Machine Learning-Based Content Analysis

Björn Burscher

Machine Learning-Based Content Analysis

Automating the analysis of frames and agendas in political communication research

Björn Burscher

University of Amsterdam

Machine Learning-Based Content Analysis: Automating the analysis of frames and agendas in political communication research

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J Maex
ten overstaan van een door het College van Promoties
ingestelde commissie,
in het openbaar te verdedigen in de Aula der Universiteit
op woensdag 5 oktober 2016, te 11:00 uur
door Björn Burscher
geboren te Lübbecke, Duitsland

Promotiecommissie

Promotores:

Prof. Dr. Claes H. de Vreese Universiteit van Amsterdam

Prof. Dr. Rens Vliegenthart Universiteit van Amsterdam

Co-Promotor:

Prof. Dr. Maarten de Rijke Universiteit van Amsterdam

Overige Leden:

Prof. Dr. Franciska de Jong Erasmus Universiteit Rotterdam

Prof. Dr. Hajo Boomgaarden Universitaet Wien

Prof. Dr. Peter Neijens Universiteit van Amsterdam

Dr. Evangelos Kanoulas Universiteit van Amsterdam

Dr. Joost van Spanje Universiteit van Amsterdam

Faculteit: Faculteit der Maatschappij- en Gedragswetenschappen to Martin, Liesel, Martijn and Stip

Contents

1

Introduction

Me: "What should I wear for Halloween?"

Siri: "Just be yourself pumpkin."

In recent years, computers radically improved in their ability to interpret human language. As the above example suggests, these days, our phones understand what we say, and they even try to be funny in their responses. Our daily lives are full of applications that interpret and act on human language. Email clients, for example, read our messages and automatically remove spam from our inboxes. And social network sites like Facebook continually analyze our online behavior in order to predict what kind of content we are interested in. This, for example, leads to personalized news feeds and search results.

Innovations in the area of natural language processing (NLP) made all of these developments possible. Natural language processing is a field at the intersection of computer science, artificial intelligence and computational linguistics concerned with the question how computers can derive meaning from human (natural) language.

As noted above, advances in NLP have significant impact on various aspects of our daily lives and, in particular, the way we interact with technology. But advances in NLP also impact the way we do scientific research. This dissertation investigates how NLP technologies can be applied to automate media content analysis in political communication research. When studying political communication, one of the central questions is how political messages (e.g., news media coverage) affect people's political attitudes (e.g., opinion about European integration) and behavior (e.g, party choice).

To investigate the effects of political communication, scholars must analyze the content of political messages and infer meaning from them [e.g., Iyengar and Simon, 1993]. Van Spanje and De Vreese [2014], for example, analyzed the content of 37,000 news articles to study the effects of media evaluations of the EU on voting for Eurosceptic parties. Such analysis of media content is traditionally done by human coders, who carefully read the messages and then interpret their content. However, due to an enormous increase in digitally available media content and advances in the field of NLP, computer-assisted analysis of political messages becomes increasingly popular [Günther and Quandt, 2015, Grimmer and Stewart, 2013].

Automating the analysis of political messages cannot only save time and resources, but also opens up new possibilities for studying the content and effects of communication. Among others, this includes investigations of the causal direction of media effects, the duration of media effects, and the conditions under which media effects occur.

The topic of this dissertation then is the question how NLP technologies can be applied to facilitate the analysis of media content in political communication research. This dissertation thus does not include media effect studies. It rather consists of methodological contributions addressing the automation of media content analysis, which can be applied in future communication research.

The method for studying political messages in communication research is called *content analysis* (CA). Holsti [1969, p.45] defines content analysis as "the process of making inferences about the antecedents and characteristics of communication". The term antecedents basically refers to the source of the communication. A relevant question is: What does the message tell us about the political ideology of the source [Barberá, 2015]?

The most studied sources of political communication are news media outlets (e.g., The New York Times), members of the public (e.g., a political blogger) and political actors (e.g., the Democratic Party). These sources can use different types of political text to carry a message. A few examples are newspaper articles, online news stories, political speeches, tweets, website posts and parliamentary records.

With regard to the characteristics of such messages, political communication scholars are interested in various phenomena. One example are the policy issues and/or political actors that are discussed in a message, and the ways they are evaluated. Relevant questions are: What is the topic of a message [Burscher et al., 2015]? And does the message portray a particular political party in a negative or positive way [Boomgaarden et al., 2012]? Generally, the content characteristic of interest is rooted in a theoretical framework and/or a particular research question.

In this dissertation, we investigate automated content analysis techniques for studying two major theories from the field of political communication: *Agenda Setting* and *Framing*. Agenda setting is the study of *what* is discussed by the media, the public and policymakers [McCombs and Shaw, 1972, Wanta and Ghanem, 2007]. The key content characteristic in agenda setting research are the topics of political messages. Framing is the study of *how* a topic is discussed in political discourse [Shah et al., 2002]. The key characteristic in framing research is called a news frame - a message's perspective on a topic.

In the following sections, we briefly review agenda setting theory and framing theory, and discuss the role of (automated) content analysis in agenda setting and framing research.

1.1 Agenda Setting

Agenda Setting research is concerned with dynamics in issue salience among the media (media agenda), citizens (public agenda) and policymakers (political agenda) [Rogers et al., 1993]. In the past 40 years, the ability to change the salience of issues (e.g., the economy, the environment, immigration or housing) on these three agendas has been at the center of much research in political communication and political science. Key questions are: What issues get most attention by news media, policymakers and the public? And how do these agendas affect each other over time?

Political communication scholars mainly focus on how news media influence the public agenda. Cohen [1963, p.13] was one of the first to phrase the classical agenda setting hypothesis. He argued that "the press may not be successful much of the time in telling people what to think, but it is stunningly successful in telling its readers what to think about". Five years later, in their seminal Chapel Hill study, McCombs and Shaw [1972] empirically tested this hypothesis for the first time. They found a correlation between what the citizens of Chapel Hill (North Carolina) said were the most important election issues and the issues getting most coverage by the news media.

During the past decades hundreds of studies have investigated and extended the classical agenda setting hypothesis [Wanta and Ghanem, 2007]. Scholars, for example, have studied agenda setting effects between different news outlets (inter-media agenda setting) [Golan, 2006] and reverse agenda setting effects that point from the public to the news media [Roberts and McCombs, 1994, Neuman et al., 2014].

Political scientists, in contrast, focus on what is called *political agenda setting*. Starting in the mid-80s, political agenda setting research has been concerned with the relationship between the media and the political agenda [Walgrave and Van Aelst, 2006]. The key question is to what extent and under what conditions government elites and policymakers set the news agenda and, in turn, news media affect the salience of policy issues on the political agenda.

Soroka [2002], for example, studied agenda setting effects between national news media and the political agenda in Canada. In his study, he classified several policy issues in Canadian newspaper articles and parliamentary questions. Results indicate a reciprocal relationship between the media agenda and the political agenda.

Three developments in particular have been shaping the evolution of agenda setting research. First, the number of policy issues that scholars distinguish between in agenda setting studies has increased during the past decades. McCombs and Shaw [1972] studied five different issues (foreign policy, fiscal policy, law & order, public welfare and civil rights) in their Chapel Hill study. Nowadays, agenda setting studies differentiate between many more topics.

A good example of this development is the Policy Agendas Project

[Baumgartner et al., 2006], in which scholars have developed a system to classify political text (e.g., bills, parliamentary questions or news articles) for policy issues. The taxonomy consists of 20 major topics (e.g., *defense*) and more than 200 subtopics (e.g., *military personnel*), which can be used to compare issue attention longitudinally, across content domains and countries. This is important, because agenda setting research has shown that political agenda setting dynamics, and the media's role in it, depend upon the type of issue [Bartels, 1996, Soroka, 2002].

Second, the number of sources that scholars analyze when studying issue salience among agendas has increased. With regard to the media agenda, for a long time scholars exclusively analyzed the content of traditional news media (newspapers and TV news). More recent studies, in contrast, investigated the media agenda by analyzing content from a multitude of sources, including online news sites [Althaus and Tewksbury, 2002] and social media sites [Meraz, 2009].

Concerning the political agenda, a wide range of parliamentary records (e.g., parliamentary questions and congressional bills) have become available digitally and have been subject of content analysis in agenda setting studies. Furthermore, agenda setting has been studied cross-national [Baumgartner et al., 2006], which involves the analysis of public opinion, media content and parliamentary records in different countries.

Third, instead of looking at the correlation in issue attention between two agendas at one point in time (e.g., media agenda and public agenda during one election), scholars try to measure media agendas, political agendas and public agendas over time - trying to understand the causal direction and dynamics of agenda setting effects.

This has led to a shift from cross-sectional research designs to longitudinal studies involving several waves of public opinion surveys [Matthes, 2008] and/or long time analysis of media content and political records [Vliegenthart and Walgrave, 2008]. Relevant questions are: Do news media set the public agenda or is it the demand of the public that dictates the media agenda? And do policymakers set the media agenda or do the media assert influence on the issues that policymakers discuss in parliament?

When studying agendas, political messages are an important data source. The described shifts toward more issues, more sources and longer periods of time, therefore, present a methodological challenge for agenda setting researchers. Especially the latter two developments lead to an enormous increase in the amount of text data that needs to be analyzed. Traditionally, this is done by means of manual content analysis, where human coders classify the dominant policy issue(s) of each document.

However, when analyzing different agendas over a long period of time, this becomes an expensive undertaking. Various scholars, therefore, advocate for the use of computer-assisted content analysis, and also started developing methodological frameworks that allow for the automatic classification of policy issues in political messages [Hillard et al., 2008, Grimmer and Stewart, 2013].

Nowadays, online publics are in a process of constant transformation, and the dynamics between public, political and media agendas are gaining complexity. The increasing accessibility of communications among politicians, online publics and the media facilitates sophisticated analyses of overtime dynamics in agenda setting, which can help addressing issues like the direction, duration and conditionality of agenda setting effects.

We know from previous research [Hillard et al., 2008] that the coding of policy issues can be automated, but it is still unclear which method is most suitable in which situation, and what exactly the limits of automated content analysis are. We address these questions in this dissertation.

In Chapter 4 and Chapter 5, we discuss in detail the role of automatic content analysis in agenda setting research. In both chapters we introduce and empirically test methods for the automatic classification of policy issues in political messages. We show that policy issues can be analyzed automatically and explain how automatic content analysis can facilitate the study of agenda setting.

1.2 Framing

Agenda-setting research focuses on *what* issues are discussed in political communication. Framing research, in contrast, addresses the question of *how* an issue is talked about. The basic idea behind framing is that news media can shape public opinion regarding an issue by emphasizing some elements of the story over others [Jasperson et al., 1998]. Such emphasis in salience of an element of a story is called an *emphasis frame* [e.g., De Vreese, 2005].

Communication scholars distinguish between different sorts of emphasis frames in news. One class of emphasis frames is called *issue-specific frames*. Such frames relate only to specific topics or events. The nuclear power debate is a prime example of an issue, which has been studied in framing research [Gamson and Modigliani, 1989]. Several frames have been identified, all of which focus on different elements of the nuclear power debate. Two popular ways of framing nuclear power are to focus on either (a) the health and environmental risks of radioactive waste [Joppke, 1991, Bickerstaff et al., 2008] or (2) presenting nuclear power as a means to satisfy energy demands and provide energy independence [Nisbet, 2009, Pidgeon et al., 2008].

Framing an issue in a specific way can affect how people think about the issue. Various studies have empirically shown that exposure to a message, which emphasizes a particular subset of relevant considerations in a debate, causes a person to focus on these considerations when forming an opinion [Chong and Druckman, 2007].

Sniderman et al. [1993], for example, found that a majority of the public supports the rights of a person with HIV when the role of civil liberties is stressed in the news and supports mandatory testing when the importance of public health is stressed.

Another class of frames is called *generic frames* [De Vreese and Semetko, 2002]. As compared to issue-specific frames, generic frames are used in coverage of different topics. A well-studied example of generic frames are episodic and thematic frames [Iyengar, 1991]. When news is framed episodically, it focuses on individual cases and discrete events. In contrast, thematically framed news focuses on general trends and highlights the bigger picture. A news article about immigration, for instance, can highlight one immigrant's personal destiny (episodic) or the general situation of immigrants in a country (thematic).

Another set of generic news frames have been introduced by Semetko and Valkenburg [2000]. In their study of media coverage on European politics, they suggested five generic frames: conflict, human interest, economic consequences, morality, and responsibility. Conflict framing, for instance, highlights conflict between individuals, groups or institutions. Prior research has shown that the depiction of conflict is common in political news coverage [Neuman et al., 1992], independent of topics discussed. As with issue-frames, exposure to generic frames can affect people's opinions and political behavior [Gross, 2008].

When studying news framing and its effects on the public, content analysis is an important research method [Matthes, 2009]. Generally, communication scholars distinguish between two content-analytical tasks in frame analysis: *frame identification* and *frame coding*.

Frame identification is the task of retrieving and defining frames adopted in political communication. This is necessary in order to explore the different ways a specific issue is framed by the news media as well as by political actors.

Traditionally, frame identification is a manual process where scholars qualitatively analyze a sample of documents about an issue and, based on this analysis, define a set of relevant issue frames [Simon and Xenos, 2000]. Although such qualitative analysis is common practice in communication science, several scholars have pointed out potential drawbacks of the approach [Carragee and Roefs, 2004, Hertog and McLeod, 2001]. The main point of criticism is that the set of identified frames can be biased by the subjective perceptions and interpretations of the researcher.

To deal with this issue, scholars have turned to computer-assisted methods, which can identify frames automatically by looking for semantic patterns in a collection of documents [Motta and Baden, 2013, Matthes, 2008]. Such methods are not only more cost-effective than manual frame identification, but also reduce the risks of human bias. Furthermore, methods for automatic frame identification scale better to big datasets, which become increasingly available as the use of social media and the availability of digital news content increases.

Various studies have investigated frameworks for automatic frame identification [Matthes, 2009]. In Chapter 3 of this dissertation, we review this research and present a method to automatically identify frames in news coverage. We empirically demonstrate that this method can be used to automatically identify issue frames.

Frame coding is the process of annotating earlier defined frames as content analytical variables. This is relevant when one already knows the different ways an issue can be framed, and wants to study the usage and effects of such frames. The majority of research investigating framing effects is experimental. However, several scholars have advocated studying framing processes outside the laboratory [Kinder, 2007]. In order to estimate framing effects in non-experimental studies, one must measure a person's exposure to various news frames. This involves the coding of defined frames in political messages.

Previous research has shown that framing effects change over time [Lecheler and De Vreese, 2013] and that frames are constantly challenged by competitor frames [Chong and Druckman, 2007]. Furthermore, most people are exposed to political messages from various sources (different newspapers and digital media). Therefore, appropriate research designs require large-scale over-time content analysis among various sources. Automatic content analysis facilitates the use of such research designs in framing studies.

In communication research, various manual and computer-assisted methods have been applied to coding of frames. In Chapter 2, we review these methods and investigate a technique to automatically code frames in news. We show that this technique can be used to code news frames with accuracy levels comparable to human coding.

1.3 Content Analysis

Having introduced agenda setting theory and framing theory - which form the theoretical foundation of this dissertation - we want to take a closer look at the method of content analysis and how it is applied in agenda setting and framing research. Various approaches toward content analysis exist [see Krippendorff, 2012, for an overview] in the social sciences. This dissertation focuses on *quantitative content analysis*. Berelson [1952, p.31] defines quantitative content analysis as "a research technique for the objective, systematic and quantitative description of the manifest content of communication".

In other words, quantitative content analysis is all about counting the presence of clear-defined (textual or visual) content characteristics in texts. Objective and systematic implies that each step in the analysis should follow explicit rules, in order to prevent subjective interpretation. Qualitative content analysis can be conducted manually - by human coders - or automatically - by means of a computer program.

Manual Content Analysis

Traditionally, content analysis is conducted by human coders, who code the content of texts by means of a code book. This approach is also called *holistic content analysis* [Semetko and Valkenburg, 2000, Wanta et al., 2004]. A code book contains questions and answer categories for the content characteristics of interest. Consider the following example:

If one studies agenda setting effects and wants to measure the salience of different policy issues in news articles, one would hire a research assistant, give him/her a collection of news articles and a code book. The code book contains a question like

"What is the main policy issue of the news article?"

and a set of categories to choose from (e.g., immigration, energy, employment, ...). Then the coder reads the articles and chooses one or multiple categories for each article. In the end, one can count the number of articles that contain each of the issues and arrive at a conclusion regarding the salience of each policy issue on the news agenda.

In some cases, the concept of interest is multi-faceted and cannot be captured by a single question. An example of this is the coding of news frames like the attribution of conflict in a news article. In such cases, it is common practice to give coders a set of related questions that address different aspects of the concept [Simon and Xenos, 2000]. The use of multiple questions helps to ensure the objectivity of the analysis.

In order to code the presence of a conflict frame, one might use the following questions [Semetko and Valkenburg, 2000]:

(1) "Does the article reflect disagreement between parties, individuals, groups or countries?"

(2) "Does the item refer to two sides or more than two sides of the problem?"

To arrive at a final coding decision, the answers to all questions can be aggregated. This is called *indicator-based content analysis* [Krippendorff, 2012], because the different questions are all indicators of the same underlying concept. Creating a codebook generally involves an iterative process in which indicator questions are tested and refined repeatedly.

Holistic and indicator-based manual content analysis are considered the gold standard in content analysis. However, these are very expensive and labor intensive procedures, especially when analyzing large datasets and/or coding multiple content characteristics; which is common practice in communication research. The content analysis of the 2004 European Parliamentary election study, for example, contained a code book with more than 50 different questions. Human coders had to answer all of these questions for each of more than 17,00 different news stories [De Vreese et al., 2006].

Another issue in manual content analysis is that human coders are not always perfectly reliable. Although, quantitative content analysis follows a very systematic approach and coders get well-trained, different coders often make different coding decisions. This can lead to poor inter-coder agreement and bias the results of a content analysis [Lombard et al., 2002].

As the amount of digital media content increases and NLP technologies improve, the demand for automatic forms of content analysis grows. Automatic content analysis scales better to large datasets as compared to manual approaches. Furthermore, automatic content analysis provides the opportunity to study such large amounts of data in real-time, which facilitates the investigation of innovative research question in political communication research.

Automatic Content Analysis

Since the 1960's scholars have introduced approaches to automate media content analysis. Traditionally, automated content analysis employs a *dictionary-based approach*. In dictionary-based content analysis, previously defined character strings are used to code messages into content categories [e.g., Schrodt et al., 1994]. This means that one creates a dictionary with search strings - a set of search terms and rules of combining them (e.g., AND, OR, NOT) - for each content category. Then a computer program counts the occurrence of these search terms in the input texts [Young and Soroka, 2012].

Dictionary-based content analysis is a highly reliable way of analyzing political messages. However, creating a good coding dictionary can be a laborious and difficult process; especially when the number of content categories is large [Hillard et al., 2008]. The biggest challenge is that most people do not know the complete set of words that indicate a particular content category and/or all ways such words can be used. Consequently, not all relevant documents can be retrieved, and results of the content analysis might be biased.

More recent research instead focuses on the use of *machine learning* methods [Alpaydin, 2004] to automate content analysis. Many modern NLP methods are based on machine learning. The paradigm of machine learning is different from that of manual and dictionary-based content analysis. As explained above, manual and dictionary-based approaches are based on the direct coding of sets of rules that have been created by human experts (e.g., questions in a code books or search strings in a coding dictionary).

In machine learning, in contrast, learning algorithms are applied to automatically learn such rules through the analysis of large corpora of real-world examples. In the next section, we will elaborate on machine learning methods and explain how they can be applied to content analysis in communication research.

1.4 Machine Learning in Content Analysis

What is machine learning (ML)? Essentially, ML is a class of methods for teaching computers to make predictions based on data [Alpaydin, 2004]. But how can a computer do that? One way to think about this is that machine learning is based on the recognition of patterns in the data. In machine learning, one generally employs an algorithm that recognizes patterns in a dataset, generalizes such patterns into a model, and then makes predictions based on the model.

Machine learning has been applied to various content-analytical tasks like coding the topic of news articles [Hillard et al., 2008], identifying frames in news [Miller, 1997, Matthes, 2008], predicting the tone of social media posts and predicting the political ideology of Twitter users [Barberá, 2015].

There are various methods for machine learning-based content analysis [Grimmer and Stewart, 2013] and it is beyond the scope of this dissertation to provide a complete review. For the sake of simplicity, here we distinguish between two forms of machine learning-based content analysis: *supervised machine learning* and *unsupervised machine learning*. Both methods have been applied in political communication research and are used in the research included in this dissertation.

In supervised learning, a computer learns from a set of already labeled example documents to automatically make a coding decision. Thereby, the set of content categories has to be defined a priori. Consider the following example: One has a labeled dataset of news articles, each of which covers either the economy or foreign policy. Using these labeled articles as training data, one can train a machine learning algorithm to automatically predict the content of a new, unseen article. See Figure

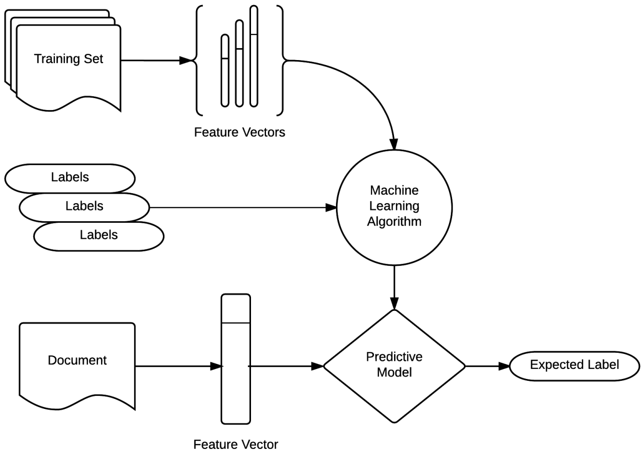

Figure 1.1: Classification - Supervised Machine Learning

1.1.[1] Because the training set consists of economy and foreign policy articles only, the algorithm does not know about other topics. Consequently, it classifies each new article as "economy" or "foreign policy", even if it is about the weather.

Once a classification model is trained it can be used as an efficient coding tool. However, for the training it requires (manually) labeled training data. Supervised learning is useful in situations where the content categories are pre-set, like for coding the main policy issue of a news article according to an a priori defined taxonomy (Chapter 4). Learning from labeled training data is a strength of supervised learning, but also makes it a costly technique in the sense than one first has to create training data by means of manual coding.

Unsupervised learning can be applied in cases where one does not possess labeled training data and/or the set of relevant content categories

[1] This figure is inspired by http://www.astroml.org/sklearn_tutorial

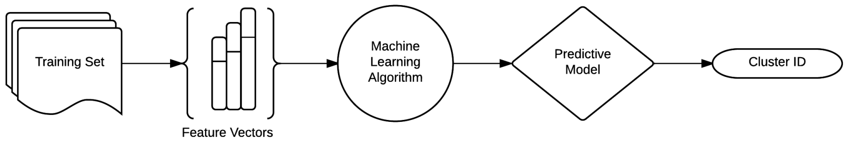

Figure 1.2: Clustering - Unsupervised Machine Learning

is unknown. Imagine one has a set of articles covering the economy, foreign policy, crime and several other topics, but they are not labelled yet. So it is for the machine learning algorithm to form certain clusters based on some notion of similarity between the articles. The machine learning algorithm looks for patterns of similarity between the articles. Based on such patterns, the algorithm creates a set of clusters and assigns each article to one cluster. See Figure 1.2 for an illustration.[2]

As unsupervised learning does not require labeled training data, it is an efficient content analysis tool that comes with very low costs. Unsupervised learning is especially useful when one wants to explore a dataset without having a very clear idea about the exact content categories. The identification of issue-specific news frames is a good example of a problem where clustering can be helpful (Chapter 3). However, as the resulting content categories are completely data-based, it bears the risk of producing categories that are not meaningful theoretically.

During the course of this dissertation different forms of supervised and unsupervised machine learning will be discussed. In Chapter 2 and Chapter 4, we use classification to predict the frames and topics of news articles. In Chapter 3, we apply clustering to identify news frames about the nuclear power debate. And in Chapter 5, we discuss an approach that

[2] This is figure is inspired by http://www.astroml.org/sklearn_tutorial

combines supervised and unsupervised learning.

1.5 Structure of the Dissertation

This dissertation contains four empirical studies. The first two studies address automatic content analysis in framing research and the latter two address automatic content analysis in agenda setting research.

As noted above, there are two main content-analytical tasks in framing research: frame identification and frame coding. In Chapter 2, we address frame coding - the annotation of already defined news frames in political messages. The method we apply is classification, a form of supervised machine learning. We conduct several experiments in which we automate the coding of four generic frames that are operationalized as a set of indicator questions.

In Chapter 3, we study automatic frame identification. Based on a large collection of news articles, we automatically identify issue frames with regard to the nuclear power debate. In doing so, we apply clustering, a form of unsupervised machine learning. We closely investigate the conceptual validity of automatically identified issue frames. Furthermore, we test a way of improving automatic frame identification, so that revealed clusters of articles reflect the framing concept more closely.

Chapter 4 and Chapter 5 deal with the automatic coding of policy issues in political messages. In both studies, we work with the topic taxonomy of the Policy Agendas Project. In Chapter 4, we apply classification, a form of supervised machine learning, in order to teach a computer to code policy issues in news articles and parliamentary questions. Furthermore, we investigate the capability of an automatic coding tool - which is based on supervised machine learning - to generalize across contexts.

In Chapter 5, we apply a dictionary-based approach to code policy issues in news articles and parliamentary questions. Constructing a dictionary with search terms for several content categories can be a difficult and laborious task. Therefore, we introduce a method to automatically expand coding dictionaries with relevant search terms. In doing so, we employ word co-occurrence statistics, which are based on word vectors from a neural network language model. We conduct several experiments in which we use this method to automatically expand dictionaries for coding policy issues. We validate our method by applying automatically constructed dictionaries to different human-coded test sets.

In Chapter 6, we summarize key findings of the dissertation and discuss more broadly the use of automated content analysis in political communication research.

The research presented in this dissertation has as goal to facilitate the study of framing and agenda setting in political communication research. With the increasing availability of digital media content, new challenges and opportunities for agenda setting and framing research have come up. These include the (real time) analysis of large and heterogenous datasets, which enables new possibilities for addressing questions of causality, duration and conditionality of media effects. However, the analysis of such large scale data asks for new research methods that can deal with its scope. The studies in this dissertation investigate the application of such methods in communication research.

1.6 Bibliography

Shanto Iyengar and Adam Simon. News coverage of the gulf crisis and public opinion: A study of agenda-setting, priming, and framing. *Communication Research*, 20(3):365–383, 1993.

Joost Van Spanje and Claes H De Vreese. Europhile media and Eurosceptic voting: Effects of news media coverage on Eurosceptic voting in the 2009 European parliamentary elections. *Political Communication*, 31(2):325–354, 2014.

Elisabeth Günther and Thorsten Quandt. Word counts and topic models: Automated text analysis methods for digital journalism research. *Digital Journalism*, 4(1):75–88, 2015.

Justin Grimmer and Brandon M Stewart. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297, 2013.

Ole R Holsti. *Content analysis for the social sciences and humanities*. Addison-Wesley, Reading, MA, 1969.

Pablo Barberá. Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political Analysis*, 23(1):76–91, 2015.

