VARYING BITS

University of Stavanger
NO-4036 Stavanger
NORWAY
www.uis.no

Cover: Newspaper by Silke Remmery 🅭🅯
https://flic.kr/p/q8M1Fy

ISBN: 978-82-8439-248-6
ISSN: 1890-1387
PhD: Thesis UiS No. 770

Varying Bits

A Computational Perspective on News Diversity and Political Parallelism

by

Erik de Vries

Thesis submitted in fulfilment of
the requirements for the degree of
PHILOSOPHIAE DOCTOR
(PhD)

University
of Stavanger

Faculty of Social Sciences
Department of Media and Social Sciences
2024

Acknowledgements

Without the help of many people, the thesis that lies before you would not have been. Their advice, jokes and support have made writing this dissertation bearable, and frequently even enjoyable! Most importantly, my thanks and gratitude go to my supervisors: Gunnar Thesen, Rens Vliegenthart and Gijs Schumacher for their endless patience, optimism and unwavering dedication to make this project into a success. But also to Helle, Raul, and all other staff at the IMS department, for providing useful comments, support and motivation throughout my PhD. And to Solveig, Ben, and all the other PhD students for being there to talk about concepts and theories, or just to have a chat about the weekend to take my mind off things.

My thanks also go out to all partners in the MaML project, for literally climbing mountains together. To the members of my larger academic family in the Comparative Agendas Project, for countless rounds of comments and feedback after yet another conference presentation. And to Stefaan, for enjoyable discussions on the nature of news diversity and its measurement.

Last, but not even remotely the least, thanks to my family and friends. To Jorien, Thijs and Carla, for being there for me in one of the darkest periods of my life. And to Karel, Pim and Quinten, for doing what friends do best, lighten up those dark periods with laughter and fun.

Summary

News media play a pivotal role in the functioning of democracies. They facilitate information exchange between elected officials and the public, have the capacity to mobilize social groups and can provide interpretation and context to the events that take place in the world around us. Considering these roles of news media, my goal with this thesis is to investigate how newspapers in Norway, The Netherlands, Denmark and the United Kingdom fulfill these roles, specifically by looking at how diverse and politically slanted their news coverage is. I do this during a period (2000-2020) in which the Internet is providing ever increasing competition for regular news media, such as newspapers. By utilizing computational text analysis methods, it is possible to consider every news article from the newspapers included in this study, and conduct analyses that span a long period of time. Through these analyses, I aim to contribute to the empirical knowledge on news diversity and political parallelism, and to investigate and improve upon the computational methods that are available to measure these concepts. Concretely, the contributions in this thesis are structured around three studies. The first is focused on the development/improvement of a method to generate sentiment dictionaries, so that it can be used for evaluating valence in political news articles. The second study investigates the diversity between newspapers in terms of lexical and valence diversity, while the third evaluates the presence of bias in the amount and valence of attention for specific political parties.

The results indicate a modest increase in news diversity, rather than the expected theory-based decline. Political parallelism either remains stable or slightly decreases, depending on the country. Both of these findings can be considered positive from a normative standpoint. Strong trends in news diversity might be detrimental to the roles that news media fulfill in democracies, resulting in either over-

or under-representation of specific issues. Similarly, newspaper readers are not exposed to increasingly slanted political news, as political parallelism remains stable. However, the amount of political parallelism is substantial, indicating that readers of different newspapers are exposed to different kinds of political slant. As for valence, the results show that there is no substantial difference between newspapers in the valence with which they discuss the same events, nor is there any political parallelism in the valence of political party coverage. The contribution of these empirical findings is twofold. First, the findings illustrate that logical and plausible theoretical assumptions do not always properly reflect the reality on which they are based. Hence, computational methods are shown to be capable of providing new and possibly surprising perspectives. The second contribution is then found in detailing the ways in which computational methods can contribute to the ongoing academic discussion on long-standing theories and assumptions.

Table of Contents

List of Figures

List of Tables

1 Introduction

The primary role of news media in democracies is to facilitate information exchange between elected officials and the public. Media inform politicians of what issues are at play in society, while at the same time they inform the public of which issues politicians prioritize and what positions they take to solve these issues. In this way the news media provide an arena for the exchange of information (van Aelst & Walgrave, 2016) and facilitate the functioning of a public sphere (e.g. Trenz, 2004). However, the impact of journalism on democracy extends beyond the reporting of factual information. News media can fulfill additional roles in society by investigating and analyzing the issues that they report on, but also by fostering social empathy and mobilizing specific groups (Schudson, 2014). Through these roles news media act as gatekeepers to and mediators of the public debate, and have the capacity to set and influence both the political and public agenda (McCombs & Shaw, 1972). Ideally, news media would use this influence to facilitate a well-informed public and well-informed politicians, but also serve as a mirror to society as a whole, by fulfilling all of the described roles.

