Automatic thematic classification of election manifestos

Suzan Verberne¹, Eva D'hondt¹, Antal van den Bosch¹, Maarten Marx²

1. Radboud University Nijmegen 2. University of Amsterdam

Introduction

- Isaac Lipschits, Dutch political scientist, annotated party programmes for the Dutch elections (1977-1998) with themes.
- ► Goals in the Political Mashup project:
- ▶ Digitize the 1977–1998 Lipschits collections
- Build an automatic classifier for more recent, unclassified editions



- 1. Convert the scanned PDFs of 1986, 1994 and 1998 to enriched publications
- 2. Use these data to train a classifier
- 3. Run the classifier on 2006–2012 data
- 4. Ask an expert to evaluate a sample of the labels for the 2006–2012 data

Converting the PDF books to enriched publications



Classification experiments

- ▶ We aim to develop a classifier which assigns themes to unseen Dutch election manifestos written after Lipschits' work
- ▶ We have to rely on the older data from the eighties and nineties for training and optimization of the classifier
- System was tuned by testing on 1998 data, while using older data as training material
- Balanced Winnow, implementation in the Linguistic Classification System (LCS)

Results on 1998 data

Table : Results for the automatic classification	of the 826 tex	ts from the	1998 data
training data	# themes	Precision	Recall
1986 and 1994 data	320	68.8%	37.3%
1986 and 1994 data, only 1994 themes	211	72.2%	37.1%

Adding bigrams, removing stopwords and/or lemmatization did not improve these results

Labelling new, unseen election manifestos

► The body of Lipschits data are the manifesto texts, labelled with numbers at the places in the text where the topic changes

At the back of the book is a register of alphabetically sorted themes with the associated texts per party by topic number.

► Lipschits classification is a multi-label classification task

• 18 Grondrechten Een effectieve verwezenlijking van de fundamentele vrijheden vereist allereerst, dat de overheid alle vrijheidsrechten (klassieke grondrechten) respecteert. Teneinde iemand in staat te stellen de aldus gewaarborgde vrijheden ook daadwerkelijk te genieten, dient het overheidsbeleid gericht te zijn op de realisering van de sociale grondrechten en dient waar nodig aanvullende wetgeving te worden vastgesteld. 12. Bij wetgeving die - in aanvulling op bestaan-

de strafbepalingen en de wetten betreffende gelijke behandeling van mannen en vrouwen in het arbeidsbestel - tot stand wordt gebracht om discriminatie tegen te gaan, zal de wetgever zich door de volgende uitgangspunten moeten laten

a. De wet biedt een ieder bescherming tegen ejegeningen die iemand wegens ras, geloof of levensovertuiging, geslacht dan wel homo-seksuele geaardheid onrechtmatig discrimine-

b. Het oordeel of iemand door discriminerend gedrag onrechtmatig heeft gehandeld, blijft ehouden aan de onafhankelijke rechter De overheid benadrukt in haar beleid het belang van mentaliteitsverandering bijvoorbeeld door zelf prioriteit te geven aan het vervolgingsbeleid terzake discriminatie en door het opnemen van personen uit de desbetreffende groepen in het personeelsbestand van de overheid en het verge makkelijken van het toegroeien naar hogere func-

De overheid mag bij haar benoemings- en ontslagbeleid geen acht slaan op de homoseksuele geaardheid en overeenkomstige levenswijze van 21 Abortus de betrokkene. 13. Voor de vrijheid van godsdienst en levens overtuiging geldt net als voor andere fundamen-tele vrijheden dat de overheid er niet mee mag

Istaan, zich zelf te onthouden van ingrepen i die vrijheid. De overheid heeft ook de plicht ervoor te zorgen dat niet de voorwaarden komen te ontbreken waaronder die vrijheid daadwerkelijk kan worden uitgeoefend: ndien mensen door bepaalde omstandighe

104, 112

porno-bioscop

EVP 40, 59

PPR 23 PSP 20, 21

GPV 46

RPF 22

SGP 7

VVD &

postbank:

CPN 17

D'66 11 GPV 113

PPR 25, 29 PSP 5, 53

PvdA 29,46 VVD 10, 11, 46

postgiro: zie: posterijen

zie: posterije

alleen dragen. • 19 Persoonsregistratie 14. De totstandkoming van een wettelijke rege

ling van de registratie van persoonsgegevens ter bescherming van de persoonlijke levenssfeer dient in de komende parlementaire periode te worden gerealiseerd. Daarbij moeten voorop een regeling van het recht van personen om ervan kennis te nemen waar, waarom en hoe hij of zij geregistreerd staat;

dien men redelijkerwijs niet mag verlanger

dat de betrokkenen de financiële lasten gehee

er dient een regeling van het recht van perso nen te komen om te weten wie of welke instar tie de bevoegdheid heeft de inhoud van de gegevensverzamelingen te raadplegen; een duidelijke regeling van de rechten van de geregistreerde, met inbegrip van een regeling voor de verbetering van onjuist gebleken ge-

effectieve controle op de naleving van deze wetgeving. Persoonsregistraties bij de overheid, die met een verschillend doel zijn opgezet, zullen alleen met inachtneming van daarvoor te stellen wettelijke regels met elkaar mogen worden gecombineerd

• 20 Beschermwaardigheid menselijk lever Eerbied voor en bescherming van het menselijk

even ongeacht het ontwikkelingsstadium of de verschijningsvorm, behoort hoeksteen te zijn van de rechtsorde