Bjorn Burscher, Joost Van Spanje, and Claes H De Vreese. Owning the issues of crime and immigration: The relation between immigration and crime news and anti-immigrant voting in 11 countries. *Electoral Studies*, 38:59–69, 2015.

Hajo G Boomgaarden, Rens Vliegenthart, and Claes H De Vreese. A worldwide presidential election: The impact of the media on candidate and campaign evaluations. *International Journal of Public Opinion Research*, 24(1):42–61, 2012.

Maxwell E McCombs and Donald L Shaw. The agenda-setting function of mass media. *Public Opinion Quarterly*, 36:176–187, 1972.

Wayne Wanta and Salma Ghanem. Effects of agenda-setting. In R Preiss, BM Gayle, N Burrell, M Allen, and J Bryant, editors, *Mass media effects research: Advances through meta-analysis*, pages 37–51. Lawrence Erlbaum, Mahwah, NJ, 2007.

Dhavan V Shah, Mark D Watts, David Domke, and David P Fan. News framing and cueing of issue regimes: Explaining Clinton's public approval in spite of scandal. *Public Opinion Quarterly*, 66(3):339–370, 2002.

Everett M Rogers, James W Dearing, and Dorine Bregman. The anatomy of agenda-setting research. *Journal of Communication*, 43(2):68–84, 1993.

Bernard C Cohen. *The press and foreign policy*. Princeton University Press, Princeton, NJ, 1963.

Guy Golan. Inter-media agenda setting and global news coverage: Assessing the influence of the New York Times on three network television evening news programs. *Journalism Studies*, 7(2):323–333, 2006.

Marilyn Roberts and Maxwell McCombs. Agenda setting and political advertising: Origins of the news agenda. *Political Communication*, 11 (3):249–262, 1994.

W Russel Neuman, Lauren Guggenheim, Mo S Jang, and Soo Young Bae. The dynamics of public attention: Agenda-setting theory meets big data. *Journal of Communication*, 64(2):193–214, 2014.

Stefaan Walgrave and Peter Van Aelst. The contingency of the mass media's political agenda setting power: Toward a preliminary theory. *Journal of Communication*, 56(1):88–109, 2006.

Stuart N Soroka. Issue attributes and agenda-setting by media, the public, and policymakers in Canada. *International Journal of Public Opinion Research*, 14(3):264–285, 2002.

Frank R Baumgartner, Christoffer Green-Pedersen, and Bryan D Jones. Comparative studies of policy agendas. *Journal of European Public Policy*, 13(7):959–974, 2006.

Larry M Bartels. Politicians and the press: Who leads, who follows? In *Proceedings of the Annual Meeting of the American Political Science Association*, San Francisco, CA, 1996.

Scott L Althaus and David Tewksbury. Agenda setting and the "new" news patterns of issue importance among readers of the paper and online versions of the New York Times. *Communication Research*, 29 (2):180–207, 2002.

Sharon Meraz. Is there an elite hold? traditional media to social media agenda setting influence in blog networks. *Journal of Computer-Mediated Communication*, 14(3):682–707, 2009.

Jörg Matthes. Need for orientation as a predictor of agenda-setting effects: Causal evidence from a two-wave panel study. *International Journal of Public Opinion Research*, 20(4):440–453, 2008.

Rens Vliegenthart and Stefaan Walgrave. The contingency of intermedia agenda setting: A longitudinal study in Belgium. *Journalism & Mass Communication Quarterly*, 85(4):860–877, 2008.

Dustin Hillard, Stephen Purpura, and John Wilkerson. Computer-assisted topic classification for mixed-methods social science research. *Journal of Information Technology & Politics*, 4(4):31–46, 2008.

Amy E Jasperson, Dhavan V Shah, Mark Watts, Ronald J Faber, and David P Fan. Framing and the public agenda: Media effects on the importance of the federal budget deficit. *Political Communication*, 15 (2):205–224, 1998.

Claes H De Vreese. News framing: Theory and typology. *Information Design Journal and Document Design*, 13(1):51–62, 2005.

William A Gamson and Andre Modigliani. Media discourse and public opinion on nuclear power: A constructionist approach. *American Journal of Sociology*, 3:1–37, 1989.

Christian Joppke. Social movements during cycles of issue attention: The decline of the anti-nuclear energy movements in West Germany and the USA. *British Journal of Sociology*, 42(1):43–60, 1991.

Karen Bickerstaff, Irene Lorenzoni, Nick F Pidgeon, Wouter Poortinga, and Peter Simmons. Reframing nuclear power in the UK energy debate: nuclear power, climate change mitigation and radioactive waste. *Public Understanding of Science*, 17(2):145–169, 2008.

Matthew C Nisbet. Communicating climate change: Why frames matter for public engagement. *Environment: Science and Policy for Sustainable Development*, 51(2):12–23, 2009.

Nick F Pidgeon, Irene Lorenzoni, and Wouter Poortinga. Climate change or nuclear power—no thanks! a quantitative study of public perceptions and risk framing in Britain. *Global Environmental Change*, 18 (1):69–85, 2008.

Dennis Chong and James N Druckman. Framing theory. *Annual Review of Political Science*, 10:103–126, 2007.

Paul M Sniderman, Richard A Brody, and Phillip E Tetlock. *Reasoning and choice: Explorations in political psychology*. Cambridge University Press, Cambridge, UK, 1993.

Claes H De Vreese and Holli A Semetko. Cynical and engaged strategic campaign coverage, public opinion, and mobilization in a referendum. *Communication Research*, 29(6):615–641, 2002.

Shanto Iyengar. *Is anyone responsible?: How television frames political issues*. University of Chicago Press, Chicago, IL, 1991.

Holli A Semetko and Patti M Valkenburg. Framing European politics: A content analysis of press and television news. *Journal of Communication*, 50(2):93–109, 2000.

W Russell Neuman, Marion R Just, and Ann N Crigler. *Common knowledge: News and the construction of political meaning*. University of Chicago Press, Chicago, IL, 1992.

Kimberly Gross. Framing persuasive appeals: Episodic and thematic framing, emotional response, and policy opinion. *Political Psychology*, 29(2):169–192, 2008.

Jörg Matthes. What's in a frame? a content analysis of media framing studies in the world's leading communication journals, 1990-2005. *Journalism & Mass Communication Quarterly*, 86(2):349–367, 2009.

Adam Simon and Michael Xenos. Media framing and effective public deliberation. *Political Communication*, 17(4):363–376, 2000.

Kevin M Carragee and Wim Roefs. The neglect of power in recent framing research. *Journal of Communication*, 54(2):214–233, 2004.

James K Hertog and Douglas M McLeod. A multiperspectival approach to framing analysis: A field guide. In SD Reese, OH Gandy, and AE Grant, editors, *Framing Public Life: Perspectives on media and our understanding of the social world*, pages 139–161. Lawrence Erlbaum, Mahwah, NJ, 2001.

Giovanni Motta and Christian Baden. Evfolutionary factor analysis of the dynamics of frames: Introducing a method for analyzing high-dimensional semantic data with time-changing structure. *Communication Methods and Measures*, 7(1):48–82, 2013.

Donald R Kinder. Curmudgeonly advice. *Journal of Communication*, 57(1):155–162, 2007.

Sophie Lecheler and Claes H De Vreese. What a difference a day makes? the effects of repetitive and competitive news framing over time. *Communication Research*, 40(2):147–175, 2013.

Klaus Krippendorff. *Content analysis: An introduction to its methodology*. Sage, Beverly Hills, CA, 2012.

Bernard Berelson. *Content analysis in communication research.* Free Press, Glencoe, Il, 1952.

Wayne Wanta, Guy Golan, and Cheolhan Lee. Agenda setting and international news: Media influence on public perceptions of foreign nations. *Journalism & Mass Communication Quarterly*, 81(2):364–377, 2004.

Claes H De Vreese, Susan A Banducci, Holli A Semetko, and Hajo G Boomgaarden. The news coverage of the 2004 European Parliamentary election campaign in 25 countries. *European Union Politics*, 7(4): 477–504, 2006.

Matthew Lombard, Jennifer Snyder-Duch, and Cheryl C Bracken. Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, 28(4):587–604, 2002.

Philip A Schrodt, Shannon G Davis, and Judith L Weddle. Political science: KEDS-a program for the machine coding of event data. *Social Science Computer Review*, 12(4):561–587, 1994.

Lori Young and Stuart Soroka. Affective news: The automated coding of sentiment in political texts. *Political Communication*, 29(2):205–231, 2012.

Ethem Alpaydin. *Introduction to Machine Learning*. MIT Press, Cambridge, MA, 2004.

M.M. Miller. Frame mapping and analysis of news coverage of contentious issues. *Social Science Computer Review*, 15(4):367–378, 1997.

2

Frame Coding through Supervised Learning

Chapter 2: Frame Coding through Supervised Learning

This chapter has been published as:

Burscher, B., Odijk, D., Vliegenthart, R., De Rijke, M., & De Vreese, C. H. (2014). Teaching the Computer to Code Frames in News: Comparing Two Supervised Machine Learning Approaches to Frame Analysis. *Communication Methods and Measures, 8*(3), 190-206.

The version presented here has been adapted to follow the overall standards and terminology included in the other chapters of the dissertation.

2.1 Abstract

We explore the application of supervised machine learning (SML) to frame coding. By automating the coding of frames in news, SML facilitates the incorporation of large-scale content analysis into framing research, even if financial resources are scarce. This furthers a more integrated investigation of framing processes conceptually as well as methodologically. We conduct several experiments in which we automate the coding of four generic frames that are operationalised as a set of indicator questions. In doing so, we compare two approaches to modelling the coherence between indicator questions and frames as an SML task. The results of our experiments show that SML is well suited to automate frame coding but that coding performance is dependent on the way the problem is modelled.

2.2 Introduction

In most framing studies, news frames are coded with indicator questions in manual Content Analysis (CA) [Matthes, 2009]. Generally, measures of several indicators are combined to cover different aspects of a frame [e.g., Simon and Xenos, 2000]. Human coders can be properly trained to code frame indicators, and through training their performance can be improved until accuracy and reliability reach satisfactory levels.

However, human coding is a time-consuming and costly process. This limits the scope of CA in framing research. Computers, in contrast, are more naturally suited for the processing of large quantities of documents and the repetitiveness of coding frames. Therefore, we introduce a computer-aided method for indicator-based frame coding; this not only decreases the effort required for CA of news frames but also helps in addressing substantial issues in communication research.

The method we apply is based on Supervised Machine Learning (SML) [Sebastiani, 2002], a technique in which a computer learns from a set of human-coded training documents to automatically predict content-analytical variables in texts. By applying SML to the coding of four generic frames, we develop a theory of how the technique should be used to automate CA in future framing studies.

We address the following issues: first, we investigate how useful it is to model indicator questions when predicting frames using SML. We compare two approaches. In the first approach, we build a classifier to automatically code frame indicators, which we then aggregate to a frame measure (indicator-based approach). In the second approach, we build a classifier to directly code the presence of a frame (holistic approach). Second, we test the generalizability of SML classifiers by applying them to news sources that were not in the training documents. Third, we investigate the relationship between the number of training documents used and the accuracy of computer-based codings.

We conclude that SML is well suited for frame coding and that it addresses several shortcomings of current approaches to automatic CA. Furthermore, we believe that future framing research can profit from SML theoretically as well as methodologically. SML can promote the incorporation of large-scale CA in framing research by making frame coding much faster and less expensive. This facilitates more integrated studies of framing processes [Matthes, 2012] as well as the analysis of large datasets that have become increasingly available. We discuss extensively the theoretical and methodological implications of our findings for framing research and CA in general.

2.3 Automatic Frame Coding

According to Gamson and Modigliani [1989], news coverage can be approached as an accumulation of "interpretative packages" in which journalists depict an issue in terms of a "central organising idea," to which Gamson and Modigliani refer as a frame. Frames in news take a central position in framing models [e.g., Scheufele, 1999]; they are the dependent variable when studying how frames emerge (frame building) and the independent variable when studying effects of frames on predispositions of the public (frame setting). When detecting frames in news media, CA is the most dominant research technique.

In communication research, various methods are applied to the CA of frames in news [see Matthes, 2009, for an overview]. When investigating the framing of news coverage, we distinguish between frame identification and frame coding. While frame identification includes operations aimed at retrieving and defining frames adopted in the news, frame coding is the annotation of frames defined earlier as content analytical variables. Coding a frame requires an operationalization, which enables the methodological assessment of the frame and allows other scholars to reliably study its use across issues, time, and space. Currently, the two most popular frame operationalizations are human coding with indicator questions and dictionary-based computer-aided coding.

Using questions as indicators of news frames in manual CA is the most widely used approach to frame coding. Indicator questions are collected in a codebook and are answered by human coders while reading the text unit to be analyzed [e.g., Simon and Xenos, 2000, De Vreese et al., 2001]. Each question is designed such that it captures the semantics of a given frame. Generally, several questions are combined to cover various aspects of a frame. Human coding of frames with indicator questions is a reliable but resource-intensive process. As the volume of digitally available media content increases significantly, computer-aided methods become desirable and even a necessity.

Most computer-aided techniques for frame coding follow a dictionary-based approach. In such an approach, previously defined character strings and rules for their combination are used to code text units into content categories [Krippendorff, 2012]. In some studies, search strings are used to directly code a frame [Roggeband and Vliegenthart, 2007]. In other studies, search strings are used to code a set of predefined concepts (e.g., an issue), and a frame is then revealed from the co-occurrence of these concepts [Ruigrok and Van Atteveldt, 2007, Shah et al., 2002].

Dictionary-based approaches to frame coding have several disadvantages. First, the researcher herself must manually build the model from which texts are coded into content categories. Therefore, she must design, pre-test, and refine search queries. Not only is this a time-intensive process, but it also may compromise semantic validity. This is because manually compiled classification rules are at risk of being biased by the subjective conceptions and limited domain knowledge of the researcher(s). A search that is too narrow in scope will omit relevant documents (false negatives), while one that is too broad will retrieve unwanted documents (false positives). Supervised Machine Learning (SML) is an alternative approach to computer-aided frame coding that addresses these shortcomings.

When applied to CA, the goal of SML is to automatically code large numbers of text documents into previously defined content categories [see Laver et al., 2003, Durant and Smith, 2007]. Basically, the computer tries to replicate the coding decisions of humans. A precondition for the application of SML is a set of documents that are already coded for the content categories of interest. We call this the training set. SML involves three steps:

First, text documents from the training set are converted so that they are accessible for computational analysis. Each document is represented as a vector of quantifiable text elements (e.g., word counts) that are called features.

Second, feature vectors of all documents in the training set, together with the documents' content labels (e.g., the presence of a frame), are used to train a classifier to automatically code the content categories. In doing so, a supervised machine-learning algorithm statistically analyses features of documents from each content category and generates a predictive model to classify future documents according to the content categories.

Finally, the classifier is used to code text documents outside the training set. For a detailed introduction to SML we refer to Russell and Norvig [2002] or Grimmer and Stewart [2013].

Using SML to automate the coding of frames is an improvement compared to dictionary-based methods [Hillard et al., 2008]. In SML, in contrast to dictionary-based approaches, a computer automatically estimates a model that classifies texts according to content categories. This is not only more efficient but also likely to be more effective because the rules used to detect frames are based on a statistical analysis of human-coded training data. Furthermore, because manually coded material is available, one can systematically assess the accuracy of computer-based annotations.

Additionally, SML is valuable to future framing research more generally. First of all, it makes CA of frames more feasible. Once a classifier is trained to code a frame, it can be effortlessly employed for real-time CA of that frame. This not only leads to savings in time and costs but also promotes integrating (large-scale) CA with experimental as well as survey research.

Furthermore, SML enables scholars to easily increase the scope of framing analysis. Comprehensive CA of mass media allows investigation of news framing and its effects over the long term and also allows more nuanced, conditional and comparative research. This is relevant because more and more media content is becoming available digitally.

Finally, because one can directly study the entire population of texts, an SML approach can decrease the risk of committing sampling errors and prevent problems related to statistical accuracy as a result of limited samples.

2.4 Research Questions

We apply SML to the coding of four generic news frames: conflict frame, economic consequences frame, human-interest, and morality [Semetko and Valkenburg, 2000].[1] In doing so, we study the following questions.

First, we empirically investigate the question of the extent to which an SML approach is suitable for automatic coding of indicator-based news frames.

Second, we investigate how we should model indicator-based frame coding as a SML task. When using machine-learning techniques to tackle methodological challenges in social science research, it is important to tailor its implementation to the specific research problem at hand. We test whether teaching the computer to code a frame directly (holistic approach) is more effective than teaching it to code a set of indicator questions from which the frame is derived by means of aggregation (indicator-based approach). Both approaches are described in detail in the following section.

Third, we investigate the generalizability of SML classifiers. The goal of automating frame coding is to be able to easily code large amounts of data from several sources. Therefore, we are interested in the question of whether our models are able to correctly predict the four frames in articles from news media not included in the training data.

Finally, we study the relationship between the amount of training data used to build a classifier and its performance to predict frames. Because manually coded training data are expensive and labor-intensive to obtain, it is important to know how much training data one needs to build a well-performing frame classifier. We expect that increasing the

[1] All four frames are introduced in detail in the next section.

number of news articles in the training set leads to an increase in coding performance.

2.5 Holistic Versus Indicator-Based Frame Coding

This study aims to increase our understanding of how SML should be used to effectively master CA problems in communication research. SML is a set of algorithms and approaches for automatic classification. Finding the optimal way of performing a specific classification task generally involves comparing various models. Previous studies have compared the performance of different SML algorithms [Joachims, 1998, Pang et al., 2002], feature types [Scharkow, 2013, Alm et al., 2005], feature selection mechanisms [Forman, 2003, Hillard et al., 2008], and validation techniques [Joachims, 2002].

We differentiate between predicting frames directly and predicting them via indicators because we expect this particular modification to impact the performance of indicator-based frame coding. This is relevant because we want to automatically code frames in news as accurately as possible.

Before presenting details of the approaches, we first define some basic concepts. We have a collection of news articles D and a set of frames U, each of which is operationalised as a set of indicator questions V. When applying SML, we predict the probability $P(u_m|d)$ that a frame $u_m \in U$ is present in an article $d \in D$. For this task we build a classifier on the basis of a training set of news articles that humans have coded for each indicator question $v \in V$.

We try to resemble the manual coding process in the indicator-based approach. First, we train a set of classifiers to predict the answer to each indicator question. In formal terms, we estimate $P(\hat{v}_n|d)$ for each indicator question $v_n \in V$ of the frame. As in manual CA, we then combine the predicted answers to the indicator questions into a single frame measure. We thus derive the probability $P(u_m|d)$ for each frame $u_m \in U$ from $P(\hat{u}_m|\hat{v}_1, ..., \hat{v}_N)$ for all indicator questions $v_M \in V$.

Answers to indicator questions can be combined in various ways. In our case, we argue that all questions indicate presence of the frame by focusing on different but equally important aspects of it [Semetko and Valkenburg, 2000]. Therefore, we claim the frame to be present when at least one of the indicators is coded "yes."

In the holistic approach we do not train classifiers to predict indicator questions. Rather, we try to predict the presence of frames directly. First, for each frame we aggregate coded indicator questions in the training data to a single frame measure. Again, a frame is considered present if at least one of the indicators is coded positive. Second, we use the resulting frame-level codings as training data to train a classifier for each frame that can predict the presence of the frame. Formally, we train a classifier to estimate $P(u_m|d)$ for each frame $u_m \in U$. In contrast to the former approach, here we completely ignore indicator-level codings in the SML process, but train our classifiers directly on frame-level codings.

Why exactly do we expect performance differences between the two approaches? This question brings us to the role of indicators in manual frame coding. Indicators are a means of measuring theoretical concepts in texts. In our case, they help coders to decide on the presence or absence of a frame aspect, from which we infer whether the frame is present. An SML algorithm, in contrast, bases its decision on a systematic statistical analysis of the vocabulary of the text. This leads to a complex model in which each unique word is associated with a probability of the text containing the frame. That is, while a human coder relies on a small set of questions as indicators, the computer relies on the presence of each word from the document collection as an indicator.

Therefore, we expect that the holistic approach might provide a better model to predict the frame variable. It is likely that text features include variables, which explain variation in the frame variables very well but do not explain variation in indicators. In the indicator-based approach, the predictive power of such unknown variables is not considered when predicting the frame, because the frame measure is based on the indicators only. In contrast, in the holistic approach, such variables are included in the model to predict the frame.

2.6 Classifiers and Document Representation

To test these SML approaches, we need to train classifiers for predicting indicator questions and frames. In doing so, we must choose a supervised machine-learning algorithm. As we code different frames with several indicators each, the applied SML algorithm must deal with considerable variation in content characteristics. Consequently, one would expect different SML algorithms to perform better, depending on the frame and indicators considered. Therefore, we propose an approach in which we combine the strengths of various SML algorithms [Dietterich, 2000, Hillard et al., 2008, Polikar, 2012]. The resulting combination of different algorithms is called an ensemble of classifiers.

Ensemble classifiers can be constructed in different ways. We applied a technique called stacked generalization, which involves training a learning algorithm to combine the predictions of several other learning algorithms. To do this, we first partitioned the data into a held-in and a held-out set. We then trained each learning algorithm on the held-in set, and obtained a vector of predictions for the held-out set. Each element of the vector corresponded to a prediction of one of the individual algorithms.

Next, we learned how to combine these predictions. We trained a logistic regression model with the individual classifiers' predictions of the held-out set as input, and the correct responses as output.[2] This way of combining predictions of various classifiers into a final predictive model is intended to be flexible in addressing the different complex characteristics of each of the frame-coding tasks. In the ensemble we combined two different Linear Support Vector Machines (SVM) [Joachims, 1998], a Polynomial SVM classifier, and a Perceptron algorithm [Lippmann, 1987].

To train classifiers and apply them to frame coding, the content of each news article must be represented quantitatively as a vector of document features. Such features are variables containing quantified information about an article that is relevant to the coding task. Selecting relevant features has a significant impact on the ability of the SML algorithms to compute a good predictive model and therefore influences coding performance when predicting the presence of frames in future news articles [Sebastiani, 2002].

When selecting document features for our frame coding task, we thus need to confront the question of which elements of a news article constitute a frame. According to Entman [1993, p.52], news frames manifest themselves in certain text attributes as "the presence or absence

[2] Instead of a single split into held-in and held-out, the vectors of predictions are obtained through 10-fold cross-validation.

of certain keywords, stock phrases, (and) stereotyped images (. . .).” Therefore, we assume that it is appropriate to represent each article as a listing of the words it contains. This is referred to as the “bag-of-words” approach and has been shown to be effective in various text classification tasks [Joachims, 1998, Sebastiani, 2002].

Strictly speaking, we represent each article as a vector of TF.IDF weights [Russell and Norvig, 2002]. This means that each word is assigned the number of times it occurs in a document (TF) and is weighted by the inversed frequency of articles in the entire collection containing the word (IDF). The idea behind TF.IDF weighting is to evaluate the power of a word to discriminate between articles. Rare words are assumed to be more discriminating and therefore are assigned higher weight.

Formally, each article $d \in D$ is represented as a vector V containing a TF.IDF weight W for each unique word $t \in T$ in the collection of articles, $V_d = (W_{d1}, W_{d2}, ..., W_{dM})$. The TF.IDF weight for each word in an article is computed as follows: $W_{td} = tf_{td} * idf_t = tf_{td} * log\frac{N}{n_t}$, where N is the total number of articles in the collection, and N_t is the number of articles in the collection that contain word t.[3]

We used the Scikit-Learn machine learning toolkit [Pedregosa et al., 2011] for computing feature representations of documents. For training and testing classification models, we used the Orange Data Mining Toolbox (Demšar et al., 2013). Both libraries are general-purpose

[3] We also tried alternative bag-of-words transformations, for example, binary-word presence, word counts, and parsimonious language models [Hiemstra et al., 2004]. Additionally, we tried representing all articles in terms of n-grams and latent topics as derived from a LDA-model [Vrehuuvrek and Sojka, 2010]. These variations in feature representation, as well as combinations of them, did not improve on TF.IDF weighting. We suggest applying syntactic (e.g., part of speech tags) or semantic features in future research.

machine-learning modules for the Python programming language.

2.7 Four Generic News Frames

We apply SML to the coding of four generic news frames. These are the conflict frame, the economic consequences frame, the human-interest frame and the morality frame.

The conflict frame highlights conflict between individuals, groups or institutions. Prior research has shown that the depiction of conflict is common in political news coverage [Neuman et al., 1992, Semetko and Valkenburg, 2000] and that it has inherent news value [Galtung and Ruge, 1965, Eilders, 1997, McManus, 1994, Staab, 1990]. Furthermore, several scholars have observed an increase in the portrayal of conflict in political reporting [Patterson, 1993, Blumler et al., 1995, Cappella and Jamieson, 1997, Vliegenthart et al., 2011]. Within the field of political communication, the conflict frame is often employed in empirical research [e.g., Vliegenthart et al., 2008].

By emphasising individual examples in the illustration of issues, the human-interest frame adds a human face to news coverage. According to Iyengar [1994], news coverage can be framed in a thematic manner, taking a macro perspective, or in an episodic manner, focusing on the role of the individual affected by an issue. Such use of exemplars in news coverage has been observed by several scholars [Semetko and Valkenburg, 2000, Neuman et al., 1992, Zillmann and Brosius, 2000] and connects to research on personalisation of political news [Iyengar, 1994].

Economic consequence framing approaches an event in terms of its economic impact on individuals, groups, countries or institutions.

Covering an event with respect to its economic consequences has been argued to possess high news value [Graber, 1993, McManus, 1994] and to increase a reader's perception of how relevant the event is [Gamson, 1992].

The morality frame focuses on moral prescriptions or moral tenets when discussing an issue or event. Morality as a news frame has been the subject of several studies and is used in the context of various issues, such as gay rights [Nisbet and Huge, 2006, Nisbet et al., 2003] and biotechnology [Brewer, 2002, 2003].

We have chosen generic news frames because generic frames, as opposed to issue-specific frames, are topic-independent. This enables us to test our SML approaches with semantically distinct frames while using the same dataset. Consequently, our findings are not limited to frames and news coverage concerning one topic.

2.8 Data

Our data consist of front-page news articles of three national Dutch daily newspapers (*De Volkskrant*, *NRC Handelsblad*, and *De Telegraaf*) between 1995 and 2011. All items were collected digitally via the Dutch Lexis-Nexis database. For each year, a stratified sample (13%) of news articles was manually coded for references to politics[4] and the presence of the four frames. Only those articles that were coded positive for references to politics were coded for the presence of the four frames. The unit of coding was the distinct news story.

To measure the extent to which the four frames appeared in stories

[4] Coders were required to answer 'yes' or 'no' to the following question: "Is the story political in nature?"

that mention politics, we used a series of 11 questions to which the coder was required to answer yes or no.[5] See Table 2.1 for the question wordings of all indicators[6] used. Frame codings were constructed by aggregating measures of indicator questions such that a frame was considered present when at least one of its indicators had been coded positive.