In this thesis I investigate two specific aspects of media content, news diversity and political parallelism. News diversity concerns the extent to which media are presenting a variety of issues in a variety of ways, and through that facilitate a broad and inclusive democratic debate (Napoli, 1999; Van Cuilenburg, 2000). Political parallelism occurs when aspects of the news production process are influenced by partisan considerations (Hallin & Mancini, 2004; Seymour-Ure, 1974). This might take the form of structurally reporting more favorable (Kahn & Kenney, 2002; Larcinese et al., 2011) or more frequently on some parties than others.

Considering the rise of the Internet, the first two decades of the 21st century are of particular relevance to this thesis, as many aspects of both the production and dissemination of (political) news have changed during this period. By the time Internet became widely accessible in the early 2000s the amount of available news gradually but substantially increased. Due to this development especially newspapers have found themselves in strong decline as a primary source of news. In response, they have adjusted their business models and content in order to remain profitable (Curran, 2010; Siles & Boczkowski, 2012). It is in this highly dynamic media landscape that I investigate trends in news diversity and political parallelism.

The prominence and distribution of topics and political actors in the news are at the core of this investigation of news diversity and political parallelism. However, the way in which topics and actors are covered is relevant as well. Therefore, the concept of sentiment (or tone) is investigated in relation to both the topics and actors that news media discuss. Knowing how a topic or political party is discussed in terms of sentiment provides additional context. After all, discussing a topic in a positive or negative way can add to the diversity with which that topic is discussed (Joris et al., 2020). Similarly, political parallelism can take the form of negative attention towards parties which have a different political alignment than the news outlet just as well as it can result in positive attention for parties with which a news outlet aligns itself.

I analyze sentiment, topic diversity and the attention for specific political parties using novel computational text analysis methods. Such an approach provides a different perspective from regular inferential statistical approaches, as it facilitates the analysis of all data over an extended period of time (Boumans & Trilling, 2016) without the need for sampling. In addition, it simplifies and structures comparative research across several countries, and enables an exhaustive comparison between them. Because studies on such a scale have not often been conducted as of yet, my analyses focus primarily on describing trends and developments in news diversity and political parallelism in a thorough way, rather than drawing causal inferences relating to these trends. The theories I describe are therefore used to provide a context, but their causal mechanisms are assumed rather than explicitly tested.

1.1 Purpose

The main goals of this thesis are to investigate and improve existing computational text analysis methods for the analysis of news diversity and political parallelism, and to contribute to the empirical knowledge on news diversity and parallelism through a comparative and longitudinal study. To that end, the contributions of this thesis are structured along three main research questions:

1. *How can developments in computational text analysis be used in a valid and affordable way to measure topic diversity and news sentiment?*
2. *How diverse are newspapers in their selection of news topics, and how has the level of topic diversity developed during the early 21st century?*
3. *To what extent do newspapers show signs of political parallelism, and how has the level of political parallelism developed during the early 21st century?*

In line with these three questions I have conducted studies to 1) produce a model to computationally analyze sentiment, 2) analyze the diversity in topics that different newspapers discuss, and 3) analyze the relative attention that political parties receive in different newspapers. The answers to RQ2 and RQ3 provide an overview of the amount of content overlap between major daily newspapers, and the amount of political preference that is visible in that content. In combination with the comparative aspect, involving countries with different political and media systems, it is then possible to define how newspapers fulfill their role as information suppliers in Western democracies. At the same time, the results can also be generalized towards media organisations under economic pressure, considering the economically challenging environment for newspapers in the early 21st century. In doing so, my findings contribute to the further development of these theories, and illustrate the potential advantages and disadvantages of computational methods in investigating both news diversity and parallelism in particular, and news content in general.

1.2 Structure and contents

This thesis consists of six chapters. Following the introduction, chapter 2 presents the theoretical foundations of the thesis. This section contains a discussion of the concepts of news diversity and political parallelism, which are central to this thesis. It is followed by a discussion on the specifics of sentiment and its analysis as an important aspect of news content, and concluded with a discussion of professionalization and commercialization in news production. The theory overview is followed by a description of the research paradigm and design, as well as a description of the data (Chapter 3). In this section there is a strong focus on the computational perspective of research design, and the paradigmatic issues that arise from such a perspective. Chapter 4 then describes the used methods in detail, while chapter 5 presents the various findings. In chapter 6 conclusions are drawn and discussed based on the findings in the previous chapters. The thesis is concluded with a list of references, the complete articles (see Table 1) and other appendices.