15. Op het terrein van de abortus is de wet, die afbreking van zwangerschap slechts toelaat als atie van de vrouw dit onontkoombaa maakt, sinds kort in werking. Op basis van door

de overheid te verlenen vergunningen en het wet telijk geregelde controlestelsel, zal er de komende aren op worden toegezien dat de wet wordt nage eefd. Het adoptiebeleid, dat er mede op gericht i CDA 53, 82 CDA 14, 19, 23, 24, 27, 29, 31, 32, 46, 73 CPN 34 CPN 1-3, 7, 9, 28, 30, 37, 40, 43, 44, 46, 48, 49 D'66 24 EVP 54 D'66 3, 14, 18, 24-26, 33, 35, 36 EVP 20, 40, 59, 60, 64, 65 GPV 9, 20, 25, 55, 111, 114, 120 PSP 38, 39 PvdA 35 PPR 6, 16, 17, 19-21, 23-26, 40, 41 RPF 115 PSP 5-10, 18-23, 42-44, 46 SGP 37 VdA 3, 5, 6, 30, 35, 39, 42 VVD 21 RPF 3, 5, 10, 24-26, 28, 29, 38, 39, 55, 56, 65, 75, privacy, bescherming van SGP 1, 2, 9, 15, 16, 18, 26, 28, 31, 62, 65 zie: persoonsgegevens, registratie van VVD 5, 7, 8, 10, 13, 15, 43, 49, 54

privatisering overheidsbedrijven en overheidsta CDA 2, 15, 36, 41, 45, 75, 77 CPN 9, 17, 27, 39 D'66 9, 11, 15, 40 zie: zedelijkheidswetgeving EVP 62 GPV 21, 22, 49, 135 PSP 5, 6, 46 RPF 3, 29, 40, 41 SGP 13, 64, 76 post-academisch onderwijs. zie: Wetenschappelijk Onderwijs

VVD 2, 3, 9, 10, 17 privatisering overheidstaken zie: privatisering overheidsbedrijven en overheidstaken pro deo-advocatuu zie: wetshandhaving produktiebeperking CDA 50 CPN 25 D'66 13 EVP 9, 15, 21, 22, 24 PPR 12, 17-19 PSP 5, 49, 50 PvdA 23, 44, 59

> prof-voetbal: zie: sportbeleid

Table : Statistics on the classification data # words vocabulary size # texts # themes # words/text # themes/text year

Table : Statistics on the unseen data							
year	# words	vocabulary size	# texts	# words/text			
2006	235,949	35,547	4,771	49.5			
2010	756,254	35,547	21,329	35.5			
2012	246,004	30,477	5,880	41.8			

Evaluation: A sample of the automatically labelled data was manually evaluated by expert (political journalist)

Text to be assessed

Ve are fully committed to employment by making European funds vailable for investment in education, research, innovation, energy fficiency, infrastructure and digitization. These investments are nanced or guaranteed by the relevant EU funding programmmes, the curopean Investment Bank, European project bonds, making better se of Structural Funds and the European Social Fund, and alternative xploitation of agricultural subsidies.	 □ This text cannot be assessed thematically (keep form empty) Assigned themes European cooperation Relevant ○ Not relevant ○ I don't know Comment (optional): employment Relevant ○ Not relevant ○ I don't know Comment (optional): infrastructure Relevant ○ Not relevant ○ I don't know Comment (optional):
	Missing theme (optional) Save and next

Evaluation

- ▶ 193 texts from 2006-2012 data were assessed by the expert
- In addition, she judged 50 text fragments from the 1998 data (manually labelled by Lipschits, but the expert was not aware of that)

Table : Results for the automatic classification of 193 text fragments from 2006 onwards into the 218 themes defined by Lipschits in 1998. The themes were manually evaluated by an expert.

1986	290,942	32,836	797	214	365	10.5
1994	269,270	32,499	951	210	283	6.9
1998	244,697	31,162	826	218	296	8.3

Example of data visualization

Dispersion matrix	<pre>c of t</pre>	the 199	98 m	nanifes	to	of the	e'pen	sioners	party'
#THEME DISPERS	SION VA	LUE DISP	ERSION	MATRIX			•		
ouderenbeleid	0.409			1 11		11	1		
sociale_voorzieningen	0.344	I I				1			
gezondheidszorg	0.276	111							
minimuminkomen	0.276	1				1			
AOW_AOWers	0.272			1					
werkgelegenheid	0.272	- I				1			
loonen_inkomensbeleid	0.258	1	1 11						
ziekenverzorging	0.258	I		1111			1		
economische_groei	0.254	11					1		
${\tt armoedebestrijding}$	0.224	I							
gehandicapten	0.224					I			
belastingen_algemeen	0.219								
Europese_samenwerking	0.219								
individuele_huursubsidie	0.219	I		I					
pensioenen	0.219								
pensioenfondsen	0.219	I							
thuiszorg	0.219	I							
WAO_WAOers	0.219								
wetshandhaving	0.219	1							
Ziekenfonds	0.219	I		1 1					

Year	# texts	# themes/text	Precision	Recall
2006 (trained on 86+94+98)	39	1.7	74.2%	45.8%
2010 (trained on 86+94+98)	122	1.9	77.4%	55.0%
2012 (trained on 86+94+98)	32	2.0	84.6%	56.1%
1998 (manually by Lipschits)	50	7.7	71.5%	89.0%

Conclusions

- 1. In the thematic classification of political texts, a high level of detail seems to be preferred by domain experts
- 2. Change of themes over the years affects recall of the learned classifier,
- 3. but precision is comparable to the precision of annotations from a human expert
- 4. Thus when using old political texts to classify new texts, work is needed to link and expand the set of themes to newer topics.

Centre for Language Studies, Radboud University Nijmegen

s.verberne@let.ru.nl