Manual coding was conducted by a total of 30 trained coders at the University of Amsterdam. All coders were native speakers of the Dutch language and received extensive training. To assess inter-coder reliability, political news articles from a random subset (N=156) were each[7] coded by two coders. We report Krippendorff's Alpha as well as pairwise agreement (in parentheses) for all frames: conflict frame = .51 (.77), morality frame = .21 (.85), economic consequences frame = .58 (.82), and human-interest frame = .29 (.64). See Table 2.1 for reliability measures for individual frame indicators.[8]

We stress that inter-coder reliability is not optimal. Performance of the classifiers likely suffers from imperfect training data, but we consider it unlikely that this biases the conclusions of our study. In the Discussion we elaborate on how the quality of the training data influences our findings and conclusions.

[5] In previous research, these questions have been shown to be reliable indicators of the four frames [e.g., Semetko and Valkenburg, 2000, De Vreese et al., 2001].

[6] We performed a principal component analysis with non-orthogonal rotation to establish the coherence of the indicator questions and their relationships to the frames. As expected, we found a four-factor solution in which all indicators show significant positive loadings (>.5) on the expected frame.

[7] Nearly all coders were involved, because multiple pairs of coders were used for reliability testing.

[8] It is a well-known issue that Krippendorff's alpha measures tend to be relatively low when assessing inter-coder agreement of binary classification tasks with unbalanced class distributions. This especially is the case with the morality frame, where we observe a substantial difference between the pairwise agreement measure and

Table 2.1: F1 Scores for SML-Based Issue Coding in News Articles and PQs

Item	*Wording*	*Kr. Alpha*
C1	Does the item reflect disagreement between parties, individuals, groups or countries?	.47 (.72)
C2	Does the item refer to two sides or more than two sides of the problem?	.41 (.70)
E1	Is there a reference to the financial costs/degree of expense involved, or to financial losses or gains?	.61 (.83)
E2	Is there a reference to economic consequences of pursuing or not pursuing a course of action?	.37 (.85)
H1	Does the item provide a human example or human face on the issue?	.20 (.75)
H2	Does the item employ adjectives or personal vignettes that generate feelings of outrage, empathy caring?	.33 (.57)
H3	Does the item mention how individuals and groups are affected by the issue or problem?	.16 (.84)
M1	Does the item contain any moral message?	.35 (.91)
M2	Does the item make reference to morality, God or other religious tenets?	.43 (.91)
M3	Does the item offer specific social prescriptions about how to behave?	.29 (.92)

Coding was performed for a large-scale research project on the influence of media coverage on parliamentarians. The final dataset consisted of 11,074 documents, of which 6,030 were political in nature. We used this set of manually coded articles to train and test our classifiers.

2.9 Evaluation Metrics and Cross Validation

We evaluated coding performance in terms of classification accuracy, receiver operating characteristics and Krippendorff's Alpha. Performance measures are reported for the automatic coding of indicators and frames.

Accuracy (AC) is the percentage of agreement between human classifications and computer-based classifications. It indicates the number of correctly classified documents. To demonstrate a classifier's improvement over chance agreement, we compared the reported accuracy measures with a random baseline.

The random baseline is a naive way of predicting the presence of an indicator or frame by chance. It randomly chooses the answer to an indicator question or whether a frame is present or not, taking into account only its prevalence in the training set. This baseline thus randomly assigns a classification without considering the document content, with a probability based only on the class distributions. Consequently, it will be more likely to randomly pick the majority class than the minority class. The classifier's accuracy improvement over the random baseline indicates its superiority to chance agreement.

Furthermore, we rely on receiver operating characteristics to evaluate classifier performance. More precisely, we report the area under the curve (AUC). AUC is a measure of how well a classifier discriminates

Krippendorff's alpha measure.

between the presence and the absence of a frame or indicator. AUC is a commonly used evaluation method for binary coding tasks [Sokolova and Lapalme, 2009]. The main advantage over other evaluation methods is its insensitivity to unbalanced datasets.

The AUC measure is based on the ROC curve, which shows the trade-off between increasing true positive rates and increasing false positive rates. The AUC indicates the probability that the classifier will rank a positive document above a negative document. A perfect classifier will score an AUC of 1 and random guessing will score an AUC of 0.5. The measure thus allows us to quantify how much better than random the classifier's choices are.

Additionally, we report Krippendorff's Alpha (KA), which is a common inter-coder agreement statistic in the field of communication science. Like the AUC measure, Krippendorff's Alpha corrects for agreement by chance.

Ten-fold cross-validation was used to obtain evaluation measures of classification performance. The dataset was partitioned into ten equal parts, one of which was reserved for testing the classifier (test set). The remaining parts were used as training data (training set). We repeated this cross-validation process ten times, such that each subsample was used once as the test set. The results from all validation rounds were averaged to produce a single estimation. This way, all observations were used for training as well as evaluation of the classifiers, but training observations were always separated from the test set.[9]

To test the generalizability of our classification models, as described in the third research question, we trained classifiers on articles from

[9] Please note that the cross-validation sample that was used to estimate weights for the ensemble of classifiers is nested in the cross-validation sample, which we used to assess coding performance.

two of the three available newspapers and then evaluated the classifiers' abilities to correctly code frames in articles from the third paper[10], which were not included in the training set. We performed this test for all possible combinations of the three newspapers.

Finally, to assess the relationship between the number of training documents and classification performance, we repeatedly trained each frame classifier while increasing the number of documents in the training set. We held out a fixed set of 1,000 articles for testing. For training, we used samples of different sizes from the held-in set. In total, we performed seven iterations with the following numbers of documents in the training set: 100, 200, 500, 1,000, 2,000, 3,000, and 4,000.

2.10 Results

To answer our research questions, we conducted a series of classification experiments in which we predicted four frames and their indicators. In Table 2.2, we report classification performance (AC, AUC and KA) per frame for the holistic and indicator-based approaches. In Table 2.3, we report classification performance for all indicators. Both tables include measures of the random baseline.

First, we address the random baseline. This baseline indicates agreement by chance in the classification process, based on the prevalence of frames in the training set. We observe a high variation in frame prevalence (M=41%, SD=23.01), with morality being the least prevalent frame (13%) and conflict the most prevalent frame (61%). Derived probabilities of correctly predicting the frames by chance range from .61 for the conflict frame to .87 for the morality frames (M=.69, SD=.13).

[10] We always used a random sample of 2,000 articles as a training set and a random

Table 2.2: F1 Scores for SML-Based Issue Coding in News Articles and PQs

	Conflict			*Economic Consequences*			*Human Interest*			*Morality*		
Prevalence	*61%*			*32%*			*59%*			*13%*		
	AC	AUC	KA	AC	AUC	KA	AC	AUC	KA	AC	AUC	KA
Baseline	.61	.50		.68	.50		.59	.50		.87	.50	
Indicator	.77	.76	.52	.85	.84	.67	.74	.74	.47	.89	.62	.33
Holistic	.80	.78	.57	.89	.85	.71	.79	.78	.55	.96	.76	.63

Second, we turn to measures of classification performance. Accuracy (AC) and AUC scores indicate high coding performance for all four frames. Therefore, we conclude that SML is suitable for frame coding. When applying the indicator-based approach, classification accuracy ranges from .74 for the human interest frame to .89 for the morality frame (M=.81, SD=.07). When applying the holistic approach, accuracy ranges from .79 for the human interest frame to .96 for the morality frame (M=.86, SD=.08).

All accuracy scores surpass the random baseline, meaning that we improve on chance agreement for each frame. Moreover, for all frames, the holistic approach outperforms the indicator-based approach in terms of classification accuracy, AUC and Krippendorff's Alpha (KA) measures. The average improvement in accuracy is about five percentage points. Therefore, we conclude that it is more effective to predict the frame variable directly, compared to predicting indicators and aggregating them afterward.

Third, we find performance differences between frames. When applying the holistic approach, AUC scores range from .76 for the morality frame to .85 for the economic consequences frame (M=.86, SD=.04). This means that, among all frames, our classifiers can most optimally differentiate between positive and negative examples of the economic consequences frame. Among the other three frames, we find little variation in AUC scores.[11]

Fourth, we investigated whether we could generalise our models to news sources that were not included in the training data. In Table 2.4, we report classification accuracy when training on data from two of the

sample of 1,000 articles as test set.

[11] We found the same pattern when applying the indicator-based approach.

Table 2.3: F1 Scores for SML-Based Issue Coding in News Articles and PQs

	C1	C2	E1	E2	H1	H2	H3	M1	M2	M3
Prevalence	*52%*	*69%*	*29%*	*16%*	*21%*	*49%*	*9%*	*7%*	*6%*	*5%*
Baseline	.52	.69	.71	.84	.79	.51	.91	.93	.94	.95
CA	.77	.75	.87	.86	.82	.76	.93	.94	.96	.95
AUC	.77	.75	.85	.78	.76	.76	.64	.59	.69	.50
KA	.54	.49	.68	.50	.49	.52	.39	.26	.49	.02

Table 2.4: Classification Accuracy for Coding Articles from Unknown Sources

	VK/NRC–>Tel	VK/TEL–>NRC	NRC/TEL–>VK
Conflict	.69	.74	.75
Economic Cons.	.88	.86	.86
Human Interest	.69	.71	.67
Morality	.97	.90	.89

three newspapers and testing on articles from the third paper. The results indicate that we can generalize our classification models to other news sources. However, in most cases, classification accuracy was slightly lower compared with predicting frames in sources that were included in the training data (see Table 2.2).

Finally, we present findings of experiments regarding the relationship between the amount of training data and coding performance. For all frames, classification accuracy is plotted in Figure 2.1. As expected, measures show that increasing the number of training documents leads to increased classification performance for all classifiers.

It is obvious immediately that compared to the other frames, classification accuracy of the morality frame increases more slowly when adding training documents. Most likely, this is because the morality frame is less prevalent in the training data. However, it stands out that classification accuracy for the economic consequences frame increases fastest when adding training documents, although it is not the most prevalent frame. This supports our finding that the SML approach works better for the economic consequences frame than for the other three frames.

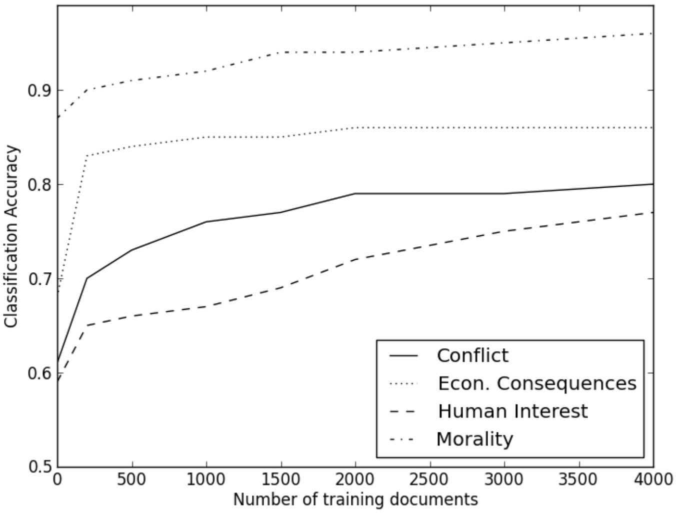

Figure 2.1: Relationship between classification accuracy and number of training documents.

2.11 Discussion

In this article we explored the application of SML to frame coding. Framing is one of the key concepts in communication science, and SML can advance future framing research by easing large-scale CA. Once a classifier is trained to code a frame, it can be employed for automatic coding of that frame in subsequent studies. Therefore, SML facilitates more integrated analyses of framing processes [Matthes, 2012, De Vreese, 2005].

Several scholars have advocated studying framing processes outside the laboratory [Kinder, 2007]. However, sophisticated designs, combining (panel) surveys and CA [e.g., Schuck et al., 2013, Wettstein, 2012], are expensive. SML-based frame coding not only facilitates the application of such mixed-methods designs but also allows the scale of its CA part to be easily increased.

Large-scale CA, which becomes more and more attractive as the amount of digitally available media content rises [Lazer et al., 2009], helps address substantial issues in framing research. Such issues include looking at frame variation over time [Matthes, 2012, Chong and Druckman, 2010] and the conditionality of framing processes [e.g., Chong and Druckman, 2007]. To what extent is a frame repeated or challenged in the media, and how does this affect the public over time? To what extent do frame usage and framing effects depend on the topic of a message, the actors with whom frames are associated in that message, and the medium used to transmit the message? Appropriate investigation of these questions requires frame coding over a long period and across various domains, respectively. SML can help the affordability of such CA without relying on small samples.

In this study, we applied SML to the coding of four widely used journalistic frames. We observed high levels of coding performance for all four frames. Using our classifiers, we can now automatically code these frames in future studies. We conclude that SML is generally suited to automate frame coding. When investigating a new frame in future studies, manual coding can be limited to that needed for training a classifier, and the remaining documents can be coded by applying the classifier. Performance levels of SML-based CA in our study are comparable to similar attempts of employing SML to automate the coding of concepts that are relevant to communication research [see, e.g., Scharkow, 2013, for SML-based coding of news values].

Our study informs the application of SML to frame coding and CA more generally in several ways. First, we conclude that SML approaches might work even if one does not possess tens of thousands of training documents, which were available in previous studies applying SML to CA [e.g., Hillard et al., 2008]. In this study, the amount of training data necessary to train a well-performing classifier varies from frame to frame.

One important factor is the overall presence of a frame. When studying a frame that occurs regularly within the text corpus used, manual coding of a few hundred documents might be sufficient to automate coding of the remaining documents. When studying an uncommon frame, active sampling [Tong and Koller, 2000] of positive examples of the frame can help keep manual coding efforts manageable. Several strategies for this are discussed in the literature [e.g., Hillard et al., 2008].

Furthermore, we conclude that some concepts are less difficult to predict than others. We found, for example, that classification performance of the economic consequences frame improves the most when increasing the size of the training set, although this frame is not the most prevalent one.

Second, we conclude that a trained classifier can be applied to automatic coding in sources other than those used for training. We provide evidence for this but also find that classification accuracy decreases for some frames. We believe the generalizability of a classifier strongly depends on the coding task and the training data used. Therefore, in future studies, similar experiments should be repeated (e.g., generalization from print to online media).

We might extrapolate these conclusions to several other concepts in communication research. This includes the coding of such concepts as sentiment, emotions, or news values, which have some conceptual similarity with frames. We recommend testing all of this in future research.

In this article, we also compared two approaches, indicator-based and holistic, to modeling the frame coding process. When applying SML it might seem appropriate to proceed as in manual CA, where we code indicators and aggregate them to frame measures. However, results of our experiments show that it is more effective to train a classifier to predict the presence of a frame directly.

In regard to generalising this finding, we would like to mention some limitations. It is difficult to say whether performance differences would be similar with other frames or even other concepts because the pattern we found might be due to properties of the data we used and the variables we coded. We compared the approaches when using a binary frame measure. When combining indicators in such a way that one gets a continuous outcome measure (e.g., by averaging them), the holistic approach might not outperform the indicator-based approach. Predicting the strength of a frame (or other concept) in a text is most likely more complicated than simply predicting its presence. Therefore, explicitly modelling indicators in the SML process might be of greater relevance.

The fact that we find the same pattern for all studied frames, which are substantially different, gives us some confidence in the generalizability of the finding. However, future studies are needed to test this. At least we can make the following argument: In some cases it works better to predict frames directly. Although we cannot establish clear rules about when this is the case based on our findings, it is worth comparing both approaches when trying to automate a coding task using SML. In future research, similar comparisons should be made using other datasets and frames.

Another limitation of our study is that we tested the SML approach with generic frames only. We believe that it would work similarly for other types of frames, such as issue-specific frames [e.g., Rhee, 1997]. The critical difference between generic frames and issue frames is that the former are used more widely and have little issue dependence. There is no reason, however, to believe that it would not work with issue-specific frames, because we expect them to be manifested in a certain vocabulary as well.

One might even expect better performance, because an issue frame might be more salient in an article than a generic frame. Moreover, with issue-specific frames, the population of texts to analyse is more uniform, which might decrease the complexity of the classification problem. Then again, it might be difficult to generate good training data, as one must deal with a limited population of texts containing the issue frame.

Another question is whether SML can be applied to more complex frames. Among the frames studied, we believe the morality frame to be the most complex. Because we are able to automatically code the morality frame with performance similar to the other three frames, we believe that an SML approach generally works with more complex frames. More advanced feature representations are likely to increase performance when coding complex frames. We leave this question for future research.

Finally, an important limitation of our study concerns inter-coder reliability. First, we are aware that we should have coded each article by more than two coders when assessing reliability. Second, the quality of our training data is not optimal. In various cases, coders disagreed on the presence of frames, as indicated by the reported reliability measures. Disagreement likely results from a combination of unsystematic coding errors and systematically different interpretation of frame indicators across coders.

The most relevant question is how the latter, especially, might influence our findings and conclusions. We expect classification performance to decrease as a result of inconsistencies in the training data. If texts with similar features are associated with different labels, it becomes more difficult for the SML algorithm to estimate a model that can clearly differentiate between two classes.

Although classification performance is most likely influenced by the moderate training data, we believe our conclusion to be largely unaffected. We conclude that SML is suited for automating frame coding, but the more error-prone the training data are, the more error-prone the automatic classifications. Moreover, our conclusion that trained classification models might fit texts from sources not included in the training data is unlikely to be affected. There is no reason to believe that models would be less generalizable if inter-coder agreement were higher.

Despite those shortcomings, this study is the first to apply SML to frame coding. Our study not only provides promising results but also provides important insights regarding the use of SML in future communication research.

2.12 Bibliography

Jörg Matthes. What's in a frame? a content analysis of media framing studies in the world's leading communication journals, 1990-2005. *Journalism & Mass Communication Quarterly*, 86(2):349–367, 2009.

Adam Simon and Michael Xenos. Media framing and effective public deliberation. *Political Communication*, 17(4):363–376, 2000.

Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.

Jörg Matthes. Framing politics: An integrative approach. *American Behavioral Scientist*, 56(3):247–259, 2012.

William A Gamson and Andre Modigliani. Media discourse and public opinion on nuclear power: A constructionist approach. *American Journal of Sociology*, 95(2):1–37, 1989.

Dietram A Scheufele. Framing as a theory of media effects. *Journal of Communication*, 49(1):103–122, 1999.

Claes H De Vreese, Jochen Peter, and Holli A Semetko. Framing politics at the launch of the Euro: A cross-national comparative study of frames in the news. *Political Communication*, 18(2):107–122, 2001.

Klaus Krippendorff. *Content analysis: An introduction to its methodology*. Sage, Thousand Oaks, CA, 2012.

Conny Roggeband and Rens Vliegenthart. Divergent framing: The public debate on migration in the Dutch parliament and media, 1995–2004. *West European Politics*, 30(3):524–548, 2007.

Nel Ruigrok and Wouter Van Atteveldt. Global angling with a local angle: How US, British, and Dutch newspapers frame global and local terrorist attacks. *The Harvard International Journal of Press/Politics*, 12(1):68–90, 2007.

Dhavan V Shah, Mark D Watts, David Domke, and David P Fan. News framing and cueing of issue regimes: Explaining Clinton's public approval in spite of scandal. *Public Opinion Quarterly*, 66(3):339–370, 2002.

Michael Laver, Kenneth Benoit, and John Garry. Extracting policy positions from political texts using words as data. *American Political Science Review*, 97(2):311–331, 2003.

Kathleen T Durant and Michael D Smith. Predicting the political sentiment of web log posts using supervised machine learning techniques coupled with feature selection. In B Massand, editor, *Advances in web mining and web usage analysis. Lecture notes in computer science, Vol. 4811*, pages 187–206, Berlin, Germany, 2007. Springer.

Stuart Russell and Peter Norvig. *Artificial Intelligence: A modern approach*. Prentice Hall, Upper Saddle River, NJ, 2002.

Justin Grimmer and Brandon M Stewart. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297, 2013.

Dustin Hillard, Stephen Purpura, and John Wilkerson. Computer-assisted topic classification for mixed-methods social science research. *Journal of Information Technology & Politics*, 4(4):31–46, 2008.

Holli A Semetko and Patti M Valkenburg. Framing european politics: A content analysis of press and television news. *Journal of Communication*, 50(2):93–109, 2000.

Thorsten Joachims. *Text categorization with support vector machines: Learning with many relevant features*. Springer, Berlin, Germany, 1998.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pages 79–86, Stroudsburg, PA, 2002. ACL.

Michael Scharkow. Thematic content analysis using supervised machine learning: An empirical evaluation using German online news. *Quality & Quantity*, 47(2):761–773, 2013.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. Emotions from text: Machine learning for text-based emotion prediction. In RJ Mooney, editor, *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 579–586, Stroudsburg, PA, 2005. ACL.

George Forman. An extensive empirical study of feature selection metrics for text classification. *The Journal of Machine Learning Research*, 3:1289–1305, 2003.

Thorsten Joachims. *Learning to classify text using support vector machines: Methods, theory and algorithms*. Dordrecht, The Netherlands, 2002.

Thomas G Dietterich. Ensemble methods in machine learning. In J Kittler and F Roll, editors, *Multiple classifier systems. Lecture Notes* *in Computer Science, Vol 1857*, pages 1–15, Berlin, Germany, 2000. Springer.

Robi Polikar. Ensemble learning. In C Zhang and Y Ma, editors, *Ensemble Machine Learning*, pages 1–34. Springer, Berlin, Germnay, 2012.

Richard P Lippmann. An introduction to computing with neural nets. *ASSP Magazine*, 4(2):4–22, 1987.

Robert M Entman. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58, 1993.

Djoerd Hiemstra, Stephen Robertson, and Hugo Zaragoza. Parsimonious language models for information retrieval. In J Kalervo and A James, editors, *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185, New York, NY, 2004. ACM.

Radim Vrehuuvrek and Petr Sojka. Software framework for topic modelling with large corpora. In *Proceedings of the LREC Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, 2010. ELRA.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, and Vincent Dubourg. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

W Russell Neuman, Marion R Just, and Ann N Crigler. *Common knowledge: News and the construction of political meaning*. University of Chicago Press, Chicago, IL, 1992.

Johan Galtung and Mari H Ruge. The structure of foreign news the presentation of the Congo, Cuba and Cyprus crises in four Norwegian newspapers. *Journal of Peace Research*, 2(1):64–90, 1965.

Christiane Eilders. *Nachrichtenfaktoren und Rezeption*. Westdeutscher Verlag, Opladen, Germany, 1997.

John H McManus. *Market-driven journalism: Let the citizen beware?* Sage, Thousand Oaks, CA, 1994.

Joachim Friedrich Staab. *Nachrichtenwert-Theorie: formale Struktur und empirischer Gehalt*. Alber, Freiburg, Germany, 1990.

Thomas E Patterson. *Out of order*. Vintage, New York, NY, 1993.

Jay G Blumler, Jay Blumler, and Michael Gurevitch. *The crisis of public communication*. Routledge, London, UK, 1995.

Joseph N Cappella and Kathleen Hall Jamieson. *Spiral of cynicism: The press and the public good*. Oxford University Press, New York, NY, 1997.

Rens Vliegenthart, Hajo G Boomgaarden, and Jelle W Boumans. *Changes in political news coverage: Personalization, conflict and negativity in British and Dutch newspapers*. Palgrave Macmillan, London, UK, 2011.

Rens Vliegenthart, Andreas RT Schuck, Hajo G Boomgaarden, and Claes H De Vreese. News coverage and support for European integration, 1990–2006. *International Journal of Public Opinion Research*, 20(4):415–439, 2008.

Shanto Iyengar. *Is anyone responsible?: How television frames political issues*. University of Chicago Press, Chicago, IL, 1994.

Dolf Zillmann and Hans-Bernd Brosius. *Exemplification in communication*. Lawrence Erlbaum Associates, Mahwah, NJ, 2000.

Doris A Graber. *Mass media and American politics*. CQ Press, Washington, DC, 1993.

William A Gamson. *Talking Politics*. Cambridge University Press, New York, NY, 1992.

Matthew C Nisbet and Mike Huge. Attention cycles and frames in the plant biotechnology debate managing power and participation through the press/policy connection. *The Harvard International Journal of Press/Politics*, 11(2):3–40, 2006.

Matthew C Nisbet, Dominique Brossard, and Adrianne Kroepsch. Framing science the stem cell controversy in an age of press/politics. *The International Journal of Press/Politics*, 8(2):36–70, 2003.

Paul R Brewer. Framing, value words, and citizens' explanations of their issue opinions. *Political Communication*, 19(3):303–316, 2002.

Paul R Brewer. Values, political knowledge, and public opinion about gay rights: A framing-based account. *Public Opinion Quarterly*, 67 (2):173–201, 2003.

Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, 2009.

Claes H De Vreese. News framing: Theory and typology. *Information Design Journal and Document Design*, 13(1):51–62, 2005.

Donald R Kinder. Curmudgeonly advice. *Journal of Communication*, 57(1):155–162, 2007.

Andreas RT Schuck, Hajo G Boomgaarden, and Claes H De Vreese. Cynics all around? the impact of election news on political cynicism in comparative perspective. *Journal of Communication*, 63(2):287–311, 2013.

Martin Wettstein. Frame adoption in referendum campaigns the effect of news coverage on the public salience of issue interpretations. *American Behavioral Scientist*, 56(3):318–333, 2012.

David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, and Myron Gutmann. Life in the network: the coming age of computational social science. *Science*, 323(5915):721, 2009.

Dennis Chong and James N Druckman. Dynamic public opinion: Communication effects over time. *American Political Science Review*, 104 (4):663–680, 2010.

Dennis Chong and James N Druckman. A theory of framing and opinion formation in competitive elite environments. *Journal of Communication*, 57(1):99–118, 2007.

Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2:45–66, 2000.

June W Rhee. Strategy and issue frames in election campaign coverage: A social cognitive account of framing effects. *Journal of Communication*, 47(3):26–48, 1997.

3

Frame Identification through Unsupervised Learning

Chapter 3: Frame Identification through Unsupervised Learning

This chapter has been accepted for publication as:

Burscher, B., Vliegenthart, R., & De Vreese, C. H. (2015). Frames Beyond Words: Applying Cluster and Sentiment Analysis to News Coverage of the Nuclear Power Issue. *Social Science Computer Review.*

The version presented here has been adapted to follow the overall standards and terminology included in the other chapters of the dissertation.

3.1 Abstract

Methods to automatically analyze media content are advancing significantly. Among others, it has become increasingly popular to analyze the framing of news articles by means of statistical procedures. In this chapter, we investigate the conceptual validity of news frames that are inferred by a combination of k-means cluster analysis and automatic sentiment analysis. Furthermore, we test a way of improving statistical frame analysis such that revealed clusters of articles reflect the framing concept more closely. We do so by only using words from an article's title and lead and by excluding named entities and words with a certain part of speech from the analysis. To validate revealed frames, we manually analyze samples of articles from the extracted clusters. Findings of our tests indicate that when following the proposed feature selection approach, the resulting clusters more accurately discriminate between articles with a different framing. We discuss the methodological and theoretical implications of our findings.

3.2 Introduction

News media can shape public opinion regarding an issue by emphasizing some elements of the broader controversy over others [Jasperson et al., 1998, Shah et al., 2002]. Research shows that aspects of an issue, which are more salient in the media, cause individuals to focus on these aspects when constructing their opinions.

Shah et al. [2002], for example, showed that public approval of President Clinton depended on whether news coverage during the Lewinsky sex scandal focused on the sexual nature of Clinton's indiscretion or the attacks of Republicans on Clinton's behavior. Furthermore, Sniderman et al. [1993] found that a majority of the public supports the rights of a person with HIV when the role of civil liberties is stressed in the news and supports mandatory testing when the importance of public health is stressed. In the literature, this phenomenon is referred to as emphasis framing.

