Information Processing and Management 50 (2014) 554–567

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

# Automatic thematic classification of election manifestos

Suzan Verberne [a,*], Eva D'hondt [b], Antal van den Bosch [b], Maarten Marx [c]

[a] Centre for Language Studies and Institute for Computing and Information Sciences, Radboud University, Nijmegen, Netherlands
[b] Centre for Language Studies, Radboud University, Nijmegen, Netherlands
[c] Informatics Institute, University of Amsterdam, Netherlands

ARTICLE INFO

*Article history:*
Received 27 March 2013
Received in revised form 30 August 2013
Accepted 26 February 2014
Available online 13 April 2014

*Keywords:*
Text classification
Political data
Expert evaluation

ABSTRACT

We digitized three years of Dutch election manifestos annotated by the Dutch political scientist Isaac Lipschits. We used these data to train a classifier that can automatically label new, unseen election manifestos with themes. Having the manifestos in a uniform XML format with all paragraphs annotated with their themes has advantages for both electronic publishing of the data and diachronic comparative data analysis. The data that we created will be disclosed to the public through a search interface. This means that it will be possible to query the data and filter them on themes and parties. We optimized the Lipschits classifier on the task of classifying election manifestos using models trained on earlier years. We built a classifier that is suited for classifying election manifestos from 2002 onwards using the data from the 1980s and 1990s. We evaluated the results by having a domain expert manually assess a sample of the classified data. We found that our automatic classifier obtains the same precision as a human classifier on unseen data. Its recall could be improved by extending the set of themes with newly emerged themes. Thus when using old political texts to classify new texts, work is needed to link and expand the set of themes to newer topics.

© 2014 Elsevier Ltd. All rights reserved.

## 1. Introduction

Isaac Lipschits (1930–2008) was a Dutch historian and political scientist. One of his works is an annotated collection of election manifestos (party programmes) for the Dutch elections between 1977–1998 (Lipschits, 1977). For each election year he compiled a book with the manifestos published by all parties that participated in that year's elections. This book was then made available for purchase nationally before the election. To facilitate voters' decision-making process when comparing parties on election issues, Lipschits manually labelled the manifestos with themes: he segmented the manifestos into coherent text fragments, gave them a unique identifier consisting of the party's acronym and a number, and added an index of themes in the back of the book referring to these identifiers.

In the Political Mashup project (Marx, 2009), Dutch political data from 1814 onwards is being digitized and indexed. The aim of the project is to bring together political information produced by political parties and information on the reception of political promises and actions, in the news as well as in user–generated content. The data are not only digitized and integrated but also disclosed to the public. This means that it will be possible to query the data and filter them on themes, persons and events. In traditional historical and political research, scientists who work with textual data have to open each

* Corresponding author. Tel.: +31 24 36 53431.
*E-mail address:* s.verberne@cs.ru.nl (S. Verberne).

http://dx.doi.org/10.1016/j.ipm.2014.02.006
0306-4573/© 2014 Elsevier Ltd. All rights reserved.

potentially relevant document or book separately, and search and browse through them using a static register. This requires a vast amount of manual analysis and time. The goal of Political Mashup is to automate part of this manual effort.

Election manifestos are traditionally considered to be a key source of information on the ideological stance of a political party on election issues (Budge, Klingemann, Volkens, Bara, & Tanenbaum, 2001). The annotated election manifestos by Isaac Lipschits are part of the Political Mashup data. The aims of the work presented in the current paper are: (1) to digitize the 1977–1998 Lipschits collections and (2) to build an automatic classifier for more recent, unclassified election manifestos. The starting points for our work are the Lipschits books, scanned as PDF files.

The challenges we face in digitizing the data and building the classifier are:

- the PDFs contain complicating OCR errors;
- the text fragments have been assigned multiple themes by Lipschits (more than 6 on average), which resulted in a large number of themes (more than 200) in total, for a relatively small corpus;
- there are inconsistencies in the labelling schemes over time (e.g. 'fishing industry' vs. 'fishing industry policy')[1];
- new themes have emerged over time (e.g. 'information technology') and the same themes may have changing content over time (concept drift);
- automatic text segmentation leads to shorter and much more segments than in the manually segmented training data.

We took the following approach: We first converted the scanned PDFs to XML data in which each text fragment has been annotated with the Lipschits themes (Section 3). We then used these data to train an automatic classifier, using the original segmentation and labelling by Lipschits (Section 4). We optimized the classifier on the task of classifying election manifestos using models trained on earlier years (Section 5). Based on these experiments, we built a classifier that is suited for classifying election manifestos from 2002 onwards using the data from the 1980s and 1990s. We then evaluated the results by having a domain expert manually assess a sample of the classified data (Section 6).

We found that our automatic classifier obtains the same precision as a human classifier on unseen data. Its recall could be improved by extending the set of themes with newly emerged themes. In addition, we found that although a smaller theme set could improve classification scores even more, a more fine-grained classification is preferred by domain experts.

## 2. Background and related work

Automatic text classification (or document categorization) in predefined categories on the basis of pre-categorized examples is a supervised machine learning task (Sebastiani, 2002). If a text in the collection has been assigned to more than one category, the task is called *multi-label* text classification. In multi-label text classification it is sensible for the classifier to rank the classes according to their estimated relevance to the text. Like any supervised machine learning task, text categorization requires a collection of (manually) labelled examples as training material. When training a classification model, each text in the collection is represented as a feature vector. The features are the terms in the collection, where *terms* can be usually read as *words*, although many attempts have been made to extend words with *n*-grams or phrases (Sebastiani, 2002). Considering each word in the corpus as an independent feature makes text categorization a classification problem in a high-dimensional feature space. The most widely used algorithms for text classification are Support Vector Machines (SVMs) and Naïve Bayes (NB), although NB is mostly considered as a baseline for more sophisticated techniques. Van Mun (1999) argues that in domains with large amounts of features it is better to use Balanced Winnow (Dagan, Karov, & Roth, 1997; Littlestone, 1988) because of its ability to discard irrelevant features. In Section 4, we describe the Winnow classifier that we used for training the Lipschits classifier in more detail.

### 2.1. Political text classification

Automatic classification of political texts using supervised learning techniques has been applied to legislative texts, parliamentary documents, manifestos, and even speeches of the Dutch Queen (Breeman et al., 2009; Hillard, Purpura, & Wilkerson, 2008; Louwerse, 2011; Purpura & Hillard, 2006). Developers of classifiers for topical classification of political texts are faced with two challenges: (1) the definition of political topics is often fine-grained, resulting in a large set of topics and (2) the set of topics and the content of these topics is not stable over time (Mourão et al., 2008). The research council of the European Union has trained a multi-label classifier on EU legislation labelled with the official EU EUROVOC thesaurus consisting of more than 7000 classes (Pouliquen, Steinberger, & Ignat, 2003). The classifier proposes a ranked list of classes for each input document. Their JEX system reaches an *R*-precision of 0.56 for English documents. On average, documents are labelled with six classes, so this means that roughly three from the top 6 proposed classes are correct. Both the classifier and the training data have recently been made available (Steinberger, Ebrahim, & Turchi, 2012). Hillard, Purpura, and Wilkerson (2007) evaluate the efficacy and accuracy of automatic text classification using a corpus of all federal public bills introduced in the U.S. since 1947. The bills have been labelled according to a hierarchical classification scheme comprising 20 major

[1] For the convenience of the reader, we translated all example themes from Dutch to English throughout the paper, except for places where the literal orthography is important.

Download English Version:

https://daneshyari.com/en/article/515851

Download Persian Version:

https://daneshyari.com/article/515851

Daneshyari.com