Table 1: Overview of articles

Title:	**The Sentiment is in the Details** *A Language-agnostic Approach to Dictionary Expansion and Sentence-level Sentiment Analysis in News Media*
Journal:	*Computational Communication Research*
Status:	Published
Author(s):	Erik de Vries
Title:	**Telling a Different Story** *A Longitudinal Investigation of News Diversity in Four Countries*
Journal:	*Journalism Studies*
Status:	Published
Author(s):	Erik de Vries, Rens Vliegenthart & Stefaan Walgrave
Title:	**Newspaper Favorites?** *A Comparative Assessment of Political Parallelism Across Two Decades*
Journal:	*The International Journal of Press/Politics*
Status:	Under review
Author(s):	Erik de Vries & Gunnar Thesen

2 Theory

As emphasized in many studies (Beckers et al., 2021; Napoli, 1999; Sjøvaag, 2016; Vogler et al., 2020), the media fulfill a central role in the functioning of democracy by providing both citizens and politicians with information about each other and the issues that are at play in society. Changes in this supply of information thus also affect the functioning of the democratic system as a whole. The concepts of news diversity and political parallelism are important in this context because they focus on processes that influence the supply of information directly. Where news diversity concerns the breadth of the provided information, parallelism is about the potential political bias in the selection and presentation of that information by a newspaper. Together, analysis of these concepts provides a wide array of information on the content of political news. Both the variation in topics and the presence of political actors are covered in this thesis, as well as the valence with which both are presented.

In the following sections the theoretical framework of this thesis is laid out. Sections 2.1 through 2.3 discuss the main concepts of news diversity, political parallelism and sentiment analysis. In section 2.4 further theoretical background is provided regarding the (historical) trends in news production practices. Section 2.5 concludes with a summary of the conflicting expectations derived from the theories and empirical findings on the professionalization and commercialization of journalism.

2.1 News diversity

News diversity in the broadest sense is about the variation in information that news outlets provide to society, as is clearly visible in the statement that news diversity "... should always be compared with relevant variations in society and social reality." (Van

Cuilenburg, 1999, p. 199). This broad scope of news diversity is also recognized by Joris et al. (2020), who identify 43 different dimensions in their literature review of news diversity. As they argue for a increased use of more explicit terms, news diversity in this thesis refers specifically to the concepts of topic and sentiment/valence diversity. Additionally, the focus is on external diversity, between different outlets, rather than internal diversity (Hallin & Mancini, 2004; Joris et al., 2020; Sjøvaag, 2016). Normatively, news diversity is evaluated based on the assumption that too much diversity leads to audience fragmentation (Roessler, 2007) and too little focus on any specific issue, while too little diversity leads to too much focus on a few specific issues at the cost of other potentially important issues (Van Cuilenburg, 1999). Hence, it is difficult to pinpoint the optimal level of news diversity, other than that it is “somewhere in the middle”. Without going deeply into the normative question of what would be too much/little diversity, balance is of importance and strong deviations are a cause for concern.

While Joris et al. (2020) illustrate with their review that there are relatively many studies investigating topic diversity, only a small portion of these studies focus on developments over time (Boczkowski & Santos, 2007 in Argentina; Vogler et al., 2020 in Switzerland; Beckers et al., 2019 in Belgium), and none do so from a cross-national perspective or by including valence in addition to topic diversity. The operationalizations of diversity also vary largely between the studies, making it difficult to compare findings. It is in this research gap, of longitudinal comparative studies of news diversity, that the results presented in this thesis make a contribution. Explanatory theories and findings relating to news diversity are discussed in sections 2.4 (Professionalization and commercialization) and 2.5 (Conflicting expectations), because they also relate to political parallelism.

2.2 Political parallelism

Like news diversity, the presence or absence of political parallelism influences the supply of information available to citizens. But where news diversity concerns the general diversity in the supply of information, political parallelism concerns the political bias in that information. It originates from the concept of press-party parallelism developed in the 1970s by Seymour-Ure (1974). During the time of press-party parallelism, newspapers had strong organizational and ideological ties to a specific political party, resulting in biased and partisan reporting by individual outlets. As pointed out by Hallin & Mancini (2004), many European countries have had such a partisan media system, where individual outlets had strong ties with specific political parties. Strict press-party parallelism has however disappeared since the middle of the last century, in favor of a parallelism of "general political tendencies" (Hallin & Mancini, 2004, p. 27). In this more general definition of parallelism the focus is less on the organizational ties and partisanship of audiences, and more on the extent to which news outlets emphasize a distinct political orientation (rather than a specific party). Here, the parallelism in content is mirrored in journalistic norms and practices (Hallin & Mancini, 2004, pp. 28–29), which can be defined as the extent to which journalists adhere to either their (outlet's) partisanship or professional journalistic norms (Patterson & Donsbach, 1996).