In this chapter, we introduce and evaluate a method to automatically analyze emphasis framing in news coverage. We use this method to identify a set of news frames within the nuclear power debate and study developments in the frames' prevalence and tone over time. Therefore, this study fits into a broader line of research investigating the use of automated content methods in social science research [Grimmer and Stewart, 2013].

More specifically, we apply a combination of k-means cluster analysis and automatic sentiment analysis. Cluster analysis can be used to group articles according to their word use. As articles that contain the same words, most likely also stress the same elements of a controversy, this technique can reveal groups of articles with a similar framing. Sentiment analysis is a way to automatically determine the polarity of the tone of an article. Together, these techniques provide a powerful tool to study dynamics in the news framing of social and political issues. To our knowledge, cluster- and sentiment analysis have not been combined before in framing research.

Furthermore, we explore a novel way of performing cluster analysis such that the resulting clusters more accurately differentiate between articles with a different framing. To do so, we apply natural language processing to select parts of a news article, which we consider highly relevant to capture the meaning of frames, and only use these parts to present articles in the analysis.

In our approach, we abstain from human involvement like defining frame elements [e.g., Motta and Baden, 2013] or manually (pre)coding data [e.g., Matthes and Kohring, 2008]. This has the advantage that the framing analysis is largely automated and mostly unaffected by potential biases of human researchers and/or coders.

In order to validate results of our analyses, we conduct a manual content analysis. We conclude that the combination of cluster- and sentiment analysis can be used to identify and code emphasis frames in news coverage automatically and validly. We discuss extensively the theoretical and methodological implications of our findings.

3.3 The Nuclear Power Debate

Our study addresses emphasis framing [e.g., Chong and Druckman, 2007] — a rather broad form of news framing, which is particularly prominent in the field of communication science. Throughout this chapter, we define framing as emphasis in salience of some elements of a story above others [e.g., De Vreese, 2005, Nelson et al., 1997].

In order to investigate the application of cluster and sentiment analysis to framing research, we must choose an issue to study. Ideally, this would be an issue that has created ample (controversial) news coverage in the past and which has been extensively studied in framing research before. Such a case would allow us to compare the frames we find by means of cluster analysis to frames that have been identified in previous studies, by means of different methods.

The nuclear power debate provides such a case. Various studies have analyzed the nuclear power debate in the past 50 years [e.g., Bickerstaff et al., 2008, Gamson and Modigliani, 1989, Nisbet, 2009]. In Table 3.1, we provide an overview of nuclear power frames. To create this overview, we reviewed the most-cited journal articles [1] that study and/or discuss the media framing of nuclear power.

3.4 Frame Analysis

Frame analysis requires (1) the identification of frames that news media focus on when covering an issue and (2) coding the presence of these frames in news articles [e.g., Jasperson et al., 1998]. Traditionally, scholars identify emphasis frames by qualitatively analyzing rather small samples of articles [e.g., Simon and Xenos, 2000]. Afterward, to measure their usage, each frame is operationalized — either in the form of indicator questions in manual content analysis [Semetko and Valkenburg, 2000] or search strings in automatic content analysis [Shah et al., 2002, Ruigrok and Van Atteveldt, 2007].

Alternatively, frames can be identified by means of statistical analysis. The most basic approach is to interpret word co-occurrences. Hellsten

[1] Articles cited more than 30 times according to Google Scholar.

Table 3.1: Nuclear Power Frames in the Literature

- risks of nuclear weapon development [b,e]
- health and environmental risks of radioactive waste [b, f, d, e]
- social progress and economic development due to nuclear power usage [a, c, e]
- terrorism threats and risks of nuclear accidents [d, e, f, c, a]
- economic risks of nuclear power production, not cost-effective [c, e]
- nuclear power to cut greenhouse gas emissions and prevent climate change [d, e]
- nuclear power to satisfy energy demands and provide energy independence [a, e, c, f]
- renewable energies as alternative to nuclear power [a, e, c]

[a]Gamson and Modigliani [1989], [b]Joppke [1991], [c]Nisbet [2009], [d]Bickerstaff et al. [2008], [e]Culley et al. [2010], [f]Pidgeon et al. [2008]

et al. [2010], for example, plotted cosine distances between words in a network graph and then interpreted agglomerations of words within the network as frames. More sophisticated approaches applied either factor analysis [Motta and Baden, 2013, Van Der Meer and Verhoeven, 2013] or cluster analysis [Matthes and Kohring, 2008, Miller, 1997].

Factor analysis describes variability among observed variables in terms of a potentially lower number of unobserved variables, which can be interpreted as frames.

Cluster analysis groups a set of articles in such a way that articles in the same group are more similar to each other than to those in other groups [Kaufman and Rousseeuw, 2009]. In other words, based on the similarity of articles, a number of clusters are created and each article is assigned to one cluster. The clusters present groups of articles with a different framing. By interpreting the most prototypical words of articles from each cluster, one can infer frames.

Cluster analysis results in a classification model, which can be used to automatically code future articles according to the created cluster structure. The method can thus be used for further analyses: One can, for example, easily compare the popularity of different frames over time and across news sources. The assignment of articles to clusters is a critical difference between factor analysis and cluster analysis. Factor analysis reduces the dimensionality of a dataset and provides information about how each factor corresponds to the original variables (e.g., words in the corpus) but does not classify the articles into groups.

In this study, we use cluster analysis to identify and code emphasis frames in news coverage about nuclear power from the past 20 years. Furthermore, we apply automatic sentiment analysis to analyze the tone of coverage. This allows us to study dynamics in the prevalence of different frames as well as dynamics in the tone of news articles containing a specific frame over time. So far, scholars have not combined cluster analysis and sentiment analysis to identify frames.

We expect that the analysis of tone improves the interpretation of clusters as emphasis frames, because earlier research has shown that news coverage of nuclear power generally focuses on benefits or risks of its usage [Gamson and Modigliani, 1989]. Moreover, frames often contain moral evaluations of policy issues [Semetko and Valkenburg, 2000]. All in all, we present a method to show how the portrayal of an issue changes over time — in terms of topical elements that are emphasized and in terms of their valence.

To validate this automatic analysis, we conduct a manual content analysis for a sample of articles and compare automatic codings to manual codings. In addition, we compare outcomes of the cluster analyses to outcomes of previous studies that investigated the framing of the nuclear power debate [e.g., Gamson and Modigliani, 1989]. This leads to the following research question: *To what extent can cluster analysis be used to infer emphasis frames from a collection of issue-specific news articles?*

By answering this question, we can determine the ability of cluster- and sentiment analysis to identify and code emphasis frames in future research and we can draw conclusions about whether cluster analysis leads to similar frames as manual approaches. To our knowledge, previous studies have not explicitly cross-validated the use of statistical frame identification with manual approaches.

3.5 Building Blocks of Frames

In cluster analysis, the quality of resulting clusters depends on the selection of document features [Kaufman and Rousseeuw, 2009]. There are various document features that might be used to compare two texts—but not all are important for the classification of interest. Some features may be redundant or irrelevant and others can misguide results of the cluster analysis.

Various articles [Dy and Brodley, 2004, Gnanadesikan et al., 1995, Hatzivassiloglou et al., 2000] have studied the question of which set of document features is most useful for several classification tasks (e.g., topic or sentiment). Among others, scholars selected words based on their frequency, part of speech, or position in the document. Furthermore, word features have been enriched by adding semantic features using Wikipedia [Hu et al., 2009] or WordNet [Sedding and Kazakov, 2004].

In statistical frame analysis, in order to find clusters (or factors) that discriminate between different frames, one must represent documents in terms of features that are indicative of such frames. According to Entman [1993, p. 52], news frames manifest themselves in certain text attributes as "the presence or absence of certain keywords, stock phrases, (and) stereotyped images." Therefore, we used word frequencies as features in our cluster analyses, which is called the "bag-of-words" approach [e.g., Hellsten et al., 2010, Miller, 1997].

This has two advantages: First, using words is highly reliable, because words are manifest features [Riff et al., 2014] and consequently, frame analysis becomes a replicable process that is unlikely to be biased by the subjective input of individual researchers. Second, it is cost-efficient, because no manual analysis is involved.

The key issue, however, is construct validity: To what extent do word-based clusters actually reflect different emphasis frames? In the literature, this is a highly debated question. On the one hand, words are widely used as features in statistical frame analysis. On the other hand, critiques have repeatedly objected to its use [Carragee and Roefs, 2004, Hertog and McLeod, 2001]. The main point of criticism is that not all words are equally important to a news frame. As Cappella and Jamieson [1997] put it, considering any production feature of verbal or visual texts as a candidate for news frames is a too broad view.

As a response, scholars started using higher level frame elements as features [e.g., Matthes and Kohring, 2008, Motta and Baden, 2013]. Matthes and Kohring [2008], for example, used Entman's popular operational definition of news frames and manually coded all articles for problem definitions, causal interpretations, moral evaluations, and/or treatment recommendations [Entman, 1993]. Afterward, the authors used these frame elements as features in a cluster analysis.

Using higher level frame elements as features has brought significant advancements to statistical frame analysis, because such features are generally more conclusive building blocks of frames and, consequently, lead to a higher construct validity when identifying frames. However, as the used frame elements are usually issue-specific, they must be defined and coded individually before each analysis.

Our aim, in contrast, is to explore a way in which we can improve statistical frame analysis but keep the analysis as inductive as possible without relying on a priori made decisions on the side of researchers or human coders. For doing so, we apply natural language processing to select such parts of a news article, which we consider highly relevant to capture the meaning of emphasis frames and only use these parts as features.

First, we only use words from the headline and the lead as features. Generally, news stories present information in terms of relative importance [Poettker, 2003]. This structure is called the inverted pyramid style. We infer from this style that the article's dominant perspective on the issue is presented at the beginning. Pan and Kosicki [1993, p. 59] argued the following: "A headline is the most salient cue to activate certain semantically related concepts in readers' minds; it is thus the most powerful framing device of the syntactical structure. A lead is the next most important device to use. A good lead will give a story a newsworthy angle". Similarly, Tankard (2001) counts headline and lead as two important framing mechanisms.

We expect that only using headline and lead as features leads to clusters that more clearly differentiate between distinct emphasis frames. This is because other elements in the remaining paragraphs of an article, which do not address the dominant frame, would act as noise in the analysis.

Related research on topic clustering has shown that giving higher weight to the title of a news article can increase the accuracy of topic clusters, because the title is more representative of the topic than the main text. In their experiments, Banerjee et al. [2007] obtained best results by doubling the weights of the terms appearing in the title of a given news article. Similarly, Bouras and Tsogkas [2012] increased the weights of terms that also appeared in the title of an article when analyzing topic clusters of news articles. We expect similar effects for frame clusters.

Second, we conduct part-of-speech tagging [Toutanova et al., 2003] to select words that are a noun, an adjective, or adverb. We believe that words from the selected classes (nouns, adjectives, and adverbs) are most indicative of frames. This is because other word classes, like verbs, conjunctions, or pronouns, are much less likely than the selected classes to add meaning to a frame. Previous research has shown that giving higher weights to nouns than other word classes can increase the quality of topic clusters [e.g., Bouras and Tsogkas, 2012, Hatzivassiloglou et al., 2000].

Third, we apply named-entity recognition [Nadeau and Sekine, 2007] to remove all names of persons, organizations, and locations as well as times and dates. Names of countries and organizations, for example, refer to very specific events, while frames are more abstract semantic concepts. Therefore, it is more likely that we obtain clusters that actually discriminate between emphasis frames, when we remove named entities from the feature space. To our knowledge, this has not been tested before in document clustering.

When representing articles in the cluster analysis, we only use the above-mentioned parts of each article as features and ignore all other words. We call this the *selection approach.* In order to see whether this way of selecting features improves the validity of the cluster analysis, we conduct a baseline analysis where we use all words from each article as features (*baseline approach*).

We compare cluster centers in the selection approach with clusters centers in the baseline approach. Furthermore, we conduct a manual content analysis to compare the accuracy of frame codings in both approaches. This leads to the following research question: *To what extent does selecting frame-related document features improve the construct validity and coding accuracy of statistical frame analysis?*

In sum, contrasting the approaches aims at finding a way of representing news articles in terms of features that are highly indicative of frames. We expect that selecting frame-specific features (selection approach) does a better job in discriminating between emphasis frames than using all words as features (baseline approach).

3.6 Data and Method

Data

Our data consisted of English-language news articles covering the issue of nuclear power, which were published in *The New York Times*, *The Washington Post*, or *The Guardian* between 1992 and 2013. We used LexisNexis to search all three sources for articles that contain the key words "nuclear power" or "nuclear energy" at least 2 times in total and at least once in the headline or lead. By applying these rather strong restrictions, we made sure that nuclear power actually is the main topic of the article. This led to 4,286 articles, which we used in the analyses.

Automatic Content Analysis

Based on this collection of news articles, we created two datasets – one for the baseline approach and one for the selection approach. In both datasets, we used all 4,286 articles and applied the following preprocessing steps.

We converted all words to their lemmas [Manning et al., 2008] and removed numbers and common English stop words. Furthermore, we removed words that appeared in less than five documents or in more than 40% of all documents. Due to their frequency of use (very high or very low), such words do not differentiate well between clusters of news articles.

As explained in the previous section, in the selection approach dataset, we also removed (a) words that did not appear in the title or lead, (b) words with a part-of speech other than noun, adjective, or adverb and (c) names of persons, organizations, and countries. For all of the above-mentioned steps, we used the Python natural language toolkit (NLTK).

Afterward, we created document vectors with TF.IDF weighted word frequencies [Manning et al., 2008] for news articles in both datasets. Each word was assigned the number of times it occurs in the document (TF), weighted by the inversed frequency of articles in the dataset containing the word (IDF). The idea behind TF.IDF weighting is to evaluate the power of a word to discriminate between articles. Rare words are assumed to be more discriminating and, therefore, are assigned higher weight. We standardized the document vectors using L2 normalization [Ng, 2004].

To reveal clusters from our datasets, we applied k-means clustering - a centroid-based clustering technique, where the number of clusters (k) must be specified a priori [Hartigan and Wong, 1979]. Given a set of articles $(x_1, x_2, ..., x_n)$, where each article is a d-dimensional vector, k-means clustering groups the n articles into k $(\leq n)$ clusters $S = \{S_1, S_2, ..., S_k\}$ so as to minimize the within-cluster sum of squares. More formerly, it aims at finding:

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \mu_i\|^2 , where$$

μ_i is the mean of points in S_i.

Each cluster is represented by a cluster center, which is described by the mean μ_i of the articles in the cluster. The algorithm defines the cluster centers and assigns each article to the cluster for which its distance to the cluster center is the smallest. We conducted separate cluster analyses for the baseline and the selection approach and used the cluster center vectors in order to identify emphasis frames in both approaches. In doing so, we listed for each cluster center the 15 document features with the highest means, that is, the most prototypical words for the cluster. Then, we gave each cluster center a frame label based on these 15 words.

A common technique to select the number of clusters (k) is the *elbow method*. We repeatedly run the analysis with different numbers of clusters (1–15) and added the amount of explained variance for each value to a scree plot (see Figure 3.1). Because the scree plot depicted an elbow at seven clusters in the baseline approach analysis, we decided to use a seven-cluster solution. In order to make both analyses more comparable, we also used a seven-cluster solution in the selection approach analysis.

Our implementation of k-means clustering makes use of the mini-batch k-means algorithm [Sculley, 2010] and the k-means++ optimization [Arthur and Vassilvitskii, 2007].[2] We used the scikit-learn machine learning library in Python for document vectorization and the cluster analyses [Pedregosa et al., 2011].

Finally, we applied the SentiWords[3] tool to automatically code the tone of articles in the selection approach dataset. SentiWords is a lexical resource containing roughly 155,000 words associated with a sentiment score between -1 (*negative*) and 1 (*positive*). Scores are learned from SentiWordNet and represent state-of-the-art computation of words' prior polarities [Baccianella et al., 2010]. See Guerini et al. [2013] for infor-

[2] In each cluster analysis, we run the k-means algorithm 10 times with different centroid seeds in each run. The final results were the best output of the 10 consecutive runs in terms of explained variance.

[3] https://hlt.fbk.eu/technologies/sentiwords

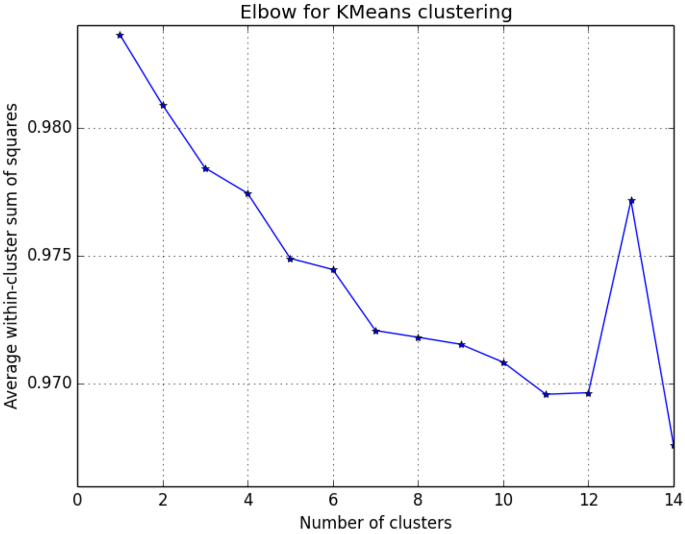

Figure 3.1: Scree Plot of Explained Variance for Baseline Approach.

mation about the method used to build SentiWords and Warriner et al. [2013] for a detailed description of the used dataset.

We annotated each word from the title and lead of all articles with the corresponding sentiment scores from the SentiWords lexicon. Then, we computed the mean of sentiment scores over all annotated words in each article and used it as a summary score for the article's tone. Words from the articles that were not included in SentiWords, received a sentiment score of zero.

Manual Content Analysis

We also conducted a manual content analysis to test whether the articles in each frame cluster actually contained the predicted frame. For the baseline and the selection approach, we sampled a random subset of 15 articles from each cluster and asked human coders to indicate the most relevant emphasis frame per article. Because we had two approaches with seven clusters each, 210 articles were manually coded.

Per article, coders could choose one frame from a list containing all unique frames that we identified for the corresponding approach. In the Results section, we describe these sets of frames more closely. Additionally, coders could code an article as "containing none of the listed frames/not primarily dealing with nuclear power."

We used two trained coders, who were fluent in English. In order to assess intercoder reliability, both coders coded 15% of the articles (N=32). Krippendorff's α for intercoder agreement was equal to .82.

3.7 Results

Baseline Approach

We performed two k-means cluster analyses: one in which we used all words of each article as features (baseline approach) and one in which we used selected parts (selection approach) of each article as features. Per analysis, we looked at the 15 features with the highest means for each of the seven cluster centers to infer emphasis frames. See Tables 3.2 and 3.4 for an overview of the cluster centers for each approach.

When using the baseline approach, we found multiple clusters that refer to the same element of the nuclear power controversy but relate to different geographical contexts. Clusters B5 and B7 are good examples of this phenomenon. Cluster B5 refers to nuclear power and the issue of weapon development in Iran. Cluster B7 also refers to nuclear power and weapon development, but in India and North Korea. B2 and B6 are another pair of examples, both clusters refer to safety issues and radiation risks of nuclear accidents. However, Cluster B2 does so in the

Table 3.2: Clusters Baseline Approach

B1	*B2*	*B3*	*B4*	*B5*	*B6*	*B7*
British	Commission	Indian	Carbon	Iran	Japan	India
Company	Chernobyl	Point	Gas	Iranian	Fukushima	Korea
Pound	Safety	Entergy	Emission	Russia	Tokyo	North
EDF	Waste	County	Wind	Weapon	Radiation	Treaty
Station	Company	Emergency	Climate	Uranium	Tepco	Weapon
Industry	Station	Westchester	Electricity	Program	Japanese	China
Price	Utility	Buchanan	Coal	Tehran	Water	Test
Cost	Official	Plan	Industry	Enrichment	Tsunami	Pakistani
Share	Site	Commission	Oil	Bushehr	Daichi	United
Million	Fuel	Siren	Station	Russian	Accident	Korean
Billion	Radiation	Steets	Renewable	United	Earthquake	Ban
Electricity	People	Federal	Cost	Sanction	DSafety	Agreement
BNFL	Industry	Official	Fuel	Official	Radioactive	Administration
Market	Public	Hudson	Renewables	Ahmadinejad	Worker	South
N=558	N=1918	N=250	N=648	N=233	N=383	N=296

context of the Chernobyl catastrophe and Cluster B6 in the context of the Fukushima disaster.

Furthermore, Cluster B3 refers to a very specific incident, instead of a more general emphasis frame—emergency evacuations of the Indian Point nuclear plant in Buchanan, New York. We can explain this clustering around specific events by the fact that the centers of the mentioned clusters mainly contain names of countries and organizations (e.g., Iran, Bushehr, India, and Fukushima). This indicates that the clusters do not primarily discriminate between distinct emphasis frames but between geographic contexts and particular incidents.

Nonetheless, several clusters uniquely refer to general elements of the nuclear power controversy. Cluster B4, for example, clearly refers to the impact of nuclear power on the climate and Cluster B1 refers to economic aspects of nuclear power usage. Cluster sizes are fairly unequally distributed (SD=552.5) with Cluster B2 as big as all other clusters combined. This suggests a residual category including articles that could not be properly assigned to other clusters.

Overall, we identified five unique frames here, which are listed in Table 3.3. We collapsed Clusters B5 and B7 (weapon development) as well as Clusters B2 and B6 (nuclear safety and accidents), because they referred to identical frames.

Selection Approach

When using the selection approach, we got a clearer cluster structure (see Table 3.4). Six of the seven clusters have coherent and unique cluster centers, all of which refer to distinct elements of the nuclear power controversy. There is little overlap as regards content between the clusters, which means that different clusters do not refer to the same

Table 3.3: Identified Frames Baseline- and Selection Approach

	Baseline Approach *Kr.Alpha = 0.52*
Frame 1	economic aspects of nuclear power production
Frame 2	safety of nuclear plants, nuclear waste, nuclear power accidents & radiation risks
Frame 3	nuclear power & weapon development
Frame 4	role of nuclear power in electricity production & effects on climate change
Frame 5	evacuation of nuclear reactors

	Selection Approach *Kr.Alpha = 0.71*
Frame 1	***safety*** of nuclear plants
Frame 2	role of nuclear power in electricity production & effects on ***climate*** change
Frame 3	***economic*** aspects of nuclear power production
Frame 4	nuclear power and ***weapon*** development
Frame 5	processing of nuclear materials & nuclear ***waste***
Frame 6	nuclear power ***accidents*** & radiation risks

emphasis frame. The cluster centers are, furthermore, easy to interpret, because they contain mostly substantial words and no names of places, persons, or organizations.

The primary question is whether we found different frames with this representation. On the one hand, some clusters are exactly the same as in the baseline approach. Two examples are Cluster S3, which deals with economic aspects of nuclear power, and Cluster S2, which deals with the effects of nuclear power on the climate.

On the other hand, we also found clusters that did not appear in the baseline approach. In the baseline approach, Cluster B2 refers to safety issues, nuclear accidents, and nuclear waste altogether. In contrast, in the selection approach, we found separate clusters for safety issues (S5), nuclear accidents (S7), and nuclear waste processing (S6).

This might be explained by the fact that we only used the title and lead of each news article as features here. When different elements of an issue are often referred to in the same article and one uses all parts of the article as features, it is likely that the elements are grouped together in one cluster. However, it is less likely that all elements are mentioned together in the title or lead. The selection approach thus provides a more nuanced grouping of articles around unique elements of the controversy.

Furthermore, compared to the baseline approach, clusters are more equally distributed with regard to size (SD=306.6). Again, one cluster (S1) is significantly larger than the average cluster size. Overall, we identified six unique frames here, which are listed in Table 3.3. We collapsed clusters S1 and S5, which both refer to safety issues of nuclear power.

Table 3.4: Clusters Selection Approach

S1	*S2*	*S3*	*S4*	*S5*	*S6*	*S7*
Station	Energy	Company	Weapon	Commission	Fuel	Reactor
State	Gas	Government	Program	Regulatory	Uranium	Radiation
Mile	Government	Price	State	Federal	Waste	Radioactive
First	Oil	Industry	President	Reactor	Plutonium	Accident
Official	Source	Pound	Country	Safety	Radioactive	Safety
Government	Renewable	Cost	Official	Regulator	Spent	Water
Plan	Climate	Reactor	Agreement	State	Reactor	Disaster
Security	Electricity	Electricity	Test	License	Rod	Leak
People	Policy	Share	Energy	Agency	State	Level
World	Emission	Plan	Foreign	Company	Enrichment	Worker
Last	Change	State	Treaty	Official	Storage	Exposure
Federal	Carbon	Utility	Nation	Problem	Company	Earthquake
Reactor	Coal	Generation	International	Utility	Material	Official
Attack	Minister	Last	World	Attack	Site	Station
Former	Generation	Energy	Uranium	Mile	Government	Operator
N=1296	N=645	N=609	N=568	N=548	N=328	N=292

Validation Analysis

We conducted three additional analyses. First, we calculated Krippendorrf's alpha as a measure of agreement between computer-based and human frame codings. Higher values of Krippendorrff's alpha indicate higher agreement between humans and the computer. As shown in Table 3.3, Krippendorrff's alpha is equal to .52 for the baseline approach and .71 for the selection approach. This shows that when using the selection approach, significantly more articles actually contained the predicted emphasis frame. The selection approach thus leads to more accurate codings of frames.

Second, we plotted the prevalence of frames from the selection approach over time. Figure 3.2 shows that frame prevalence varies considerably over time. Several peaks in the graph correspond to real-life events. We see, for example, a peak in the weapon development frame (Frame 4) around 1998, when the Indian government conducted the Pokhran 2 nuclear bomb tests.

Moreover, we see a peak in the prevalence of the accidents and radiation risks frame (Frame 6) around 2011, followed by a peak in the safety frame (Frame 1) shortly afterward. This is most probably related to the Fukushima disaster and debates about the safety of nuclear power it caused in the media. All in all, these findings confirm our conclusion that we identified a valid set of emphasis frames.

Third, we analyzed the tone of articles in the selection approach dataset. In Figure 3.3, we show overtime variation in tone for articles from each frame cluster. The graph indicates clear differences in tone across clusters. Articles focusing on the processing of nuclear waste and materials (Frame 5) or accidents and radiation risks (Frame 6), for example, are much more negative than articles that focus on the effects

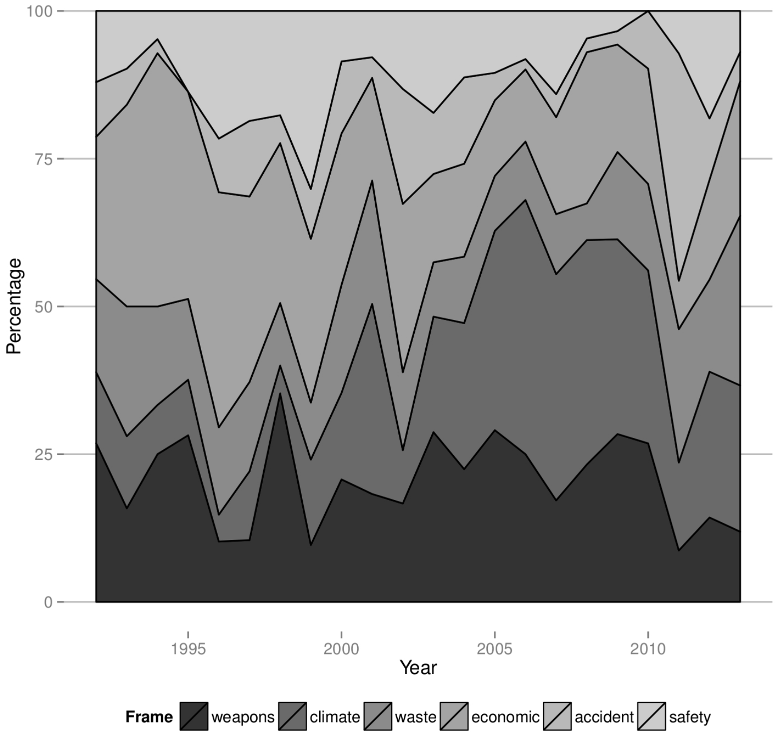

Figure 3.2: Stacked Area Plot of Frame Prevalence between 1990 and 2013.

of nuclear power on climate change (Frame 2) or economic aspects of nuclear power (Frame 3).

