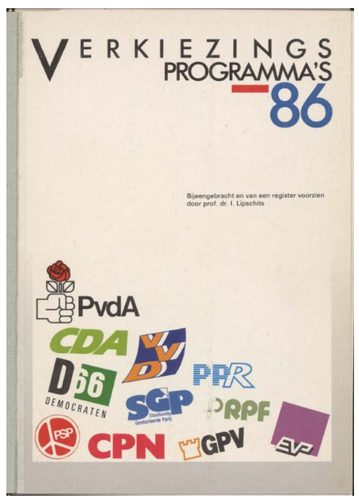
Automatic thematic classification of election manifestos

Suzan Verberne[1], Eva D'hondt[1], Antal van den Bosch[1], Maarten Marx[2]

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

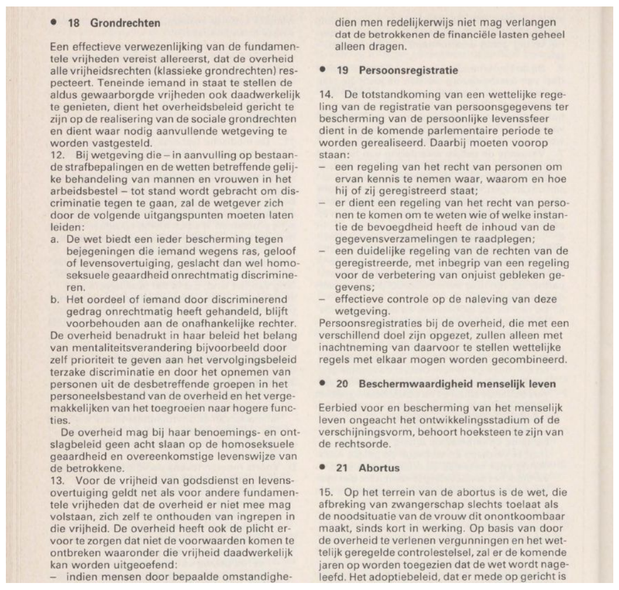


Approach

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

- 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

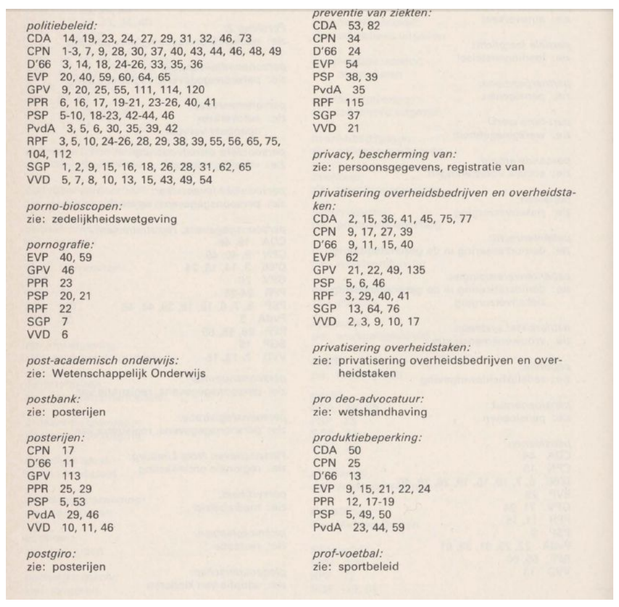
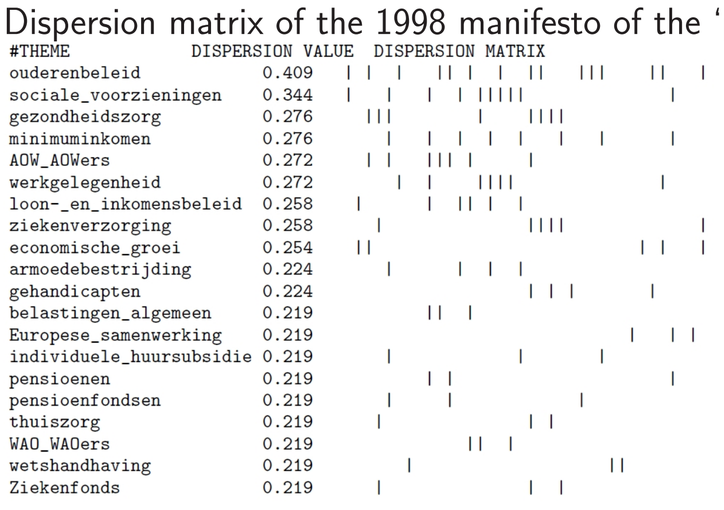


Table : Statistics on the classification data

year	# words	vocabulary size	# texts	# themes	# words/text	# themes/text
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 of the 1998 manifesto of the 'pensioners party'



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 texts 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%

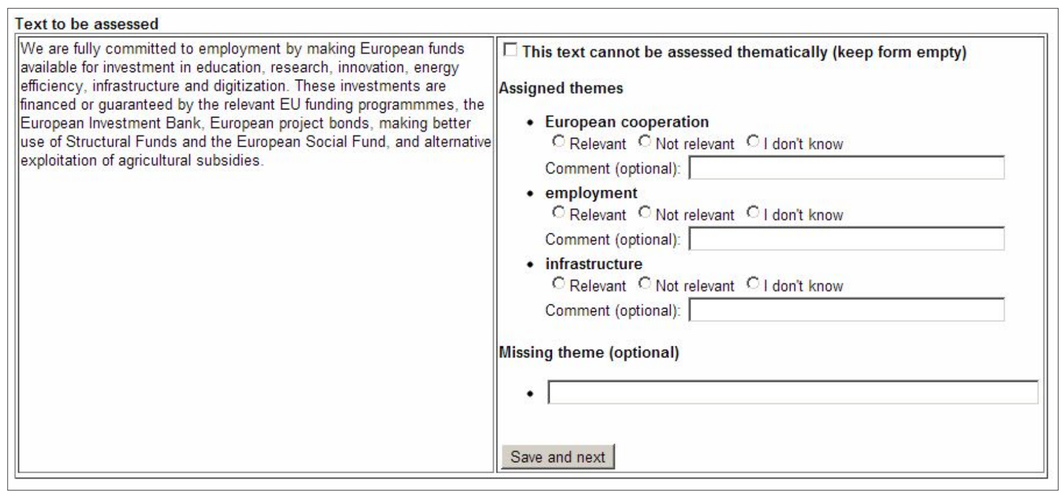
Adding bigrams, removing stopwords and/or lemmatization did not improve these results

Labelling new, unseen election manifestos

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)



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.

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