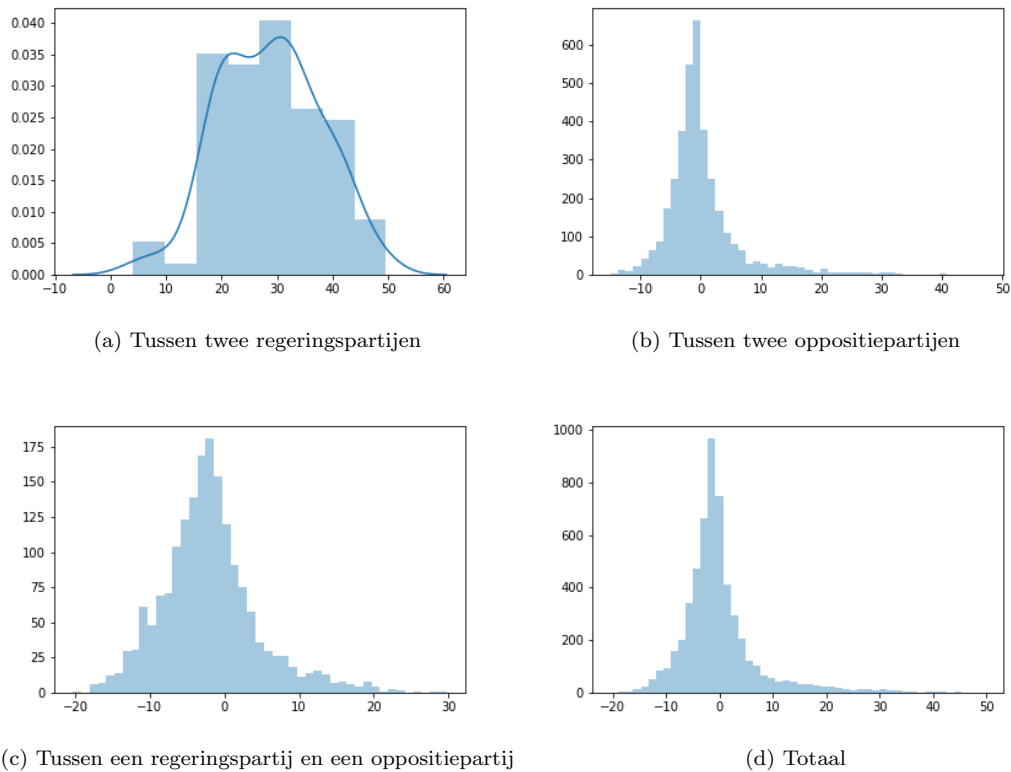
| 50PLUS | CDA | ChristenUnie | D66 | GroenLinks |
|------------------------|--------------------|---------------------------|--------------|---------------------|
| gepensioneerden | PARTIJNAAM fractie | dementie | mijn fractie | zou |
| ouderen | inwoners | voedselverspilling | natuurlijk | schone energie |
| plussers | regering | inderdaad | buitengewoon | persoonsgebonden |
| oudere | PARTIJNAAM | gezinnen | mijn | belastingontwijking |
| koopkrachtontwikkeling | limburg | mensenhandel | het kabinet | kamer hierover te |
| 50 plussers | de regering | zullen | vandaag | in elk geval |
| 50 | echt | horeca | fractie | elk geval |
| werkenden | hier | onder | belangrijk | in elk |
| overwegende | diverse | het ingezetenen criterium | minister | vluchtelingen |
| overwegende dat | zorginstellingen | onder meer | kansen | budget |

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 |
|--------------|---------------------|-------------------|-----------------------|---------------|--------------|
| islamitische | circulaire economie | dieren | mevrouw de voorzitter | huurders | aruba |
| natuurlijk | circulaire | aarde | mevrouw de | zegt | volgens mij |
| al | jongeren | de aarde | beantwoording | voorstellen | PARTIJNAAM |
| nederland | en om | burgers | punt | segregatie | regelgeving |
| miljard | open standaarden | milieu | voor de beantwoording | bezuinigingen | speelveld |
| miljarden | kinderen | natuur | de beantwoording | de bevolking | banen |
| onze | mijn partij | constaterende dat | bewindslieden | mensen | essentieel |
| de islam | redelijk | constaterende | nadrukkelijk | armoede | aangegeven |
| islam | daarbij | de bio | wel | mening dat | ondernemers |
| brussel | de regering tevens | bio | je | bevolking | PARTIJNAAM f |

4.1.3 Deelvraag 3

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



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

4.2 Discussie

4.2.1 Deelvraag 1

Dit onderzoek heeft zich beperkt tot methoden genoemd in vergelijkbare onderzoeken en 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 vergelijkbaar onderzoek voor politieke tekst classificatie zijn daarom niet getest. Omdat dus niet alle opties getest zijn, kan geen uitsluitel gegeven worden dat dit daadwerkelijk het classificatiemodel is. Voor vervolgonderzoek kan daarom gekeken worden naar meer verschillende methoden.

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 het onderzoek van Hirst et al. 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

4.2.3 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] Conal Sathi Maneesh Bhand, Dan Robinson. Text classifiers for political ideologies, 2009.
- [6] Bjørn Høyland, Jean-François Godbout, Emanuele Lapponi, and Erik Vellidal. 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.

- 467 [7] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source
468 scientific tools for Python, 2001–. [Online; accessed `today`].
- 469 [8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Gri-
470 sel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas,
471 A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay.
472 Scikit-learn: Machine learning in Python. *Journal of Machine Learning*
473 *Research*, 12:2825–2830, 2011.
- 474 [9] *NIST/SEMATECH e-Handbook of Statistical Methods*. NIST/SEMA-
475 TECH, April 2012.
- 476 [10] Andrea Volkens, Pola Lehmann, Theres Matthieß, Nicolas Merz, Sven Re-
477 gel, and Bernhard Weßels. The manifesto data collection. manifesto project
478 (mrg/cmp/marpor). version 2017b, 2017.
- 479 [11] Bei Yu, Stefan Kaufmann, and Daniel Diermeier. Classifying party affilia-
480 tion from political speech. *Journal of Information Technology & Politics*,
481 5(1):33–48, 2008.

482 A Slides