This is in line with the literature, where the former two frames are depicted as risk frames and the latter as opportunity frames [Gamson and Modigliani, 1989]. Furthermore, we also see within-cluster variation in tone. Articles in the weapon development cluster (Frame 4) as well as articles in the economic aspects cluster (Frame 3) become much more positive over time. Again, the patterns of variation correspond to actual events. In the aftermath of the Fukushima disaster, news coverage not

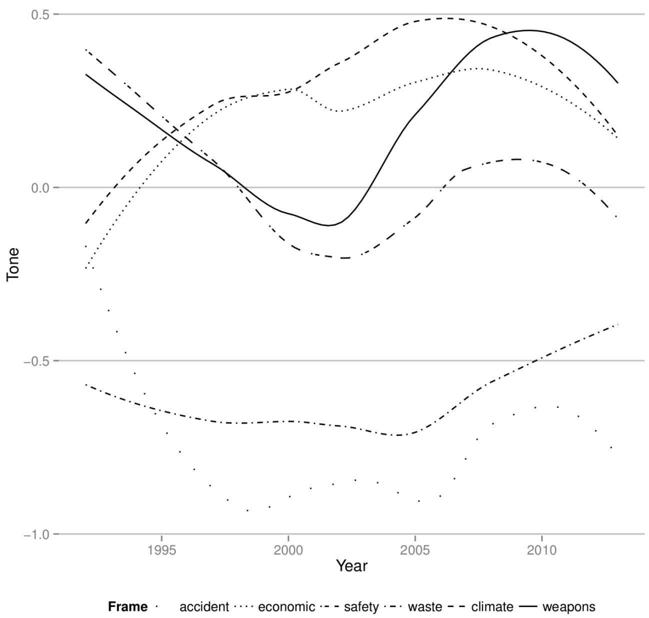

Figure 3.3: Tone of Articles from each Frame Cluster between 1990 and 2013.

only focused much more on safety issues (Frame 1) of nuclear plants, but articles focusing on safety issues also became more negative.

3.8 Discussion

In this chapter, we applied cluster- and sentiment analysis to identify and code news frames. Statistical frame analysis has several advantages over holistic approaches, all of which can guide future framing research.

First, it is more cost-efficient, because no manual content analysis is required. Second, it scales better to big datasets, which become increasingly available as the use of social media and the availability of digital news content increases. Third, it reduces risks of bias caused by human perceptions and interpretations [Matthes and Kohring, 2008].

Cluster analysis, in particular, has the advantage that it automatically classifies articles into groups and that it provides a model for doing so in future research. This is a more efficient and sophisticated way of coding documents for frames, as compared to manually creating coding rules, either on the basis of results from holistic analyses or based on key words from a factor analysis.

Although statistical techniques are widely used among communication scholars to identify news frames, they are criticized for not being able to do so in a conceptually valid manner [Carragee and Roefs, 2004, Hertog and McLeod, 2001]. For this reason, we explored a way of improving the cluster analysis of frames such that the resulting clusters more closely resemble emphasis frames. We found that when using all words of an article as features, clusters are often centered on individual actors and events instead of more abstract elements of the controversy. In addition, different elements are referred to in the same cluster, and different clusters overlap as regards the elements they refer to.

In contrast, when using only words from the title and lead as features and when removing all named entities from the feature space, clusters more accurately discriminate between distinct elements of the controversy. In addition, when using this selection of highly indicative features, more articles get accurately coded for frames. Generally, we conclude that the vast majority of articles are correctly clustered for the frame they contain when selecting features. In other words, most articles within a frame cluster actually contain the predicted frame.

The frames we identified by means of cluster analysis closely match frames that scholars found in earlier research, applying holistic methods (Table 3.1). From this, we conclude that our method is suited to identify frames. However, the frames we found are less detailed interpretations of the nuclear power controversy than those discussed in holistic studies.

First, frames from holistic studies often contain valence elements: They focus on either positive or negative aspects of the issue. The valence of frames is not properly revealed by our cluster analysis, as the cluster centers mostly describe topical aspects. However, when adding sentiment analysis, we can reproduce the valence of holistically identified frames in most cases.

Second, manually identified frames often express causal relations. These are not directly visible in our cluster centers and sentiment scores. We believe that, based on plain word features, a cluster analysis cannot reveal complex semantic and logical relationships like causality. It should be a challenge for future research to improve automatic frame clustering such that causality can be accounted for. In computational linguistics research, this problem has been addressed [e.g., Girju and Moldovan, 2002], but it is still difficult to automatically reveal the exact relation between two concepts in a sentence.

There are several limitations to this research. First, we only focused on three newspapers from two countries. Second, it is challenging to validate the found frames, as there is no ground truth about what is a frame and what is not a frame. Similarly, there is no true reference list of frames that are used in the nuclear power debate. Third, k-means clustering is a nondeterministic method and, therefore, results slightly vary each time the analysis is conducted. However, after repeated analyses, we observed comparable results in the sense of similar frames in the majority of the runs.

Finally, it might be misleading to argue that our approach is completely inductive, because we interpret the words in the cluster centers as frames. However, this interpretation is very straightforward. When applying the selection approach, for the vast majority of clusters, the words from the cluster centers clearly indicate one (topical) element of the nuclear power debate.

We believe that our approach to statistical frame analysis facilitates the use of mixed-methods designs (e.g., the combination of panel surveys and content analysis) in framing research (e.g., de Vreese, 2012), because it is very cost-efficient. Furthermore, it allows for increases in the scale of frame analysis. This allows scholars to reliably study developments in framing over long time frames and between different sources in an efficient manner [e.g., Vliegenthart and Roggeband, 2007].

This approach is useful for certain applications in particular, including studying the mapping of topical aspects of social and political issues, with an interest in long-term dynamics of how issues are presented in the news. If one is interested in getting a highly detailed and in-depth account of single events that span a limited amount of time, traditional (holistic) approaches might be a better choice. Furthermore, the generalization of this approach is limited to issues that receive a certain amount of coverage and which are sufficiently contested in news coverage.

Finally, findings of this study suggest implications for framing effects on public opinion. Since we identify a different set of frames when only looking at the title and lead of articles, this could have implications if people only read the headline or lead of news stories. Framing research has shown that exposure to different news frames can affect peoples' opinions about an issues and also their behavior [e.g., Nelson and Oxley, 1999, Van Spanje and De Vreese, 2014]. Furthermore, title and lead are considered the most important framing devices of a news story [Tankard, 2001]. Therefore, one can conclude that framing effects might be stronger among people, who only read title and lead of a news story.

3.9 Bibliography

Amy E Jasperson, Dhavan V Shah, Mark Watts, Ronald J Faber, and David P Fan. Framing and the public agenda: Media effects on the importance of the federal budget deficit. *Political Communication*, 15 (2):205–224, 1998.

Dhavan V Shah, Mark D Watts, David Domke, and David P Fan. News framing and cueing of issue regimes: Explaining Clinton's public approval in spite of scandal. *Public Opinion Quarterly*, 66(3):339–370, 2002.

Paul M Sniderman, Richard A Brody, and Phillip E Tetlock. *Reasoning and choice: Explorations in political psychology*. Cambridge University Press, Cambridge, UK, 1993.

Justin Grimmer and Brandon M Stewart. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21:267–297, 2013.

Giovanni Motta and Christian Baden. Evolutionary factor analysis of the dynamics of frames: Introducing a method for analyzing high-dimensional semantic data with time-changing structure. *Communication Methods and Measures*, 7(1):48–82, 2013.

Jörg Matthes and Matthias Kohring. The content analysis of media frames: Toward improving reliability and validity. *Journal of Communication*, 58(2):258–279, 2008.

Dennis Chong and James N Druckman. Framing theory. *Annual Review Political Science*, 10:103–126, 2007.

Claes H De Vreese. News framing: Theory and typology. *Information Design Journal and Document Design*, 13(1):51–62, 2005.

Thomas E Nelson, Rosalee A Clawson, and Zoe M Oxley. Media framing of a civil liberties conflict and its effect on tolerance. *American Political Science Review*, 91(03):567–583, 1997.

Karen Bickerstaff, Irene Lorenzoni, Nick F Pidgeon, Wouter Poortinga, and Peter Simmons. Reframing nuclear power in the UK energy debate: nuclear power, climate change mitigation and radioactive waste. *Public Understanding of Science*, 17(2):145–169, 2008.

William A Gamson and Andre Modigliani. Media discourse and public opinion on nuclear power: A constructionist approach. *American Journal of Sociology*, 95:1–37, 1989.

Matthew C Nisbet. Communicating climate change: Why frames matter for public engagement. *Environment: Science and Policy for Sustainable Development*, 51(2):12–23, 2009.

Christian Joppke. Social movements during cycles of issue attention: The decline of the anti-nuclear energy movements in West Germany and the USA. *British Journal of Sociology*, 42(1):43–60, 1991.

Marci R Culley, Emma Ogley-Oliver, Adam D Carton, and Jalika C Street. Media framing of proposed nuclear reactors: An analysis of print media. *Journal of Community & Applied Social Psychology*, 20 (6):497–512, 2010.

Nick F Pidgeon, Irene Lorenzoni, and Wouter Poortinga. Climate change or nuclear power—no thanks! a quantitative study of public perceptions and risk framing in Britain. *Global Environmental Change*, 18 (1):69–85, 2008.

Adam Simon and Michael Xenos. Media framing and effective public deliberation. *Political Communication*, 17(4):363–376, 2000.

Holli A Semetko and Patti M Valkenburg. Framing european politics: A content analysis of press and television news. *Journal of Communication*, 50(2):93–109, 2000.

Nel Ruigrok and Wouter Van Atteveldt. Global angling with a local angle: How US, British, and Dutch newspapers frame global and local terrorist attacks. *The Harvard International Journal of Press/Politics*, 12(1):68–90, 2007.

Iina Hellsten, James Dawson, and Loet Leydesdorff. Implicit media frames: Automated analysis of public debate on artificial sweeteners. *Public Understanding of Science*, 19(5):590–608, 2010.

Toni GLA Van Der Meer and Piet Verhoeven. Public framing of organizational crisis situations: Social media versus news media. *Public Relations Review*, 39(3):229–231, 2013.

M Mark Miller. Frame mapping and analysis of news coverage of contentious issues. *Social Science Computer Review*, 15(4):367–378, 1997.

Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, New York, NY, 2009.

Jennifer G Dy and Carla E Brodley. Feature selection for unsupervised learning. *The Journal of Machine Learning Research*, 5:845–889, 2004.

R Gnanadesikan, JR Kettenring, and SL Tsao. Weighting and selection of variables for cluster analysis. *Journal of Classification*, 12(1): 113–136, 1995.

Vasileios Hatzivassiloglou, Luis Gravano, and Ankineedu Maganti. An investigation of linguistic features and clustering algorithms for topical document clustering. In N Belkin, MK Leong, and P Ingwersen, editors, *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 224–231, New York, NY, 2000. ACM.

Xiaohua Hu, Xiaodan Zhang, Caimei Lu, Eun K Park, and Xiaohua Zhou. Exploiting Wikipedia as external knowledge for document clustering. In J Elder and FS Fogelman, editors, *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 389–396, New York, NY, 2009. ACM.

Julian Sedding and Dimitar Kazakov. Wordnet-based text document clustering. In V Pallotta and A Todirascu, editors, *proceedings of the 3rd workshop on robust methods in analysis of natural language data*, pages 104–113, Stroudsburg, PA, 2004. ACL.

Robert M Entman. Framing: Towards clarification of a fractured paradigm. *Journal of Communication*, 43:51–58, 1993.

Daniel Riff, Stephen Lacy, and Frederick Fico. *Analyzing media messages: Using quantitative content analysis in research.* Routledge, New York, NY, 2014.

Kevin M Carragee and Wim Roefs. The neglect of power in recent framing research. *Journal of Communication*, 54(2):214–233, 2004.

James K Hertog and Douglas M McLeod. A multiperspectival approach to framing analysis: A field guide. In SD Reese, OH Gandy, and AE Grant, editors, *Framing public life: Perspectives on media and our understanding of the social world*, pages 139–161. 2001.

Joseph N Cappella and Kathleen Hall Jamieson. *Spiral of cynicism: The press and the public good.* Oxford University Press, New York, NY, 1997.

Horst Poettker. News and its communicative quality: the inverted pyramid—when and why did it appear? *Journalism Studies*, 4(4):501–511, 2003.

Zhongdang Pan and Gerald M Kosicki. Framing analysis: An approach to news discourse. *Political Communication*, 10(1):55–75, 1993.

Somnath Banerjee, Krishnan Ramanathan, and Ajay Gupta. Clustering short texts using Wikipedia. In W Kraai and A De Vries, editors, *Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval*, pages 787–788, New York, NY, 2007. ACM.

Christos Bouras and Vassilis Tsogkas. A clustering technique for news articles using WordNet. *Knowledge-Based Systems*, 36:115–128, 2012.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In HT Ng and E Riloff, editors, *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 173–180, Strousburg, PA, 2003. ACL.

David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. Cambridge University Press, Cambridge, UK, 2008.

Andrew Y Ng. Feature selection, l 1 vs. l 2 regularization, and rotational invariance. In C Brodley, editor, *Proceedings of the twenty-first international conference on Machine learning*, page 78, New York, NY, 2004. ACM.

John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Applied Statistics*, 2:100–108, 1979.

David Sculley. Web-scale k-means clustering. In M Rappa and P Jones, editors, *Proceedings of the 19th international conference on World wide web*, pages 1177–1178, New York, NY, 2010. ACM.

David Arthur and David Vassilvitskii. k-means++: The advantages of careful seeding. In H Gabow, editor, *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035, Philadelphia, PA, 2007. SIAM.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, and Vincent Dubourg. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12: 2825–2830, 2011.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In N Calzolari and K Choukri, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 2200–2204, Valletta, Malta, 2010. European Language Resources Association (ELRA).

Marco Guerini, Lorenzo Gatti, and Marco Turchi. Sentiment analysis: How to derive prior polarities from SentiWordNet. In A Moschitti, editor, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP'13)*, pages 1259–1269, Stroudsburg, PA, 2013. ACL.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4):1191–1207, 2013.

Roxana Girju and Dan I Moldovan. Text mining for causal relations. In S Haller and G Simmons, editors, *Proceeding of the FLAIRS Conference*, pages 360–364, Palo Alto, CA, 2002. AAAI Press.

Rens Vliegenthart and Conny Roggeband. Framing immigration and integration relationships between press and parliament in the Netherlands. *International Communication Gazette*, 69(3):295–319, 2007.

Thomas E Nelson and Zoe M Oxley. Issue framing effects on belief importance and opinion. *The Journal of Politics*, 61(04):1040–1067, 1999.

Joost Van Spanje and Claes H De Vreese. Europhile media and Eurosceptic voting: effects of news media coverage on Eurosceptic voting in the 2009 European Parliamentary elections. *Political Communication*, 31(2):325–354, 2014.

James W Tankard. The empirical approach to the study of media framing. In SD Reese, OH Gandy, and AE Grant, editors, *Framing public life: Perspectives on media and our understanding of the social world*, pages 95–106. Taylor and Francis, Mahwah, NJ, 2001.

4

Generalization of Classifiers in Agenda Setting Research

Chapter 4: Generalization of Classifiers in Agenda Setting Research

This chapter has been published as:

Burscher, B., Vliegenthart, R., & De Vreese, C. H. (2015). Using Supervised Machine Learning to Code Policy Issues Can Classifiers Generalize across Contexts? *The ANNALS of the American Academy of Political and Social Sciences, 659*(1), 122-131.

The version presented here has been adapted to follow the overall standards and terminology included in the other chapters of the dissertation.

4.1 Abstract

Content analysis of political communication usually covers large amounts of material and makes the study of dynamics in issue salience a costly enterprise. In this chapter, we present a supervised machine learning approach for the automatic coding of policy issues, which we apply to news articles and parliamentary questions. Comparing computer-based annotations with human annotations shows that our method approaches the performance of human coders. Furthermore, we investigate the capability of an automatic coding tool, which is based on supervised machine learning, to generalize across contexts. We conclude by highlighting implications for methodological advances and empirical theory testing.

4.2 Introduction

Social scientists increasingly use supervised machine learning (SML) to automatically analyze media content [e.g., Hillard et al., 2008]. SML is a technique in which a computer learns from a set of human-coded training documents to automatically predict variables (e.g., the topic of a news article) in texts. In this chapter, we apply SML to the coding of policy issues, which is central to the study of agenda setting—a major paradigm in various social sciences [e.g., Baumgartner and Jones, 2010].

As agenda setting research is concerned with dynamics in issue salience among the media, politicians, and citizens [Rogers et al., 1993], it requires large-scale over-time content analysis (CA) across different types of political texts. An automatic coding tool should be able to correctly predict policy issues in different sorts of political texts, from various sources and time periods.

To investigate this, we conducted a series of validation experiments in which we employed SML to code policy issues in unknown datasets. Furthermore, we studied how a classifier's ability to predict the primary policy issue of a news article changes when using only words from its lead section in the training data. When it is necessary to code only a small fraction of each training document manually, SML becomes more cost efficient.

We found that SML is well suited to automatically code the primary policy issue of political texts. The ability of an SML model to generalize across contexts, however, is limited and depends on the characteristics of available training data. We conclude by discussing the strengths and limits of SML as compared to other approaches to automatic CA.

4.3 Computer-Aided Content Analysis

Scholars have followed different approaches to automatically code policy issues. In dictionary-based CA, previously defined character strings are used to code textual units into content categories [e.g., Schrodt et al., 1994].

This approach may compromise semantic validity, because manually compiled classification rules are at risk of being biased by the subjective conceptions and limited domain knowledge of the researcher(s). Furthermore, most people are not very good at determining how many different ways (e.g., senses, parts of speech) a word can be used when prompted with a specific category. This can lead to incomplete search strings and result in wrong predictions.

When applying unsupervised machine learning, issues are not defined a priori but are inductively extracted from the data by clustering documents that share the same words [e.g., Quinn et al., 2006]. This is an efficient approach, because it requires very little guidance. However, for each identified cluster, a person needs to manually infer its meaning afterward. This can be a difficult task, because the found clusters might not necessarily represent the desired content categories.

In the case of policy issues, some clusters might represent multiple issues, or might represent concepts other than policy issues (e.g., news coverage regarding a specific political actor or country). This poses a problem when one wants to code political texts according to a priori defined issues.

In SML, documents are automatically coded according to previously defined content categories by training a computer to replicate the coding decisions of humans [e.g., Hillard et al., 2008]. A premise for the application of SML is a set of documents that have been manually coded for the content categories of interest. This is called the training set. SML involves three steps:

First, documents from the training set are converted in such a way that they are accessible for computational analysis. Each document is represented as a vector of quantifiable textual elements (e.g., word counts), which are called features.

Second, feature vectors of all documents in the training set, together with the documents' content labels, are used to train a classifier to automatically code the content categories. In doing so, an SML algorithm statistically analyzes features of documents from each content category and generates a classifier to predict the content categories in future documents. Third, the classifier is used to code text documents outside the training set.

In SML, in contrast to dictionary-based CA, a computer automatically estimates a model that classifies texts according to content categories. This is likely to be more effective, because the rules used to identify the primary policy issue of a document are based on statistical analysis of human-coded training data. Compared to unsupervised machine learning, SML can apply a previously defined coding scheme. Being able to work with the same coding scheme in different studies facilitates the comparison as well as integration of findings across research contexts [John, 2006].

4.4 Research Questions

In this study, we applied SML to the coding of policy issues in political texts. The aim of the study was twofold. First, we investigated the generalizability of policy issue classifiers across research contexts. To do so, we conducted a series of validation experiments, in which we applied classifiers to unknown datasets. As Grimmer and Stewart [2013, p.268] argue, the "performance of any one classifier can vary substantially across context, so validation of a classifier's accuracy is essential to establish the reliability of supervised learning methods".

Information on the generalizability of classifiers helps scholars to decide on the suitability of a SML method. This is particularly relevant in comparative and longitudinal research, where documents from several outlets and time periods must be coded. In this chapter, we studied the generalizability of classifiers across two sorts of political texts (news articles and parliamentary questions [PQs]), across three different newspapers, and across a time frame of 15 years.

Second, we investigated how a classifier's ability to predict the primary policy issue of a news article changes when using only words from its lead section as features in the training set. Being able to reach similar classification accuracy with a training set in which only a small fraction of each article must be coded manually would significantly decrease the costs of applying SML to CA.

The chosen fraction must comply with two requirements. For an SML classifier, the fraction must be indicative of the primary policy issue. For human coders, the fraction must contain sufficient information to determine the primary policy issue when reading it. We chose to use the first 10 percent of words from each article, because in news articles facts are generally presented in descending order of importance [Poettker, 2003]. Hence, this fraction of an article should inform human readers about the main policy issue discussed, and it should include words that are highly indicative of that policy issue.

Third, we studied the relationship between the amount of training data used to build a classifier and its performance to predict the primary policy issue. As manually coded training data are expensive and labor-intensive to obtain, it is important to know how much training data one must possess to build a well-performing issue classifier.

4.5 Data

To investigate our research questions, we used data that consist of front-page news articles of the three most-read Dutch newspapers (*Volkskrant*, *NRC Handelsblad*, and *Telegraaf*) and Dutch PQs for the period between 1995 and 2011. All news articles were collected digitally via the Dutch Lexis-Nexis database. PQs were downloaded from the official website of the Dutch government.[1]

In the Netherlands, PQs are questions that members of parliament can direct to the government. Each question must be delivered in written form to the president of the House of Representatives, and must be orally answered by the addressed representative of the government during a weekly public session.

For each year, a stratified sample of news articles (13 percent) and written PQs (N = 500) were manually coded for the main policy issue discussed. For each article/PQ, coders could choose one out of twenty different policy issues. The coding scheme that we used was developed by the Policy Agendas Project [Baumgartner et al., 2006]. See Table 4.1 for an overview of all issue categories. The unit of coding was the distinct news article/PQ. Some PQs contained subquestions, which were grouped together. The resulting datasets consisted of 11,089 manually

[1] See http://www.officielebekendmakingen.nl

coded news articles and 4,759 manually coded PQs.

Manual coding was conducted by thirty trained coders. All coders were native Dutch speakers. To assess intercoder reliability, a random subset of articles (N=198) and PQs (N=200) was each coded by two coders. Krippendorff's alpha for issue category codings was equal to .69 for news articles and .60 for PQs. The coding was done as part of a large-scale research project about the influence of media coverage on parliamentarians.

4.6 Validation Experiments

First, we tested whether our classifiers could replicate the hand coding of documents from the original datasets of news articles (N=11,089) and PQs (N=4,759). In doing so, we used a stratified random sampling procedure to split each dataset into a training set (80 percent), on which we trained the classifier, and a test set (20 percent), on which we evaluated the classifier.

Second, to test a classifier's ability to correctly predict policy issues in another sort of political texts, we trained a classifier on a stratified random sample of four thousand news articles and tested the classifier on all PQs. Similarly, we trained a classifier on a stratified random sample of four thousand PQs and tested it on all news articles.

Third, we tested whether a classifier could correctly predict the main policy issue in documents from unknown sources. We split the news dataset into two subsets, one included a stratified random sample of four thousand articles from two of the three newspapers, and the other included all articles from the third newspaper. Then, we used the former as the training set and the latter as the test set. We repeated this exercise for all possible combinations of newspapers.

Finally, we tested whether a classifier could correctly predict the main policy issue in documents from unknown time frames. We split the news dataset in two subsets: a training set, which contained a stratified random sample of four thousand articles from 1995 to 2003, and a test set that contained all articles from 2004 to 2011. We also did this in the reverse.

4.7 SML Implementation

We compared two different SML implementations: one in which we used all words from each document in the training set as features, and one in which we used only the first 10 percent of words from each document in the training set as features. We compared the performance of both implementations when classifying news articles. When classifying PQs, we always used all words of the document.

For both news articles and PQs, we applied the following processing steps. First, we tokenized all documents and applied stemming to each token using the Frog natural language processing modules [Bosch et al., 2007].

Then, contingent on the implementation, we used either all tokens of the document, or selected the first 10 percent of its tokens. From this selection of tokens, we removed punctuation, single-letter words, and common Dutch stop words.

Then, we extracted all unique unigrams and bigrams from the remaining tokens and applied TF.IDF weighting [Russell and Norvig, 2002][2] to

[2] We also tried other bag-of-words implementations such as binary word presence and word counts. Findings showed that using TF.IDF weights was the most effective approach. When applying TF.IDF weighting, we normalized all data using the L2

them. Therefore, each unigram or bigram was assigned the number of times it occurs in a document (TF), weighted by the inversed frequency of documents in the entire collection containing the unigram/bigram (IDF). The idea behind TF.IDF weighting is to evaluate the power of a word to discriminate between documents.

In each classification task, we employed the Passive Aggressive learning algorithm[3], which is known to perform well in various text classification tasks [Crammer et al., 2006].[4]

Our main evaluation measure was the F1 score, which is equal to the harmonic mean of recall and precision. Recall is the fraction of relevant documents that are retrieved, and precision is the fraction of retrieved documents that are relevant. The F1 score is a standard evaluation measure for SML applications and provides a good indication of classification performance.

To assess the relationship between the size of the training set and classification performance, we plotted learning curves for the classification of news articles and PQs. We used a stratified cross-validation generator to split the whole dataset five times into training (80 percent) and test data (20 percent). Subsets of the training set with varying sizes were used to train the classifier, and F1 scores for each training subset size and the test set were computed. Afterward, the scores were averaged over all runs for each training subset size.

In all steps of the analysis, we used the scikit-learn machine learning library for the Python programming language [Pedregosa et al., 2011].

norm.

[3] We set the C-parameter to 100. This parameter trades off misclassification of training examples against simplicity of the decision surface.

[4] We tried different state-of-the-art algorithms for text classification as well as an ensemble of classifiers. However, the Passive Aggressive algorithm outperformed all tested alternatives.

4.8 Results

In Table 4.1, we report F1 measures of coding performance per policy issue for news articles and PQs. In these analyses we split each of the datasets into a training set (80 percent) and a test set (20 percent), and then used the former for training and the latter for testing.

When using all words of each document in the training set as features, average coding performance is equal to F1 = .71 for news articles and F1 = .69 for PQs. When using only the first 10 percent of words from each document in the training set as features, classification performance is equal to F1 = .68 for news articles. This is only marginally lower as compared to using all words of each article as the training data.

When looking at individual issue categories, we see that classification performance is higher for those issues that are more prevalent in the data. The correlation between F1 scores and the number of positive examples among the policy issues is equal to r = .40 for news articles and r = .50 for PQs.

Next, we turn to the validation experiments. To make results of all validation experiments comparable to one another, we set the training size in each validation experiment to four thousand documents. To make them comparable to the general analyses presented above, we report general classifier performance when using only four thousand training documents as a baseline measure in Table 4.2. Results of all validation experiments are based on the implementation in which we used all words of each document in the training set as features.[5]

First, we present results of experiments, in which we used newspaper articles as training data and PQs as test data (and then PQs for training

[5] Results are nearly identical when using only the first 10 percent of words from each document in the training set as features.

Table 4.1: F1 Scores for SML-Based Issue Coding in News Articles and PQs

Issue		*News Articles*			*PQs*
		All Words	*Lead Only*		*All Words*
Features	*N*	*F1*	*F1*	*N*	*F1*
Macroeconomics	413	.54	.63	172	.46
Civil Rights and Minority Issues	327	.34	.28	192	.53
Health	444	.70	.71	520	.81
Agriculture	114	.72	.76	159	.66
Labor and Employment	217	.43	.49	174	.58
Education	188	.79	.71	229	.78
Environment	152	.34	.44	237	.59
Energy	81	.35	.59	67	.66
Immigration and Integration	150	.50	.57	239	.78
Transportation	416	.58	.67	306	.81
Law and Crime	1198	.70	.69	685	.77
Social Welfare	115	.33	.34	214	.54
Community Development and Housing	113	.45	.44	136	.72
Banking, Finance and Commerce	622	.62	.67	188	.58
Defense	393	.59	.55	196	.71
Science, Technology and Communication	426	.64	.59	57	.53
International Affairs and Foreign Aid	1106	.70	.64	352	.65
Government Operations	1301	.71	.72	276	.48
Other Issue	3322	.84	.80	360	.51
Total	11089	.71	.68	4759	.69

and news articles for testing). Table 4.2 reports F1 measures of such tests. Measures show that classification accuracy significantly decreases when applying a classifier to a different sort of political text on which it is not trained. When predicting PQs with a classifier that is trained on news articles, F1 is equal to .50. When predicting news articles with a classifier that is trained on PQs, F1 is equal to .49.

Second, we turn to results of experiments in which we predicted the policy issues of news articles from unknown papers and time periods. F1 measures for predicting news articles from another newspaper range from .59 to .65, which is clearly lower compared to measures for predicting papers that were included in the training data.

Also, when predicting news articles from another time period, classification accuracy decreases. When training on the first half of the available time frame (1995–2003) and testing on the second half (2001–2011), F1 is equal to .59. When training on the second half of the available time frame and testing on the first half, F1 is equal to .63.

Finally, we turn to the relationship between the amount of training data and classification performance. The results are plotted in Figure 4.1. For news articles and PQs, classification performance increases as the amount of training data increases. This relationship, however, is not linear. After reaching a training size of around two thousand documents, coding performance increases only slowly when adding additional training documents. Moreover, the learning curve for PQs has a higher slope than the one for news articles. This indicates that PQs are easier to classify than news articles.

Table 4.2: F1 Scores for Validation Experiments

	Baseline (*N = 4000*)		*Other Text Sort*		*Other Newspaper* (*News Dataset*)			*Other Time Frame* (*News Dataset*)	
	News-> News	PQs-> PQs	News-> PQs	PQs-> News	VK/TEL-> NRC	NRC/TEL-> VK	VK/NRC-> TEL	’95/’03-> ’04/’11	’04/’11-> ’95/’03
F1	.67	.68	.50	.49	.59	.63	.65	.59	.63

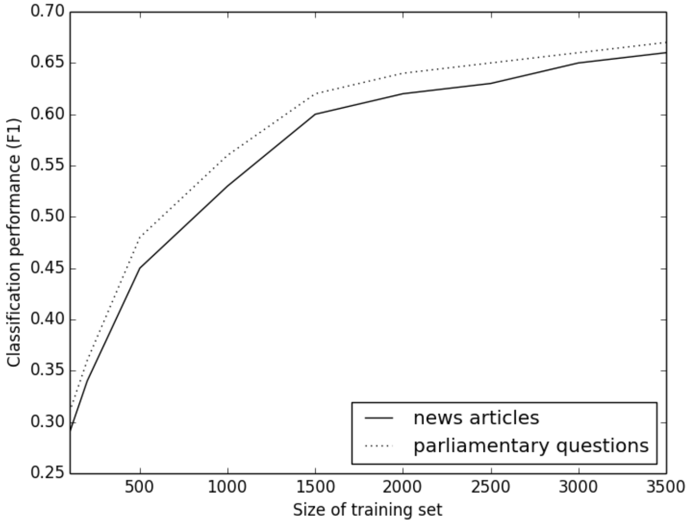

Figure 4.1: Learning Curves for the Classification of News Articles and PQs.

4.9 Discussion

Here we focused on two aspects of SML-based content analysis: the validation of SML classifiers across research contexts and the costs of training an SML classifier. To test the former, we applied policy issue classifiers to several unknown datasets. We found that classification accuracy decreases slightly when applying a classifier to an unknown newspaper, and strongly when applying it to articles from unknown time periods and content domains. From this, we conclude that training data must be representative of all outlets, time periods, and document types that one wants to study.

When this is infeasible, a dictionary-based approach might be preferred over an SML approach. An SML-based classification model is very specific to the word use within the training set. In a dictionary-based approach, in contrast, the classification model is more general. Therefore, it most likely performs more consistently across different contexts. Future research should focus on ways to improve the generalizability of policy issue classifiers by selecting less context-dependent features (e.g., names of persons and places).

To investigate the costs of training a policy issue classifier, we plotted the learning curves for both news articles and PQs. Based on the curves, we conclude that one does not need several thousand training documents to train a policy issue classifier. Actually, adding more hand-coded documents to the training set increases average coding performance only slowly after reaching a threshold of around two thousand training documents. Instead, it would be more effective to selectively sample positive examples for underrepresented categories. Several strategies for this are discussed in the literature [Hillard et al., 2008, Tong and Koller, 2002].

Furthermore, we found that whether one uses all words of a news article or only words from its leading paragraph when presenting it in the training set has little effect on classification performance. This implies that, when creating training data, it might be sufficient to code only the leading paragraphs of each article. This makes SML-based coding of policy issues more cost-efficient, and facilitates the coding of more representative samples from several sources and time periods, which most likely will increase the robustness and generalizability of a policy issue classifier. This way, SML becomes more attractive compared to other approaches to automatic CA, which require no manually coded training data.

Finally, we are aware that the quality of our training data is not optimal. Disagreement between coders likely results from a combination of unsystematic coding errors and systematically different interpretation of policy issues across coders. The most relevant question is how this might influence our findings and conclusions. We expect classification performance to decrease as a result of inconsistencies in the training data. If texts with similar features are associated with different policy issues, it becomes more difficult for the SML algorithm to estimate a model that can clearly differentiate between content categories. Although classification performance is most likely influenced by the quality of the training data, we believe our conclusion to be largely unaffected.

4.10 Bibliography

Dustin Hillard, Stephen Purpura, and John Wilkerson. Computer-assisted topic classification for mixed-methods social science research. *Journal of Information Technology & Politics*, 4(4):31–46, 2008.

Frank R Baumgartner and Bryan D Jones. *Agendas and instability in American politics*. University of Chicago Press, Chicago, IL, 2010.

Everett M Rogers, James W Dearing, and Dorine Bregman. The anatomy of agenda-setting research. *Journal of Communication*, 43(2):68–84, 1993.

Philip A Schrodt, Shannon G Davis, and Judith L Weddle. Political science: KEDS—a program for the machine coding of event data. *Social Science Computer Review*, 12(4):561–587, 1994.

Kevin M Quinn, Burt L Monroe, Michael Colaresi, Michael H Crespin, and Dragomir R Radev. An automated method of topic-coding legislative speech over time with application to the 105th-108th US Senate. In *Proceedings of the Midwest Political Science Association Meeting*, pages 1–61, Austin, TX, 2006. MPSA.

Peter John. The policy agendas project: a review. *Journal of European Public Policy*, 13(7):975–986, 2006.

Justin Grimmer and Brandon M Stewart. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21:267–297, 2013.

Horst Poettker. News and its communicative quality: the inverted pyramid—when and why did it appear? *Journalism Studies*, 4(4):501–511, 2003.

Frank R Baumgartner, Christoffer Green-Pedersen, and Bryan D Jones. Comparative studies of policy agendas. *Journal of European Public Policy*, 13(7):959–974, 2006.

Antal van den Bosch, Bertjan Busser, Sander Canisius, and Walter Daelemans. An efficient memory-based morphosyntactic tagger and parser for dutch. *LOT Occasional Series*, 7:191–206, 2007.

Stuart Russell and Peter Norvig. *Artificial Intelligence: A modern approach*. Prentice Hall, Upper Saddle River, NJ, 2002.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *The Journal of Machine Learning Research*, 7:551–585, 2006.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, and Vincent Dubourg. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12: 2825–2830, 2011.

Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2:45–66, 2002.

5

Automatic Dictionary Expansion in Content Analysis

Chapter 5: Automatic Dictionary Expansion in Content Analysis

5.1 Abstract

Dictionary-based content analysis is the most popular approach to study message characteristics in political communication research. Constructing a dictionary with search terms for several content categories can be a difficult and laborious task. Therefore, in this chapter, we introduce a method to automatically expand coding dictionaries with relevant search terms. In doing so, we employ word co-occurrence statistics that are based on word vectors from a neural network language model. We conduct several tests in which we use this method to automatically expand dictionaries for coding policy issues. We validate our method by applying automatically constructed dictionaries to different human-coded test sets. Results show that we can significantly increase the performance of a coding dictionary by automatically adding search terms. We discuss theoretical and methodological implications of our research.

5.2 Introduction

In political communication research, automatic content analysis is gaining popularity as more and more news coverage, user-created content and parliamentary records become digitally available [Günther and Quandt, 2015]. Dictionary-based coding is the most common form of automatic content analysis. In dictionary-based coding, previously defined character strings are used to code textual units into content categories [e.g., Schrodt et al., 1994]. In doing so, a scholar creates a dictionary with search terms for each content category and then a computer program counts the occurrence of these search terms in the input texts.

Although statistical approaches like supervised machine learning have gained increased popularity in the past years [Grimmer and Stewart, 2013], coding dictionaries are still widely used among scholars. This is because dictionaries perform well in various coding tasks [e.g., Young and Soroka, 2012] and, at the same time, have a high ease of use.

However, creating a good coding dictionary can be a laborious and difficult process; especially when the number of content categories is large [Hillard et al., 2008]. The biggest challenge is that most scholars can not think of all relevant words that indicate a content category, and they can not think of all ways such words can be used in language. Consequently, relevant search terms are missing in the coding dictionary, and not all documents can be coded correctly.

In this chapter we present a method to automatically expand coding dictionaries with additional search terms. The method aims at facilitating the creation of coding dictionaries, such that it becomes less laborious and the performance of automatic content analysis increases. Given a basic set of initial search terms (e.g., environment, pollution, climate) for a content category (e.g., environmental news) one can use the proposed method to expand such initial search terms with semantically related words and phrases (e.g., global warming, waste, forest). This increases the performance of a coding dictionary in content analysis.

In natural language processing (NLP) research, various approaches have been used to retrieve semantically related words [See Senellart and Blondel, 2008, for an overview]. Generally, the most successful approach is to infer semantic closeness between two words from their co-occurrence in natural language [e.g., Weerkamp et al., 2012]. This approach is based on the distributional hypothesis [Firth, 1957], which suggests that the more two words occur together (e.g., appear in the same document), the more semantically similar they are.

In this chapter, we follow such a *distributional approach*. More specifically, we make use of a neural network language model [Mikolov et al., 2013a] to compute similarity statistics between more than a million English words and phrases. The model is trained on a large text corpus of news articles. Given a basic coding dictionary, which contains just a few search terms per category, the model can be used to automatically expand the dictionary with relevant search terms.

We present a case study, in which we use this method to create dictionaries for coding 13 policy issues in political documents. In doing so, we compare several ways of identifying new search terms based on the similarity measures created by the neural network language model. In order to validate automatically created dictionaries, we use them to code policy issues in documents from two human-coded testsets: a set of New York Times news articles and a set of UK parliamentary questions.

Results show that we are able to automatically expand coding dictionaries with words that are relevant indicators of the policy issues. We conclude that the introduced method is very helpful in constructing dictionaries for automatic content analysis. It can help to improve the validity of content analysis and ease the process of constructing coding dictionaries. We discuss the theoretical and methodological implications of our study for future communication research.

5.3 Dictionary-Based Content Analysis

For several decades, dictionary-based content analysis has been applied to the coding of a broad range of concepts in political communication research. It has been used to code policy issues [e.g., Albaugh et al., 2013], news frames [e.g., Roggeband and Vliegenthart, 2007], tone [e.g., Soroka et al., 2015], emotions [e.g., Cho et al., 2003], and references to political actors. While some scholars have created small-scale search queries for an individual study [Roggeband and Vliegenthart, 2007], others developed sophisticated coding dictionaries as part of larger research projects.

A well-known example of such a large-scale project is the KEDS TABARI project [Schrodt et al., 1994], where scholars created a set of computational rules for coding texts into various event categories. Researchers use the resulting system to analyze changing attention to events in (international) media coverage.

A more recent example is the Policy Agendas Project [Baumgartner et al., 2006], in which scholars have developed a system to classify political texts for policy issues. The taxonomy consists of 20 major topics and more than 200 subtopics. In several countries, scholars have used this taxonomy to manually code policy issues in thousands of documents from various domains (e.g., media content, legislation, judicial decisions and public opinion). This has facilitated innovative research on the interplay between the agendas of publics, media and policymakers [Vliegenthart and Walgrave, 2008, Walgrave et al., 2007].

Several studies investigated methods to automatically code political texts according to the issue taxonomy of the Policy Agendas Project [Hillard et al., 2008, Burscher et al., 2015]. Albaugh et al. [2013], for example, created the Lexicoder Topic Dictionary - a dictionary with search terms for each of the major topics of the taxonomy. The dictionary has gone through several iteration over a number years.

In this chapter, we investigate how coding dictionaries can be created semi-automatically. We create dictionaries for a selection of policy issues, which are based on the taxonomy of the Policy Agendas Project. Among others, we compare our semi-automatically created dictionaries to the original Lexicoder topic dictionary. But first, we want to discuss some characteristics of dictionary-based content analysis.

The goal of creating a coding dictionary is to find a set of words that indicate a specific content category (e.g., economic news) and that can be used as search terms for identifying documents from that category. Ideally, this set of words is as discriminative as possible between the category (economic news) and all remaining categories (e.g., immigration news, crime news, environmental news, ...).

Coding dictionaries have various pros and cons. First, no hand-coded training data is necessary to create a coding dictionary. This makes dictionaries a good choice when it is not feasible or impossible to hand-code documents as part of a supervised machine learning approach [Hillard et al., 2008]. Second, coding dictionaries are transparent. The process of how texts are associated with a content category can be judged by the face-validity of the search terms in the dictionary. Third, like supervised machine learning approaches, dictionary-based coding is reliable, because it is based on a deterministic model.[1] As long as the search queries in the dictionary do not change, each coding decision is replicable.

However, creating a dictionary is a laborious and demanding process. It requires a careful selection of search terms, which can be challenging for several reasons:

First, for an individual it is difficult to come up with the "complete" set of words that indicate a content category. Each person's domain knowledge is limited by personal experience and everybody has his or her own associations with a particular concept. Second, for most people it is very difficult to think about all the different ways (e.g., senses, parts of speech) a word can be used when prompted with a specific category. Third, if a word in a document is misspelled or an abbreviation is used, the search-term would not match the word. This can be particularly problematic when analyzing social media, because people make frequent use of abbreviations and spoken language.

The performance of a coding dictionary can be evaluated in various ways. *Precision* and *recall* are two popular metrics, both of which are standard information retrieval measures [Davis and Goadrich, 2006]. Precision is the percentage of documents that have been labeled as belonging to a content category, which actually belong to that category. Recall is the percentage of documents belonging to a content category, which have been labeled as belonging to that category. In terms of these metrics, a coding dictionary should identify as many relevant documents as possible (high recall), while resulting in as few false positives as possible (high precision).

[1] This is not the case with unsupervised machine learning models [e.g., Quinn et al., 2006], where each fit of a model differentiates form the previous one.

This study presents a method to facilitate the process of creating a coding dictionary. By automatically expanding a dictionary with additional search terms it helps dealing with the above listed challenges. This significantly increases document recall and overall coding performance.

5.4 Automatic Dictionary Expansion

Different strategies can be applied to automatically expand coding dictionaries. Basically, the task is to retrieve sets of semantically related words. For example: Given the content category *economic news* and the search term *economy*, we want to find additional search terms that refer to concepts from that content category, like *growth*, *inflation* and *debt*. But we also want to find synonyms of such words (e.g., *liability* as a synonym for *debt*), and different (mis)spellings of the words (e.g., *ecnomy* instead of *economy*).

The task of identifying semantically related words is a well-studied subject in the fields of natural language processing (NLP) and information retrieval (IR). It has mainly been addressed in the context of *automatic thesaurus generation* [Grefenstette, 2012] and *automatic query expansion* [Carpineto and Romano, 2012].

Thesaurus generation is the task of creating lists of words, which are grouped by semantic similarity [Zohar et al., 2013]. A thesaurus can be applied to various problems in NLP and IR. A common application is automatic query expansion [e.g., Bai et al., 2005]. The aim of query expansion is to automatically expand a search query with semantically related words - in order to improve the retrieving of relevant documents.

In this chapter, we basically do not study a retrieval problem, but a classification problem. The goal of dictionary-based coding is to classify documents into previously defined content categories. In the communication literature, using dictionaries of search terms is a common approach for classifying documents [Grimmer and Stewart, 2013]. Our goal, thus, is not to identify semantically related words for expanding search queries (or building a thesaurus), but for expanding coding dictionaries. To solve this problem, we build on query expansion studies and research in automatic thesaurus generation.

How can dictionary expansion improve content analysis? By adding additional search terms to a dictionary, we increase the chance of correctly coding a relevant document that does not contain the original search terms. As a result, recall of relevant documents and, therefore, the coding performance of the dictionary increases.

However, by adding additional search terms, one also runs the risk that precision decreases. This is the case when the newly added search terms are not (exclusively) indicators of the content category they are intended for, and called query drift [Mitra et al., 1998]. The word *defense*, for example, might refer to the ministry of defense or a football game and, therefore, can lead to false positives. Therefore, the challenge in dictionary expansion is to find a good trade-off between increasing recall by adding additional search terms, but not loosing too much precision.

A wide range of techniques have been applied to find semantically related words (see Senellart and Blondel [2008] for an overview). The most basic strategy is to make use of *thesauri* like the *WordNet* database [Miller, 1995]. WordNet is a digital thesaurus and can be used to find synonyms of words from various content categories.[2]

However, research has shown that adding synonyms from a thesaurus

[2] Varelas et al. [2005] provide an overview of how semantic similarity between two words can be calculated based on WordNet features.

only marginally improves the performance of a query [Navigli and Velardi, 2003]. One explanation is that thesauri are too general to provide useful synonyms for analyzing a domain-specific collection of documents [Zohar et al., 2013].

Another class of methods is called *distributional methods* [Chen and Lynch, 1992]. Such methods infer the semantic similarity of two words from their co-occurrence in a document collection. Therefore, distributional methods are able to capture the domain-specific meaning of a word. The theoretical foundation underpinning this approach is the distributional hypothesis [Sahlgren, 2008], which states that words are similar if they are used in the same context. Various distributional methods have been applied in thesaurus generation [Church and Hanks, 1990] and search query expansion [Bast et al., 2007, Hu et al., 2006].

In this chapter, we follow a distributional approach to find semantically related words. Based on the co-occurrence of words in a large collection of English news articles, we identify search terms that can be used to expand coding dictionaries. This leads to better coding decisions.

When arguing that words are similar if they are used in the same context, the term context refers to the local, textual environment in which a word is used (e.g., a sentence or document). Different definitions of context and different measures of similarity between contexts have been applied in previous research [Senellart and Blondel, 2008]. One definition of similarity with respect to a context is that two words are semantically related if they appear in the same document.

However, computing co-occurrence of words in the whole document has the disadvantage that a word's position in the document is not considered. Two words that co-occur within the same sentence are more similar than two words that occur distantly within the same document.

A solution to this problem is using a finite word window (e.g., the five direct neighbours of a word) as a context [Xu and Croft, 1996].

When modeling word meaning in terms of word contexts, one generally represents each word as a vector. Each element of such a word vector then is the probability that the word co-occurs with a specific other word in the same context. This leads to word vectors where the dimensionality of the vector is equal to the amount of unique words in the vocabulary of the document collection.

Generally, the performance of distributional methods increases with the size of the corpus. This is because a model has more material from which it can learn how words and phrases are distributed in natural language. But by increasing the corpus size also the size of the vocabulary and, therefore, the dimensionality of the vector space increases [Curran and Moens, 2002].

Because most words do not appear in most contexts, the word vectors are very sparse, and mostly contain zeros. Modeling such high-dimensional and sparse data is a computationally complex problem. It can be solved by representing words in a lower-dimensional space.

Basically, there are two ways this can be achieved. One solution is dimensional reduction - projecting the high-dimensional word vectors into a lower-dimensional space. Popular techniques to reduce the dimensionality of word vectors are singular value decomposition [Deerwester et al., 1990, Yang and Powers, 2008] or random indexing [Henriksson et al., 2014]. The reduced word vectors are called *word embeddings*.

Another way of representing words as vectors in a low-dimensional space is the use of *neural network language models* [Bengio et al., 2003]. Such models provide a way to directly learn low-dimensional word embeddings from a collection of text documents.

In this chapter, we employ a neural network language model to create word embeddings. In the next section we explain the workings of neural network language models and how they can be used to create word embeddings. But first, we want to elaborate on the actual advantages of low-dimensional word embeddings.

Working with low-dimensional word embeddings has two advantages: it reduces computational complexity, and improves the semantic quality of the word vectors [Henriksson et al., 2014]. In the reduced vector space, "terms that do not necessarily co-occur directly in the same contexts (...) will nevertheless be clustered about the same subspace, as long as they appear in similar contexts, i.e. have neighbors in common" [Henriksson et al., 2014, p.4].

Consider the following example: if using vectors from a term co-occurrence matrix, the words *astronaut* and *cosmonaut* would have low similarity, because the words are unlikely to appear in the same document or word-window. But they are most likely used in combination with the same neighboring words (e.g., *space*, *rocket* or *moon*).

In this study we enrich coding dictionaries with additional search terms. In doing so, we follow a distributional approach to identify semantically related words. In our approach, we represent words as word embeddings, which have been learned by means of a neural network language model [Mikolov et al., 2013a]. This language model has been trained on a collection of English news articles.

We expect that expanding a coding dictionary with semantically related words increases the performance of the dictionary and, therefore, improves dictionary-based content analysis. This leads to the following research question: *To what extent can we increase the performance of a coding dictionary by automatically adding additional search terms?*

In the next section, we will take a closer look at the specifics of our approach.

5.5 Word2Vec Word Embeddings for Dictionary Expansion

In the previous section, we introduced the concept of word embeddings. In this chapter we applied a neural network language model (NNLM) [Bengio et al., 2003] to create such word embeddings from a collection of news articles. Neural network language models can directly learn low-dimensional word embeddings from a text corpus using a word's neighbours within a sentence as context. In the resulting vector space, word embeddings of semantically related words are close to each other. Consequently, the closeness of two word embeddings can be used as a measure of semantic similarity, and semantically related words can be identified based on this measure [Mikolov et al., 2013b].

Generally, language models are probability distributions over word sequences in natural language [Lavrenko and Croft, 2001]. Neural network language models are a particular sort of language models, which represent word sequences in an artificial neural network [Bengio et al., 2003].

Given some input data, artificial neural networks reveal hidden patterns in the data [Yegnanarayana, 2009]. Based on such hidden patterns, they create higher-level representations of the data consisting of more abstract concepts. Such representations are called hidden layers and can be used as low-dimensional word embeddings.

We applied a recently developed neural network language model to compute word embeddings for the vocabulary of a corpus of news articles. This model is called the *skip-gram negative sampling model* (SGNS) and originates from the Word2Vec tool [Mikolov et al., 2013a]. Word2Vec is an open-source tool for computing and analyzing word vectors. It has been implemented in various programming language. We used the Gensim library [Řehůřek and Sojka, 2010], which is a Python implementation of Word2Vec.

Word embeddings from the SGNS model are very good in capturing semantic relationships between words in large-scale text corpora. When giving enough training data, one can even do arithmetical operations on the embeddings and reveal analogies like in the following example [Mikolov et al., 2013b]:

$$Qu\vec{e}en - Wo\vec{m}an + M\vec{a}n = Ki\vec{n}g$$

In previous research, Word2Vec-based word embeddings have been used to produce state-of-the-art results in several NLP tasks like synonym extraction [Wolf et al., 2014], sentiment analysis [Tang et al., 2014], part of speech recognition [Santos and Zadrozny, 2014] and query rewriting [Grbovic et al., 2015].

How does the SGNS model work? The SGNS model takes a document collection as input and learns a word embedding for each unique word (or phrase) in the vocabulary. A word embedding is a distributed representation of a word in a N-dimensional space. Each element of the embedding captures semantic characteristics of that word, based on the distributional properties of the document collection. The cosine similarity between two word embeddings can be used as a way to measure the semantic similarity of the two words.

To compute the word embeddings, SGNS trains a neural network model with one hidden layer. The goal of the model is to predict the neighbouring words of a word w_i, given w_i [Mikolov et al., 2013a, Goldberg and Levy, 2014]. In the training process, the model iterates over raw text from the training documents and uses pairs of individual words and their neighbours as training instances. The model consists of an input layer, a hidden layer and an output layer. At each time T, the input is a single word w_i and the output is the words in w_i's context $c_{w_i} = w_{i_1}, ..., w_{i_M}$ defined by a word window of size M.

At each iteration in the training process, the weight matrix, which is the projection from the input layer to the hidden layer is updated. Each row in this weight matrix is the vector representation $\vec{v_{w_i}}$ of a word w_i. This way the model learns a word embedding for each unique word in the vocabulary.

To get word embeddings with the desired semantic properties, the SGNS model should be trained on a large dataset, which to some extent represents the population of texts one is interested in analysing. We used a pre-trained Word2Vec model, which has been trained on a huge corpus of news articles. The model is publicly available and can be downloaded at the official website of the Word2Vec tool.[3]

The training data consists of an internal Google corpus of English news articles with a length of more than 100 billion words. The resulting model contains 300-dimensional vectors for about 3 million words and phrases. This includes unigrams, bigrams and trigrams. The phrases were obtained using a simple data-driven approach described in Mikolov et al. [2013c].

[3] https://code.google.com/p/word2vec/

5.6 Dictionary Expansion Approaches and Validation

We conducted a series of tests in which we used word embeddings from the SGNS model to automatically expand a dictionary for coding 13 policy issues: (1) *economy*, (2) *civil rights*, (3) *health*, (4) *education*, (5) *environment*, (6) *energy*, (7) *transportation*, (8) *crime*, (9) *welfare*, (10) *defense*, (11) *housing*, (12) *agriculture* and (13) *science & technology*. These issues are based on the taxonomy of the Policy Agendas Project. [4] Please note that the original taxonomy of the Policy Agendas Project consists of even more topics. For practical reasons we brought this number down to 13 by merging and excluding some of the original categories.[5]

In order to expand a dictionary, we first need to create a basic dictionary with a set of initial search terms for each issue category. Then, we can use the SGNS model to identify semantically related words for each initial search term in this basic dictionary.

We created such a basic dictionary B, which contains a set of three initial search terms $I_b = i_{b_1}, i_{b_2}, i_{b_3}$ for each policy issue $b \in B$. In each case, we used the category name (e.g., environment) as one of the three initial search terms. The other two words were based on the issue descriptions in the Policy Agendas code book.[6] See Table 5.1 for an overview of the initial search terms.

[4] See http://www.policyagendas.org for more information about the policy issue taxonomy and the Policy Agendas project.

[5] First, we merged all economic issues (*macroeconomics*, *labor & employment*, *banking*, *finance & commerce* and *foreign trade*. Second, we merged the issues *agriculture* and *land & water management*. Third, we removed the issue *government operations*, because it strongly overlapped with several other issues.

[6] See http://www.policyagendas.org

Table 5.1: Initial Search Terms Basic Dictionary

Issue	**Initial Search Terms**
Economy	economy, employment, business
Civil Rights	rights, minorities, discrimination
Health	health, disease, medical
Education	education, student, learning
Environment	environment, pollution, climate
Energy	energy, fuel, reactor
Transportation	transportation, airport, road
Crime	crime, prison, police
Welfare & housing	welfare, social services, poverty
Defense	military, war, weapons
Housing	housing, condo, tenant
Agriculture	agriculture, farming, crop
Science & technology	science, technology, internet

We realize that three terms per policy issue is an arbitrary number, but the purpose of these tests is to show that we can automatically expand coding dictionaries with as little human input as possible.[7]

In each test we estimated the effect of the expansion of the basic dictionary on its coding performance. For this, we used the basic as well as the expanded dictionaries to code two reference datasets of political documents. All documents in the reference datasets have been annotated manually for the same policy issues. This way we can compare human codings to dictionary-based automatic codings.

We compared two baselines with two experimental approaches. In the first baseline we used the basic dictionary, which contained three search terms per policy issue. In the second baseline we used the Lexicoder topic dictionary.[8]

[7] Results are similar when using four or five initial search terms per category.

[8] Because the Lexicoder topic dictionary contained search terms for each of the original topics of the Agendas Project's taxonomy, we merged the search terms for

In the two experimental approaches, we compared two different ways of using the word embeddings from the SGNS language model to expand the basic dictionary with new search terms: a) *the single-term similarity approach* and b) the *two-step similarity approach*. We continue with a detailed account of these approaches.

Single-term similarity

In the *single-term similarity* approach, we expanded the initial set of search terms I_b for each issue $b \in B$ in the basic dictionary as follows: we retrieved a set of N most similar words S_{i_b} for each word $i_b \in I_b$ and added each newly retrieved word $s_{i_b} \in S_{i_b}$ to the initial set of search terms I_b for that category. We run several variations of this approach with different amounts of most similar words.

We used the cosine distance between two word vectors as a measure of similarity. If we would retrieve the 10 most similar words for the term *economy*, we would compute the cosine similarity between the word embedding of *economy* and each other word embedding in the model. Then, we would choose the top 10 words from the corpus whose word embeddings are closest to the embedding of the word *economy*.

Two-step similarity

In the "two-step similarity" approach, we had two iterations of the procedure described in the single-term similarity approach. First, we retrieved a set of N most similar words S_{i_b} for each word $i_b \in I_b$. Then, in a second iteration, we retrieved a set of O most similar words $T_{S_{i_b}}$ for every word $s_{i_b} \in S_{i_b}$. Finally, we added all newly retrieved words from

categories that fall into the same category in our 13-issue taxonomy.

the two iterations to the initial set of search terms I_b for that category. We run several variations of the approach with different numbers for N and O.

The rationale behind this approach is to retrieve a more diverse set of search terms. In order to expand our initial set of search terms such that it covers all aspects of a policy issue, we must not only retrieve words that are similar to the initial search terms. Instead, we must retrieve a more diverse set of words that expands to relevant subtopics. We expect to achieve this by performing two iterations of expansion. However, the approach can also lead to increased query drift and result in completely unrelated terms. This can have a negative effect on precision.

Validation

In order to test our baselines and the experimental approaches, we used each of the created dictionaries for coding two reference datasets. Both datasets have also been coded manually for policy issues from the Policy Agendas taxonomy.

The first dataset contains abstracts of English language news articles. It is a systematic random sample (N = 49,201) of the New York Times Index from 1946 to 2008. The sample includes the first entry on every odd-numbered page of the index. Each entry was coded by Policy Agendas major topics. The unit of analysis is an entry and each entry was assigned one and only one content code. Inter-coder agreement exceeded the level of 90 %. For more details on the coding procedure and the New York Times index, we refer to the website of the Policy Agendas Project [9], where we downloaded the dataset.

[9] http://www.policyagendas.org

Second, we used a set of British Parliamentary Questions (PQs, N = 9,062). PQs are a parliamentary convention where the Prime Minister answers questions in the House of Commons from members of parliament. PQs were blind-coded by two researchers; assigning one Policy Agendas major topic code to each question. Each entry includes the complete text of the question. Again, the unit of analysis was an entry. Inter-coder agreement exceeded 80%. For additional details on the dataset and coding procedure, we refer to [Bevan and John, 2015]. The dataset is publicly available. We downloaded it from the website of the UK Policy Agendas Project.[10]

Because both datasets were coded for each of the original major policy issues of the Policy Agendas Project taxonomy (and additional categories in some cases), we recoded all documents such that the topic codes match our 13-issue taxonomy.

Document Coding

So far we have discussed our baseline and experimental coding dictionaries. In this section, we describe the way we used such dictionaries to actually code documents in the datasets. For this, we employed the elastic-search software.[11] Elastic-search is an open-source search engine with interfaces for several programming languages. It can be used to code documents based on search terms and rank the resulting matches by relevance.

Given a document collection D and a set of policy issues T, we coded each document $d \in D$ for all policy issues. Please note that we excluded all documents from the two datasets (news articles and PQs),

[10] http://www.policyagendasuk.wordpress.com

[11] http://www.elastic.co

which were not assigned a specific topic by the human coders. Per policy issue $t \in T$, we followed the same 3-step procedure:

(1) We took the set of search queries Q_t for issue t from the dictionary. As the vocabulary of our SGNS model includes unigrams, bigrams and trigrams, each query contained up to three words.

(2) We performed a separate collection-wide 'OR' search for each individual search query $q_t \in Q_t$. For each document $d \in D$, these searches resulted in a vector $\vec{v_{dt}}$ containing a relevance score r_{dq_t} for each query $q_t \in Q_t$,

$$\vec{v_{dt}} = (r_{dq_{t_1}}, r_{dq_{t_2}}, ..., r_{dq_{t_M}})$$

If a document contained none of the words in the search query, the relevance score was equal to zero. If the document contained one or more of the words in the search query, the relevance score was equal to the return value of the Lucene practical scoring function (which is explained below).

(3) We computed for each document $d \in D$ an issue-score u_{dt}, which is equal to the sum of the relevance scores in $\vec{v_{dt}}$,

$$u_{dt} = \sum_{i=1}^{N} \vec{v_{dt_i}}$$

The issue-score indicates the likelihood that document d covers issue t.

Individual relevance scores are based on the Lucene practical scoring function: The relevance score of document d for query q is equal to

$$r_{dq} = coord(d, q) \cdot queryNorm(q) \cdot \sum_{t\ in\ q} tf(t) \cdot idf(t)^2, where$$

$tf(t)$ is the term frequency of term t in document d and $idf(t)$ is the inverse document frequency of term t in the collection of documents D.

Furthermore, $queryNorm(q)$ is a query normalization factor, which is equal to

$$\sum_{t\ in\ q} idf(t)^2, and$$

$coord(d, q)$ is a coordination factor, which is equal to the number of words that occur both in q and d. After applying this procedure, each document has been assigned one policy issue - the one with the highest relevance score.

5.7 Results

To answer our research questions, we conducted a series of tests in which we automatically expanded dictionaries for coding policy issues. We compared two baselines and two experimental approaches. In doing so, we coded two reference datasets using each of the baseline- and experimental dictionaries. The datasets were set of news articles and a set of parliamentary questions. We compared dictionary-based codings with manual codings, which we consider the gold-standard.

We report coding performance in terms of precision, recall and the F1-score. We already introduced precision and recall. Precision is the fraction of retrieved documents that are relevant and recall is the fraction of relevant documents that were retrieved. The F1-score is the harmonic mean of the two, and is a standard information retrieval metric. It can take values between 0 and 1 (see equation below). An overview of results for the baseline dictionaries and expanded dictionaries can be found in Table 5.2.

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

Table 5.2: Overview Results

	PQs			New York Times		
	Pre	Rec	F1	Pre	Rec	F1
Baseline 1	.74	.32	.44	.80	.16	.27
Baseline 2	.57	.45	.48	.57	.40	.42
Single-term 150	.66	.64	.64	.61	.51	.54
Two-step 100/50	.68	.66	.66	.65	.63	.63

Baseline Approaches

We compared two baselines. In the first baseline, we used the basic dictionary, which contained three search terms for each policy issue. See Table 2. When using the basic dictionary, precision was equal to 0.74 for news articles and 0.80 for PQs, and recall was equal to 0.32 for news articles and 0.16 for PQs. F1-scores were equal to 0.27 for news articles and 0.44 for PQs.

These numbers show that precision was fairly high for all issues, while recall was generally low. This is what one would expect, because documents that contain the initial search terms are most likely about the issue that the terms indicate. However, many relevant documents were not retrieved, because the initial search terms do not cover most aspects of the policy issues.

In the second baseline, we used all words from the Lexicoder topic dictionary as search terms. The Lexicoder topic dictionary performed much better than the basic dictionary. Precision was equal to 0.57 for news articles and 0.55 for PQs, and recall was equal to 0.40 for news articles and 0.45 for PQs. F1-scores were equal to 0.42 for news articles and 0.48 for PQs. Compared to the first baseline, where we used our basic dictionaries, this is a strong performance improvement in terms of recall and F1-scores.

Experimental Approaches

Next, we present results from the two experimental approaches, in which we automatically expanded the baseline dictionary. In the first approach ("single-term similarity"), we expanded the baseline dictionary by adding the top N most similar words and phrases for each initial search term. We did this for various N's: 25, 50, 100, 150 and 200.

When coding the PQs dataset, results show that this approach was most effective for $N = 150$. Performance, in terms of F1-scores, steadily increased as we increased N and peaked at 150.[12] When adding the 150 most similar words and phrases for each initial search term, precision was equal to 0.64, recall was equal to 0.66 and the F1-score was equal to 0.64. As compared to the basic dictionary (first baseline), this is a 106% increase in recall and a 45% increase in the F1-score. Precision, however, slightly decreased (10%) when expanding the dictionary. Overall, we found a positive impact on coding performance.

When coding abstracts from the New York Times index, we found a similar pattern. Again, performance peaked when expanding the basic dictionary with the 150 most similar words and phrases. Precision was equal to 0.51, recall was equal to 0.61 and the F1-score was equal to 0.54. As compared to the basic dictionary (first baseline), this is a 219% increase in recall and a 100% increase in the F1-score. Precision, decreased by about 24% when expanding the dictionary. As with the PQs dataset, we found a positive overall impact on coding performance.

In Table 5.3 and Table 5.4, we take a closer look at the coding performance for individual policy issues. Generally, we found substantial differences between issues.

[12] F1-scores for $N = 150$ are only marginally higher than when adding the $N = 50$ most similar terms.

Table 5.3: Results Basic Dictionaries "Baseline 1"

	PQs				New York Times			
	Pre	Rec	F1	N	Pre	Rec	F1	N
Economy	0.70	0.17	0.28	1241	0.87	0.18	0.29	9358
Welfare/Housing	0.69	0.28	0.40	642	0.79	0.17	0.28	1951
Defense	0.89	0.31	0.46	1163	0.83	0.20	0.33	2769
Science/Technology	0.41	0.10	0.16	88	0.45	0.04	0.07	1147
Civil Rights	0.46	0.16	0.23	329	0.67	0.22	0.33	951
Health	0.82	0.49	0.61	864	0.76	0.23	0.36	1535
Education	0.79	0.43	0.56	568	0.74	0.13	0.22	1657
Environment	0.58	0.15	0.24	452	0.87	0.08	0.14	1301
Energy	0.65	0.43	0.52	137	0.79	0.14	0.18	629
Transportation	0.71	0.43	0.54	336	0.58	0.05	0.09	1456
Crime	0.78	0.55	0.65	778	0.84	0.22	0.35	2728
Housing	0.78	0.55	0.65	778	0.84	0.22	0.35	2728
Agriculture	0.78	0.55	0.65	778	0.84	0.22	0.35	2728
Average	0.74	0.32	0.44	6598	0.80	0.16	0.27	

Table 5.4: Results Expanded Dictionaries "single-term similarity"

	PQs				New York Times			
	Pre	Rec	F1	N	Pre	Rec	F1	N
Economy	0.60	0.69	0.64	1241	0.79	0.50	0.62	9358
Welfare	0.44	0.43	0.44	642	0.17	0.51	0.25	1951
Defense	0.85	0.66	0.74	1163	0.67	0.63	0.65	2769
Science/Technology	0.35	0.34	0.35	88	0.30	0.22	0.26	1147
Civil Rights	0.42	0.37	0.39	329	0.36	0.52	0.42	951
Health	0.83	0.76	0.79	864	0.66	0.51	0.58	1535
Education	0.74	0.75	0.75	568	0.53	0.63	0.58	1657
Environment	0.32	0.55	0.40	452	0.34	0.44	0.38	1301
Energy	0.36	0.74	0.49	137	0.38	0.75	0.51	629
Transportation	0.67	0.42	0.52	336	0.31	0.36	0.33	336
Crime	0.66	0.70	0.78	778	0.55	0.49	0.52	2728
Housing	0.44	0.53	0.48	778	0.36	0.52	0.42	2728
Agriculture	0.66	0.54	0.60	778	0.48	0.52	0.50	2728
Average	0.66	0.64	0.64	6598	0.61	0.51	0.54	

In the second approach ("two-step similarity"), we performed two iterations of single-term similarity expansions. In the first iteration, we added the N most similar words for each search term in the baseline dictionary. In the second iteration, we added the M most similar words for all terms that were added in the first iteration. We did this with different numbers for N and M.

When analyzing PQs, we found the best performance for $N = 100$ and $M = 50$. With these values for N and M, recall was equal to 0.66, precision was equal to 0.68 and the F1-score was equal to 0.66. As compared to the "single term similarity" approach, this is a 3 % increase in both recall and the F1-score.

Also when analyzing New York Times abstracts, we found the best performance for $N = 100$ and $M = 50$. Here, recall was equal to 0.63 and precision was equal to 0.65. The F1-score was equal to 0.63. As compared to the "single term similarity" approach, this is a 24% increase in recall and a 17% increase in the F1-score.

These results show that conducting two iterations of search query expansion ("two step similarity") is more effective than one iteration ("single term similarity"). The gained additional performance is small for the PQs, but large for the news articles. One possible explanation for this difference is that news coverage of a specific policy issue covers a broader set of sub-issues and events than PQs about the same policy issue. Therefore, adding a more diverse set of additional search terms has a larger effect on the coding of news articles than on the coding of PQs.

With regard to the Lexicoder topic dictionary, results of our tests show that automatically expanded dictionaries can improve on the manually compiled Lexicoder dictionary. When comparing the best-performing expanded dictionary ("two-step similarity") with the Lexicoder topic dictionary, we found an 51% increase in recall and a 38% increase in the F1-score for PQs. For news articles, we found a 58% increase in recall and a 50% increase in the F1-score. Furthermore, precision increased for PQs (19%) as well as news articles (14%).

5.8 Discussion

We introduced and tested a method to semi-automatically expand coding dictionaries for content analysis. Results of our tests show that using the method can improve the recall and overall coding performance of a dictionary.

What are the implications for future communication research? Foremost, this method is a tool to scholars that (1) increases the performance of content analysis and (2) makes content analysis a less laborious task. The method's strength is its ability to reveal words referring to aspects of a content category with which one is less familiar. In doing so, it helps scholars finding words and phrases that one would not think about in the first place. Consequently, a more complete and better-performing coding dictionary can be created, and the validity of content analysis increases.

Traditionally, scholars read samples of relevant documents to gain domain knowledge and identify search terms [Günther and Quandt, 2015]. Our method presents a more efficient way of doing so. This is especially useful at the initial stage of the dictionary creation process. Automatically retrieved search terms can directly be used as a dictionary, but they can also be used as a pool of candidate terms from which one can select.

A particular strength of the method is that the resulting dictionary is tailored to the specific use of language within a domain (e.g., legislation). Earlier research has shown that the vocabulary used to describe a content category is domain-specific [Zohar et al., 2013]. Consequently, a dictionary developed for one domain often performs less well when applied to another domain. Using our method, a dictionary can easily be expanded with domain-specific search terms.

But the method also provides an efficient way of improving or updating an existing dictionary. This is important, because the vocabulary used to describe a concept in language changes constantly. When training the neural network language model on an up to date corpus, relevant new search terms can be identified and included in the dictionary.

Automatic dictionary expansion is especially useful for research projects in which documents from different languages are studied. The European Election studies, where content analysis is conducted in all EU countries, are such a case [De Vreese et al., 2006]. Another case is comparative agenda setting research [Green-Pedersen and Wilkerson, 2006]. Developing automatic content analysis tools for coding topics in different languages is one of the biggest challenges of the Policy Agendas Project.

In the above cases, a language-specific dictionary can automatically be created by means of the query-expansion approach presented in this chapter. This way our query expansion approach can enrich theorizing in agenda setting research. It can help studying agenda setting effect in a cross-national context.

Furthermore, coding dictionaries can be used to code the same document into different content categories and rank them by relevance. One can compute a relevance score for each combination of a topic and a content category by counting the number of matching search terms for each category [Albaugh et al., 2013]. But one can also employ more advanced relevance scores, like we did with the Lucene scoring function [McCandless et al., 2010].

A ranked multi-label coding is relevant for many concepts in communication research - particularly when studying longer documents like entire news articles. Then, most likely, different topics or events are discussed in the same document and such topics and events are framed in various ways.

There are several limitations to our research. In our tests, we only applied the method to the coding of policy issues. An important question is whether the same method can be used to create dictionaries for coding other concepts in political communication research. We believe that this is the case, but that certain limitations apply.

A relevant concept in political communication research are frames [Entman, 1993]. Different types of frames have been defined [De Vreese, 2005]. This method can be useful to construct dictionaries for the coding of issue-specific frames [Shah et al., 2002], but it can also be used to code generic news frames like the attribution of conflict. In that case, one could use the method to retrieve words that indicate, for instance, various sorts of conflict. This should be studied in future research.

What are other limitation of the study? First, we tested it on relatively short documents. The New York Times Index consists of very brief abstracts for each article. Similarly, the PQs are relatively short documents, which generally consist of one to three sentences. But how does the method perform when applied to longer documents like full-length news articles?

With policy issues we see that even when coding relatively short texts, one document contains search terms from different policy issues.

This is because policy often involves multiple issues at the same time. This makes the coding of policy issues a challenging test of the method. Generally, our search procedure provides a good strategy to find out which issue is more dominant by counting and weighting the matching search terms from all issues. However, handling overlap in topics remains a challenge in dictionary-based content analysis, especially when coding longer documents.

Second, our choice of the initial search terms for the basic dictionary might seem arbitrary. What is the influence of the choice and the number of initial search terms on the performance of our method? For each content category we used the name of the policy issue as the first search term. The remaining two words are based on the descriptions from the code book of the Policy Agendas Project. When using another set of words, results change by definition. However, this change is marginal. We have tested this for a few cases, and saw that results are very similar. This means that the method is rather robust.

We used not more than three initial search terms in order to reveal the full potential of the method. We expect the performance of automatically expanded dictionaries to increase if we use a larger set of initial search terms. When selecting initial search terms, it is important that the terms cover some of the topical diversity within a category.

Third, we discuss several approaches for retrieving semantically related words in the theoretical framework. However, we do not empirically compare these approaches with our neural network language model. Comparing different solutions to the problem of automatically expanding coding dictionaries is beyond the scope of this chapter, but should be addressed in future research.

Finally, this study raises several other questions that should be addressed in future research: First, the effects of combining different datasets for training the language model should be explored in more detail. Earlier research on search query expansion has shown that one can improve performance by using different training sets [Senellart and Blondel, 2008].

Second, future research should focus on improving the search procedure, such that one can find a better fit between increasing recall but preserving precision. We used the standard scoring function from the Elastic-search software. However, we expect that a different scoring model can further improve the coding procedure.

Although results of our study suggest that the proposed method is able to automatically create a well-performing coding dictionary with very little manual input or supervision, we conclude that it can best be used in combination with human expert knowledge. We consider human expert knowledge very important to the process of building a good dictionary. Therefore, we merely see the method as a tool to facilitate the process of building a coding dictionary, rather than to completely automate it.

5.9 Bibliography

Elisabeth Günther and Thorsten Quandt. Word counts and topic models: Automated text analysis methods for digital journalism research. *Digital Journalism*, 4(1):75–88, 2015.

Philip A Schrodt, Shannon G Davis, and Judith L Weddle. Political science: KEDS—a program for the machine coding of event data. *Social Science Computer Review*, 12(4):561–587, 1994.

Justin Grimmer and Brandon M Stewart. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21:267–297, 2013.

Lori Young and Stuart Soroka. Affective news: The automated coding of sentiment in political texts. *Political Communication*, 29(2):205–231, 2012.

Dustin Hillard, Stephen Purpura, and John Wilkerson. Computer-assisted topic classification for mixed-methods social science research. *Journal of Information Technology & Politics*, 4(4):31–46, 2008.

Pierre Senellart and Vincent D Blondel. Automatic discovery of similar words. In *Survey of Text Mining II*, pages 25–44. Springer, 2008.

Wouter Weerkamp, Krisztian Balog, and Maarten de Rijke. Exploiting external collections for query expansion. *ACM Transactions on the Web*, 6(4):18, 2012.

John R Firth. A synopsis of linguistic theory. *Studies in Linguistic Analysis*, pages 1–32, 1957.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Y Bengio and Y Lecun, editors, *Proceedings of the International Conference on Learning Representations*, pages 1–12, Scottsdale, AZ, 2013a.

Quinn Albaugh, Julie Sevenans, Stuart Soroka, and Peter John Loewen. The automated coding of policy agendas: A dictionary-based approach. In *Proceedings of the 6th Annual Comparative Agendas Conference*, Antwerp, Belgium, 2013. CAP.

Conny Roggeband and Rens Vliegenthart. Divergent framing: The public debate on migration in the Dutch parliament and media, 1995–2004. *West European Politics*, 30(3):524–548, 2007.

Stuart N Soroka, Dominik A Stecula, and Christopher Wlezien. It's (change in) the (future) economy, stupid: Economic indicators, the media, and public opinion. *American Journal of Political Science*, 59 (2):457–474, 2015.

Jaeho Cho, Michael P Boyle, Heejo Keum, Mark D Shevy, Douglas M McLeod, Dhavan V Shah, and Zhongdang Pan. Media, terrorism, and emotionality: Emotional differences in media content and public reactions to the September 11th terrorist attacks. *Journal of Broadcasting & Electronic Media*, 47(3):309–327, 2003.

Frank R Baumgartner, Christoffer Green-Pedersen, and Bryan D Jones. Comparative studies of policy agendas. *Journal of European Public Policy*, 13(7):959–974, 2006.

Rens Vliegenthart and Stefaan Walgrave. The contingency of intermedia agenda setting: A longitudinal study in Belgium. *Journalism & Mass Communication Quarterly*, 85(4):860–877, 2008.

Stefaan Walgrave, Stuart Soroka, and Michiel Nuytemans. The mass medias political agenda-setting power: A longitudinal analysis of media, parliament, and government in Belgium (1993 to 2000). *Comparative Political Studies*, 41(6):814–836, 2007.

Bjorn Burscher, Rens Vliegenthart, and Claes H De Vreese. Using supervised machine learning to code policy issues can classifiers generalize across contexts? *The ANNALS of the American Academy of Political and Social Science*, 659(1):122–131, 2015.

Kevin M Quinn, Burt L Monroe, Michael Colaresi, Michael H Crespin, and Dragomir R Radev. An automated method of topic-coding legislative speech over time with application to the 105th-108th US Senate. In *Proceedings of the Midwest Political Science Association Meeting*, pages 1–61, Austin, TX, 2006. MPSA.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and ROC curves. In W Cohen and A Moore, editors, *Proceedings of the 23rd international conference on machine learning*, pages 233–240, Pittsburgh, PA, 2006. ACM.

Gregory Grefenstette. *Explorations in automatic thesaurus discovery*. Kluwer Academic Publishers, Norwell, MA, 2012.

Claudio Carpineto and Giovanni Romano. A survey of automatic query expansion in information retrieval. *ACM Computing Surveys (CSUR)*, 44(1):1–13, 2012.

Hadas Zohar, Chaya Liebeskind, Jonathan Schler, and Ido Dagan. Automatic thesaurus construction for cross generation corpus. *Journal on Computing and Cultural Heritage (JOCCH)*, 6(1):4, 2013.

Jing Bai, Dawei Song, Peter Bruza, Jian-Yun Nie, and Guihong Cao. Query expansion using term relationships in language models for information retrieval. In O Herzog and HJ Schek, editors, *Proceedings of the 14th ACM international conference on information and knowledge management*, pages 688–695, New York, NY, 2005. ACM.

Mandar Mitra, Amit Singhal, and Chris Buckley. Improving automatic query expansion. In WB Croft and A Moffat, editors, *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval*, pages 206–214, New York, NY, 1998. ACM.

George A Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

Giannis Varelas, Epimenidis Voutsakis, Paraskevi Raftopoulou, Euripides GM Petrakis, and Evangelos E Milios. Semantic similarity methods in WordNet and their application to information retrieval on the web. In A Bonifati and L Dongwon, editors, *Proceedings of the 7th annual ACM international workshop on Web information and data management*, pages 10–16, New York, NY, 2005. ACM.

Roberto Navigli and Paola Velardi. An analysis of ontology-based query expansion strategies. In N Lavrac and D Gamberger, editors, *Proceedings of the 14th European Conference on Machine Learning*, pages 42–49, Cavtat-Dubrovnik, Croatia, 2003.

Hsinchun Chen and Kevin J Lynch. Automatic construction of networks of concepts characterizing document databases. *IEEE Transactions on Systems, Man and Cybernetics*, 22(5):885–902, 1992.

Magnus Sahlgren. The distributional hypothesis. *Italian Journal of Linguistics*, 20(1):33–54, 2008.

Kenneth W Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1): 22–29, 1990.

Holger Bast, Debapriyo Majumdar, and Ingmar Weber. Efficient interactive query expansion with complete search. In AHF Laender, editor, *Proceedings of the sixteenth ACM conference on information and knowledge management*, pages 857–860, New York, NY, 2007. ACM.

Jiani Hu, Weihong Deng, and Jun Guo. Improving retrieval performance by global analysis. In YY Tang and P Wang, editors, *Proceedings of the 18th International Conference on Pattern Recognition*, pages 703–706, Washington, DC, 2006. IEEE.

Jinxi Xu and W Bruce Croft. Query expansion using local and global document analysis. In HP Frei, editor, *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 4–11, Zurich, Zwitzerland, 1996. ACM.

James R Curran and Marc Moens. Scaling context space. In P Isabelle, editor, *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 231–238, Philadelphia, PA, 2002. Association for Computational Linguistics.

Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *JAsIs*, 41(6):391–407, 1990.

Dongqiang Yang and David M Powers. Automatic thesaurus construction. In D Gillian and M Bernard, editors, *Proceedings of the thirty-first Australasian conference on Computer science*, pages 147–156, Darlinghurst, Australia, 2008. Australian Computer Society.

Aron Henriksson, Hans Moen, Maria Skeppstedt, Vidas Daudaravicius, and Martin Duneld. Synonym extraction and abbreviation expansion with ensembles of semantic spaces. *Journal of Biomedical Semantics*, 5(6), 2014.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137–1155, 2003.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In L Vanderwende, editor, *Proceedings of the 2103 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 746–751, Atlanta, GA, 2013b. ACL.

Victor Lavrenko and W Bruce Croft. Relevance based language models. In DH Kraft and WB Croft, editors, *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 120–127. ACM, 2001.

B Yegnanarayana. *Artificial neural networks*. PHI Learning, Delhi, India, 2009.

Radim Řehůřek and Petr Sojka. Software framework for topic modelling with large corpora. In R Witte and H Cunningham, editors, *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, 2010. ELRA.

Lior Wolf, Yair Hanani, Kfir Bar, and Nachum Dershowitz. Joint word2vec networks for bilingual semantic representations. *International Journal of Computational Linguistics and Applications*, 5(1): 27–44, 2014.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. Learning sentiment-specific word embedding for Twitter sentiment classification. In L Toutanova and H Wu, editors, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1555–1565, Baltimore, MA, 2014. ACL.

Cicero D Santos and Bianca Zadrozny. Learning character-level representations for part-of-speech tagging. In EP Xing and T Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1818–1826, Beijing, 2014.

Mihajlo Grbovic, Nemanja Djuric, Vladan Radosavljevic, Fabrizio Silvestri, and Narayan Bhamidipati. Context-and content-aware embeddings for query rewriting in sponsored search. In R Baeza-Yates, editor, *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 383–392, Santiago, Chile, 2015. ACM.

Yoav Goldberg and Omer Levy. word2vec explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*, 2014.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013c.

Shaun Bevan and Peter John. Policy representation by party leaders and followers: What drives UK Prime Minister's Questions? *Government and Opposition*, 51:1–25, 2015.

Claes H De Vreese, Susan A Banducci, Holli A Semetko, and Hajo G Boomgaarden. The news coverage of the 2004 European Parliamentary election campaign in 25 countries. *European Union Politics*, 7(4): 477–504, 2006.

Christoffer Green-Pedersen and John Wilkerson. How agenda-setting attributes shape politics: basic dilemmas, problem attention and health politics developments in Denmark and the US. *Journal of European Public Policy*, 13(7):1039–1052, 2006.

Michael McCandless, Erik Hatcher, and Otis Gospodnetic. *Lucene in Action: Covers Apache Lucene 3.0*. Manning Publications, London, UK, 2010.

Robert M Entman. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58, 1993.

Claes H De Vreese. News framing: Theory and typology. *Information design journal+ document design*, 13(1):51–62, 2005.

Dhavan V Shah, Mark D Watts, David Domke, and David P Fan. News framing and cueing of issue regimes: Explaining Clinton's public approval in spite of scandal. *Public Opinion Quarterly*, 66(3):339–370, 2002.

6

Conclusion

The previous chapters consisted of four empirical studies, all of which discuss the application of automated content analysis methods in framing and agenda setting research. In this final chapter, we summarize key findings of the dissertation and discuss more broadly the methodological and theoretical implications of automated content analysis for future communication research.

6.1 Key Findings

We used machine learning to study policy issues and frames in political messages. With regard to frames, we investigated the automation of two content-analytical tasks: frame coding and frame identification. We found that both tasks can be successfully automated by means of machine learning techniques. Frame coding can be automated through supervised machine learning (SML). Results show that the performance of SML-based frame coding approaches the performance of human coders (Chapter 2).

Furthermore, we have shown that frames can be automatically identified through clustering, a form of unsupervised machine learning. We used this method to identify issue frames in the nuclear power debate. We found that automatically identified frames closely resemble frames that have been identified in previous studies, by means of qualitative approaches (Chapter 3).

In addition, we have shown that policy issues can be coded by means of SML (Chapter 4) as well as through semi-automatically created dictionaries (Chapter 5). Again, automatic coding approaches the performance of human coders. Moreover, we demonstrated that SML and dictionary-based coding can be applied to different types of political messages (e.g., news articles and parliamentary records).

Feature Selection

An important aspect of machine learning based content analysis is feature selection. Feature selection refers to the choice of elements of a text document that are used to represent the document in the machine learning process. Generally, these are the words of a text document. But not all words are even useful features in each content-analytical task. We investigated the use of different feature sets when studying policy issues (Chapter 4) and frames (Chapter 3).

We found that whether a word is a good indicator of a frame or policy issue depends on the type of the word (e.g, whether or not it is a named entity) and on the position of the word in the document (e.g., whether it is part of the title). We, therefore, conclude that machine learning based content analysis can be improved by selecting additional features that are not commonly used now. In the following paragraphs we explain our findings in detail.

When using clustering to identify issue frames in news articles, careful feature selection is essential to get clusters that differentiate between different issue frames (Chapter 3). In previous research, scholars argued that not all words are equally important to a news frame [Carragee and Roefs, 2004, Hertog and McLeod, 2001]. We found that some words in a news article are indeed better indicators of the article's framing than others, and that frame identification improves when choosing these words as features.

First, we found that frame clustering improves if one removes names of persons, countries and organizations from the feature space. This is because such words lead to clusters that are centered around specific spaces and events. Second, we found that nouns and adjectives are more indicative of frames than words with other parts of speech. Third, our findings indicate that words from the title and lead of a news article are highly indicative of issue frames. This is because news articles are structured in a way that they present information in terms of relative importance.

We, furthermore, found that the relevance of words from title and lead also holds when studying other characteristics of news coverage. When analyzing policy issues, we showed that words from the title and lead are particular indicative features. When using only the first ten percent of the words of a news article as features, classification performance is nearly identical as when one uses all words of the article.

Generalization

Generalization is another important issue in machine learning based content analysis. Remember that the basic idea behind supervised learning is using already annotated example documents to train a classification model for a specific content characteristic. Once the model is trained it can be used to code this characteristic in unseen documents. But what if the unseen documents systematically differ from the training examples?

We empirically addressed this question by applying classifiers to the coding of documents from unseen time periods, sources and message types. We found that classification accuracy decreases slightly when applying a classifier to an unknown newspaper (Chapter 2 and 4), and strongly when applying it to documents from unknown time periods (Chapter 4) or to another type of political message (e.g., news articles vs. parliamentary questions). From this, we conclude that to achieve good performance training data should be representative of the different outlets, time periods, and message types that one wants to study.

We found that our classifiers are particularly bad in generalizing across time. This is because attention within a specific policy domain constantly shifts toward new events and problems. Consequently, also the set of words indicating the policy issue changes. In order to deal with such changes, one should keep training sets up to date and retrain classifiers at regular intervals.

Generalization is a relevant property of classifiers, because one goal of automated content analysis is to facilitate the analysis of large and heterogeneous datasets. This is particularly relevant in comparative and longitudinal research, because documents from several sources and time periods have to be coded. This often is the case in framing as well as agenda setting research, where the causal direction of effects and the conditions under which effects occur are theoretically relevant questions.

Training Data

When making use of SML, it is useful to know how much training data one needs. This is because collecting training data is a costly undertaking, whereas the reason for using machine learning is to decrease the costs of coding. We studied the relation between the amount of training data used and the performance of a classifier when coding frames (Chapter 2) and policy issues (Chapter 4).

Overall, results show that increasing the number of training documents leads to increased classification performance. This relationship, however, is not linear. After reaching a certain training size coding performance increases only slowly when adding additional training documents. This might be due to the fact that one has reached a performance peak and additional training documents provide little new information for the classifier.

Another explanation is that some content categories are underrepresented in the training set. When building a classifier that differentiates between a larger number of content categories and some of these categories are rare, performance for such categories might be poor even though one has a large overall training set. In that case, instead of just increasing the size of the training set, it is more effective to selectively sample positive examples for underrepresented categories. This is, for example, the case when coding certain policy issues from the Policy Agendas Project (e.g., Civil Rights and Minority Issues).

But variation in coding performance between content categories can also be due to the fact that some phenomena are more difficult to learn than others. Results of our studies show that this is the case for certain frames and policy issues. Classification performance of the economic consequences frame, for instance, improves the most when increasing the size of the training set, although this frame is not the most prevalent one. This indicates that some categories are more complex phenomena. We see a similar phenomenon in manual coding, where inter-coder agreement is higher for some content categories than for others (Chapter 2).

Indicator-Based Content Analysis

The application of machine learning methods in political communication research is not always straight forward. In manual content analysis, communication scholars often combine answers to multiple questions in order to measure one content characteristic. This is called indicator-based content analysis, and is often applied when coding news frames [Semetko and Valkenburg, 2000]. Generally, measures of several questions are used to cover different aspects of a frame.

We investigated how useful it is to model each individual indicator question when automatically coding frames using SML (Chapter 2). In doing so, we compared two approaches. In the first approach, we built a classifier to automatically code each of the indicator questions, which we then aggregated to a single frame measure. In the second approach, we built a classifier to directly predict the presence of a frame.

Results of our experiments show that it is more effective to train a classifier to predict the presence of a frame directly. From this, we conclude that when applying SML, it is not always appropriate to proceed as in manual content analysis.

Comparing approaches like this is relevant for developing knowledge about how machine learning should be used to effectively master content analysis problems in communication research. SML is a set of algorithms and approaches for automatic classification. Finding the optimal way of performing a specific classification task generally involves comparing various models.

Automatic Dictionary Creation

So far, we presented findings of studies where we used machine learning to directly analyze text documents. But we also investigated another application of machine learning - as a way to facilitate dictionary-based content analysis (Chapter 5).

Dictionary-based coding is a popular form of automatic content analysis in communication research. The biggest challenge in dictionary-based coding is the creation of the coding dictionary. Most scholars cannot recall all relevant words that indicate a content category, and they cannot recall all ways such words can be used in language. Consequently, relevant search terms are missing in the coding dictionary, and not all documents can be coded correctly.

We found that machine learning techniques can be applied to deal with this problem and, therefore, facilitate the creation of coding dictionaries. Given a corpus of news articles, a neural network language model can be used to compute similarity statistics between all words from the vocabulary of the corpus. Based on such similarity statistics, one can retrieve semantically similar words and use them to expand a coding dictionary with relevant search terms.

Given a basic coding dictionary, which contains just a few search terms per category, the approach can be used to automatically expand the dictionary and improve its performance. The method is useful when building a new coding dictionary, but can also be used to update and/or improve an existing dictionary.

This approach is especially useful in situations where no training data is available to apply supervised learning, or where creating training data is not feasible. One example are large cross-national studies like the European election study, which involves content analysis in more than

25 countries.

6.2 Discussion

Huge amounts of political media content are produced every day and become digitally available [Günther and Quandt, 2015]. This stimulates the development of new methods for automatic content analysis [Boumans and Trilling, 2016]. Turning to research methods from disciplines such as computer science, artificial intelligence and computational linguistics provides new opportunities - for empirical analysis and theory development. In this dissertation, we investigated how such methods can be used best in order to study the contents of political communication.

We created knowledge regarding which forms of automated content analysis and, in particular machine learning, can be applied to framing and agenda setting research. We investigated how such methods should be implemented, how well they perform, and what their limitations are.

All in all, we conclude that machine learning and automated content analysis have great potential for advancing theory development in communication research. With regard to framing, clustering can enhance frame identification by offering a way to objectively and validly identify frames based on large datasets. This is relevant as it facilitates theorizing about the persistence of issue frames under various conditions (e.g., different sorts of media).

Furthermore, automated frame coding facilitates large scale content analysis in framing research. This makes it easier to study the causal direction and conditionalities of framing effects outside the laboratory, where people are exposed to a variety of different, opposing media messages [Chong and Druckman, 2007].

Moreover, in agenda setting studies automated coding can advance theorizing about the long-term dynamics between the media, the public and the political agenda, and provide new insights regarding the direction of agenda setting effects.

Finally, machine learning methods facilitate the study of uncommon concepts outside the laboratory. It is a challenge for content analysis that some categories are less prevalent than others in political communication. If a phenomenon is very rare, it is difficult to measure its occurrence and assess its effects. Machine learning provides a solution to this problem, because it allows us to easily increase the scope of content analysis.

In the following sections, we elaborate on several practical issues related to automated content analysis. This includes questions like "When should we use computational methods?, "How accessible are computational methods?" and "How can we improve tools for automatic content analysis?".

Spoilt for choice?

We discussed three different methods for automated content analysis - supervised learning, unsupervised learning and dictionary-based coding? But in which situation should we use each of these methods? For the main part, the choice of method should depend on the research question at hand. Nonetheless, we can formulate a few guidelines based on the research in this dissertation.

When choosing between these methods one important question is whether one already knows the content categories to code messages into. For instance, have we already defined a set of frames or policy issues that we want to study? If yes, supervised machine learning (SML) or dictionary-based coding are the most obvious choices. Both approaches require a finite set of a priori defined content categories and can be used to code each document into one of these categories.

Another important question is whether one possesses (or can generate) labeled training data? If so, and if the quality of the training data is sufficient, SML can be applied. However, often training data is not available or expensive to generate. This can be the case in situations where one has rare content categories or if the research project requires the coding of documents in several languages. Then, dictionary-based coding can be the more practical choice.

Finally, unsupervised machine learning is mostly suited for cases in which one wants to explore the underlying structure of a data set, and identify relevant content categories. Frame identification is a good example for the application of unsupervised learning. But unsupervised learning can also be applied in agenda setting research to explore the underlying topic structure of a document collection.

It is important to note that unsupervised learning methods can also be applied to the coding of documents (Chapter 3). The important difference with supervised learning is that one has little influence on the final set of content categories. Once a clustering algorithm has distinguished a certain set of clusters, new documents can only be coded according to this cluster structure.

Can I do machine learning myself?

Automatic content analysis has become increasingly popular among communication scholars. While dictionary-based approaches have been applied for several decades, the use of machine learning is a more recent phenomenon [Günther and Quandt, 2015, Grimmer and Stewart, 2013]. Having discussed several key findings regarding the application of machine learning, we want to discuss a more practical aspect - the accessibility of machine learning methods.

This is important, because most machine learning algorithms are not part of the standard statistical packages that the majority of communication scholars use (e.g., SPSS or Stata), and most communication scholars have not been trained in using machine learning methods.

For dictionary-based content analysis several software packages have been released in the past decades (e.g., Lexicoder [1]). Such packages are all very similar in their core functionality, which is the counting of search terms in text documents. Generally, these packages have graphical user interfaces and are easy to use.

However, when it comes to machine learning, things are less straight forward. Machine learning not only involves the application of a learning algorithm, but also the extensive processing of text documents. The latter includes all the steps needed to gather text data and bring it into the right format.

Common tasks are downloading documents from various digital sources, cleaning and pre-processing the raw text (e.g., parsing, stopword removal and stemming), partitioning the data (e.g., creating train and test sets), and transforming data so that a machine learning algorithm can read it (e.g., feature extraction and vectorization).

There is machine learning software that comes with a graphical interface (e.g., Weka [2]) and there are also graphical tools for most text processing tasks. However, in order to tap the full potential of machine learning and to create problem-tailored solutions, we strongly recommend the use of a programming language. This is because the majority

[1] http://www.lexicoder.com/

[2] https://weka.wikispaces.com/

of all machine learning and natural language processing packages are only available as libraries for programming languages.

Furthermore, working in a programming environment is often much more efficient than using different graphical programs. This is because one can easily integrate the functionality of various libraries in a single task-tailored script.

Several programming languages come with extensive libraries for machine learning and natural language processing. In this dissertation, we did all machine learning tasks in the Python programming language. Python has several libraries for machine learning (e.g., Scikit-Learn and Gensim) and natural language processing (e.g., NLTK). However, other programming languages (R, Java, C++, etc.) provide similar functionality.

We conclude that in order to apply machine learning in content analysis one has to familiarize oneself with the different methods and algorithms available. Furthermore, it is extremely helpful to acquire some programming skills. Programming is not a skill to acquire within a day, but neither does it take years. Various books, tutorials and online courses provide excellent introductions to programming and machine learning.

We hope that programming will become a standard part of (academic) education in the future. This is important in order to remove obstacles for using advanced methods for automated content analysis. We already see a trend in the social sciences towards using open-source programming tools like R or Python for statistical analysis. We strongly encourage communication scholars to use such tools and also teach their students to use them.

Sharing is Caring

In the past years, various communication scholars have developed software for automated content analysis [Young and Soroka, 2012, Baccianella et al., 2010]. We believe that this is an important trend. Therefore, we encourage scholars to keep on doing this in the future, and also make their software available to the research community. This includes applications for text processing, implementations of machine learning algorithms, and classification models. Sharing software is necessary in order to promote the adoption of innovative research methods in the field of communication.

We, furthermore, encourage scholars to publish their applications as open source-software. This can be in the form of scripts or fully implemented packages for open programming languages like Python on R. We speak out against the use of closed software, as it cannot be easily combined with existing machine learning applications.

However, in order to advance automatic content analysis, we not only need to develop new tools and methods, but should also share data. It should become a collective effort among communication scholars to develop and maintain data sets, which can be used by the community at large. This includes the systematic manual coding of large amounts of political messages from various sources and time periods as well as in different languages.

Such training data is critical for building robust coding tools, which can become a validated standard in the field. This is important for theory development in agenda setting research as well as framing research. In order to better understand how agenda setting and framing effects vary across countries, political systems and media systems, we need coding tools that perform similarly across languages and sources.

6.3 Bibliography

Kevin M Carragee and Wim Roefs. The neglect of power in recent framing research. *Journal of Communication*, 54(2):214–233, 2004.

James K Hertog and Douglas M McLeod. A multiperspectival approach to framing analysis: A field guide. In SD Reese, OH Gandy, and AE Grant, editors, *Framing public life: Perspectives on media and our understanding of the social world*, pages 139–161. Lawrence Erlbaum, Mahwah, NJ, 2001.

Holli A Semetko and Patti M Valkenburg. Framing European politics: A content analysis of press and television news. *Journal of Communication*, 50(2):93–109, 2000.

Elisabeth Günther and Thorsten Quandt. Word counts and topic models: Automated text analysis methods for digital journalism research. *Digital Journalism*, 4(1):75–88, 2015.

Jelle W Boumans and Damian Trilling. Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital Journalism*, 4(1): 8–23, 2016.

Dennis Chong and James N Druckman. Framing theory. *Annual Review of Political Science*, 10:103–126, 2007.

Justin Grimmer and Brandon M Stewart. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21:267–297, 2013.

Lori Young and Stuart Soroka. Lexicoder sentiment dictionary. *McGill University, Montreal, Canada. Available online at: www. lexicoder. com*, 2012.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In N Calzolari and K Choukri, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 2010. European Language Resources Association (ELRA).

7

Summary

This dissertation contains four empirical studies. The first two studies address automatic content analysis in framing research and the latter two address automatic content analysis in agenda setting research. It follows a summary of each study.

Study 1

In the first study (Chapter 2), we address frame coding - the annotation of already defined news frames in political messages. The method we apply is supervised machine learning (SML). By automating the coding of frames in news, SML facilitates the incorporation of large-scale content analysis into framing research. This furthers a more integrated investigation of framing processes conceptually as well as methodologically.

We conduct several experiments in which we automate the coding of four generic news frames that are operationalised as a set of indicator questions. In doing so, we compare two approaches to modelling the coherence between indicator questions and frames as an SML task. The results of our experiments show that SML is well suited to automate

frame coding but that coding performance is dependent on the way the problem is modelled.

Study 2

In the second study (Chapter 3), we investigate automatic frame identification. Based on a large collection of news articles, we automatically identify issue frames with regard to the nuclear power debate. For this, we apply clustering, a form of unsupervised machine learning. Furthermore, we test a way of improving statistical frame analysis such that revealed clusters of articles reflect the framing concept more closely. We do so by only using words from an article's title and lead and by excluding named entities and words with a certain part of speech from the analysis.

To validate revealed frames, we manually analyze samples of articles from the extracted clusters. Findings of our tests indicate that when following the proposed feature selection approach, the resulting clusters more accurately discriminate between articles with a different framing.

Study 3

The third study (Chapter 4) deals with the automatic coding of policy issues in news articles and parliamentary questions. We apply supervised machine learning for this. Comparing computer-based annotations with human annotations shows that our method approaches the performance of human coders.

As agenda setting research is concerned with dynamics in issue salience among the media, politicians, and citizens it requires large-scale over-time content analysis across different types of political texts.

Therefore, we investigate the capability of an automatic coding tool, which is based on supervised machine learning, to generalize across contexts.

Study 4

In the last study (Chapter 5), we apply a dictionary based approach to code policy issues is news articles and parliamentary questions. Constructing a dictionary with search terms for several content categories can be a difficult and laborious task. Therefore, in this study, we introduce a method to automatically expand coding dictionaries with relevant search terms. In doing so, we employ word co-occurrence statistics, which are based on word vectors from a neural network language model.

We conduct several tests in which we use this method to automatically expand dictionaries for coding policy issues. We validate our method by applying automatically constructed dictionaries to different human-coded test sets. Results show that we can significantly increase the performance of a coding dictionary by automatically adding search terms.

8

Samenvatting

Dit proefschrift bestaat uit vier empirische onderzoeken. De eerste twee behandelen automatische inhoudsanalyse in framing-onderzoek, en de laatste twee behandelen automatische inhoudsanalyse in agenda-setting-onderzoek. Hier volgt een korte samenvatting per onderzoek.

Onderzoek 1

In het eerste onderzoek (Hoofdstuk 2) behandelen we framecoderen – het annoteren van bestaande nieuwsframes in politieke berichten. Hiertoe wordt *supervised machine learning* (SML) toegepast. Door middel van het automatisch coderen van frames in krantenartikelen, maakt SML het mogelijk grootschalige inhoudsanalyse toe te passen op framingonderzoek. Dit bevordert geintergreerd onderzoek van framingprocessen, op zowel conceptueel als methodologisch gebied.

We voeren verscheidene experimenten uit waarin we vier algemene nieuwsframes geautomatiseerd coderen. Deze frames zijn geoperationaliseerd als een reeks indicatorvragen. We vergelijken hiermee twee verschillende benaderingen voor het modelleren van de samenhang tussen indicatorvragen en frames als een SML-taak. De resultaten van de experimenten tonen aan dat SML geschikt is voor het automatisch coderen van frames, maar dat de kwaliteit van de codering afhangt van de specifieke toepassing van SML.

Onderzoek 2

In onderzoek twee (Hoofdstuk 3) behandelen we de automatische indentificatie van frames. Gebaseerd op een grootschalige collectie van krantenartikelen identificeren we *issue frames* met betrekking tot het kernenergiedebat. Hiertoe passen we *clustering* toe – een vorm van *unsupervised machine learning*. Vervolgens testen we een verbeterde methode van statistisch frameonderzoek waarbij de gevonden clusters van artikelen het framingconcept beter afspiegelen. Dit wordt gerealiseerd door enkel woorden van titels en eerste alinea's te gebruiken en door het uitsluiten van namen en bepaalde woordsoorten in de analyse.

Om de gevonden frames te valideren, analyseren we steekproefgewijs en handmatig artikelen uit gevonden clusters. Onze resultaten tonen aan dat het volgen van de voorgestelde feature-selectie leidt tot een beter onderscheid tussen artikelen met verschillende framings.

Onderzoek 3

Onderzoek drie (Hoofdstuk 4) behandelt de automatische codering van beleidskwesties in nieuwsartikelen en Kamervragen. We passen supervised machine learning toe. Vergelijken van computergebaseerde en menselijke annotaties toont aan dat onze aanpak menselijke prestaties evenaart.

Daar *agenda setting*-onderzoek betrekking heeft op de dynamiek van beleidskwesties tussen media, politici en burgers, vereist onderzoek grootschalige en over breed tijdvlak lopende inhoudsanalyse van diverse soorten politieke teksten. Daarom onderzoeken we de generaliseerbaarheid van een methode voor automatische codering – gebaseerd op supervised machine learning – over verschillende contexten.

Hoofdstuk 4

In onderzoek vier (Hoofdstuk 5) passen we een woordenboek-gebaseerde methode toe om beleidskwesties in krantenartikelen en Kamervragen te coderen. Samenstellen van een woordenboek van zoektermen voor verscheidene inhoudscategorieen is een uitdagende en arbeidsintensieve aangelegenheid. Derhalve introduceren we in dit onderzoek een methode voor het automatisch uitbreiden van coderingswoordenboeken met relevante zoektermen. Hiervoor maken we gebruik van statistieken in woordgebruik, verkregen uit woordvectoren uit een neuraalnetwerk-gebaseerd taalmodel.

We voeren verschillende experimenten uit om – gebruikmakend van de voorgestelde methode – automatisch coderingswoordenboeken voor beleidskwesties uit te breiden. We valideren onze methode door automatisch gegenereerde woordenboeken te testen op verschillende menselijk gecodeerde testsets. Onze resultaten tonen aan dat we de prestaties van coderingswoordenboeken significant kunnen verbeteren door automatische toevoeging van zoektermen.

9

Acknowledgements

It is a pleasure to thank those who made this thesis possible and who accompanied me during the past four years. First, I must thank my promoters Claes, Rens and Maarten for their continuous advice and confidence in me. Claes and Rens, you always trusted me to make my own decisions and gave me the freedom to develop myself as a researcher and person. Furthermore, I want to thank Joost, Peter, Hajo, Franciska and Evangelos for taking part in my dissertation committee. I really appreciate the time you dedicate to my work.

I would also like to thank Joost for being my mentor and colleague for the past seven years. As a student I did my research internship with you and you were my thesis supervisor. Now I am glad to have you as a member of my dissertation committee. I have always valued your feedback, and am thankful for working together with you on several projects through the years.

While working toward my PhD I had the opportunity to collaborate with different people. I am grateful to be part of the COMMIT project and I want to thank all my co-authors for their interest in my work. It was a pleasure to work with you Daan and Damian.

Furthermore, I would like to thank my fellow PhD students at ASCoR. Together, we had a lot of fun during the past years. We went to conferences, criticised each other's work during PhD club, and helped each other through the up's and down's of the PhD life. But most importantly, we became a very close group of friends which I do not want to miss. Thank you very much for everything Tom, Jasper, Guus, Carlos, Alina, Jelle, Toni, Nadine and Michael.

Finally, I want to thank my parents and my boyfriend. Mom and Dad, without your unconditional love and support I would not be the person I am today. I am very proud of you. Martijn, thank you for supporting me, loving me, and pushing me towards learning new things.