While many have investigated political parallelism, for example through audience partisanship (van Kempen, 2007), most studies use data from the European Media Systems Survey (EMS) (Popescu et al., 2011) in their investigations of parallelism in news content (Lelkes, 2016; van Dalen et al., 2011) . The data in the EMS is based on national expert scoring of the extent to which newspapers advocate for or are influenced by specific parties and policies, and shows substantial levels of (variation in) political bias across Europe (Popescu et al., 2011). Other studies investigate parallelism directly in media content, through dimensions such as the relationship between party endorsement and negative coverage (US: Kahn & Kenney, 2002; Larcinese et al., 2011; UK: Brandenburg, 2006), or through increased attention to issues that are on the agenda of the parties their readers vote for (van der Pas et al., 2017).

A different – and surprisingly under-explored – aspect of content parallelism relates to biases in the volume of news attention to specific parties. This is relevant for a number of reasons. For example, quantity and quality of news attention influences party support (Hopmann et al., 2010). From the perspective of party cue effects (Bullock, 2011, 2020) news media fulfill an intermediary role in the transmission of such effects between parties and voters (Nordø, 2021). In a content analysis study conducted in Austria, Haselmayer et al. (2017) find that newspapers report more on press releases from parties that their readers identify with. Similarly, the focus here is on the proportionality of the amount of attention that different political parties receive, and the valence of the attention. And like with news diversity longitudinal and comparative studies into political parallelism are rare, emphasizing the contribution of the political parallelism findings presented here.

2.3 Sentiment analysis

In the literature on news diversity there is a strong focus on topic diversity (see Joris et al., 2020 for an overview), such as the study by Boczkowski & Santos (2007) in Argentina. Similarly, political parallelism is investigated through dimensions such as the events that news outlets cover (Hopmann et al., 2012). Knowing the amount of attention for specific events and variation in that attention between outlets however says little about the character of the attention. Characterization of attention in terms of sentiment has repeatedly been shown to influence the interpretation of news by the public.

Negative sentiment has been shown to have a stronger effect than positive news (de Vreese et al., 2011; Soroka & McAdams, 2015) and news sentiment in general has been shown to affect election results (Hopmann et al., 2010). Considering this it is no surprise that sentiment is frequently investigated in news content studies (see Boukes et al., 2020 for an overview). And while sentiment is a relevant aspect in some political parallelism studies (Brandenburg, 2006; Kahn & Kenney, 2002; Larcinese et al., 2011), it is rarely investigated in the context of news diversity (Joris et al., 2020).

Sentiment can generally be described as the 'attitude towards a particular target or topic' (Mohammad, 2016, p. 201). It indicates appraisals of a situation, expressed as either positive or negative evaluations, or as more specific discrete emotions representing the feelings one has. Analyzing either emotions or evaluations is however highly complex, considering the computational methods that were available at the time the studies were conducted. Evaluations in particular are a challenge to interpret (e.g. van Atteveldt et al., 2017), as illustrated by the sentence "I am sad that Hillary lost the presidential race". In this sentence the negative valence of the words "sad" and "lost" relates to the feelings of the author towards the situation. Linguistically "sad" and "lost" are however negative evaluations of "Hillary", even though the author has a clear positive stance towards Hillary Clinton (example from Aldayel & Magdy, 2021, p. 5).

At the core of evaluations such as the example in the previous paragraph is valence, the positive or negative connotation of words (e.g. de Vreese et al., 2011). A common way of analyzing sentiment as valence is through the use of sentiment dictionaries. When using a dictionary for sentiment analysis, there are two main aspects to consider: 1) the construction and content of the dictionary, and 2) the specific domain to which it is going to be applied. As the meaning of words changes between domains, sentiment dictionaries need to be domain-specific to some extent (Young & Soroka, 2012). Manually created dictionaries tend to work reasonably well when they are very domain-specific (Muddiman et al., 2019), but in more general applications performance drops significantly (Boukes et al., 2020). A balance between specificity and general applicability thus needs to be found when constructing a sentiment dictionary. The question then is whether computational methods can be used to optimize the procedure of dictionary construction and increase the performance of the dictionary.

2.4 Professionalization and commercialization

The theories discussed in this section are used to derive assumptions concerning trends in news diversity and political parallelism. These assumptions serve as context to the findings presented in this thesis. They are however not explicitly tested because the collection and analysis of data on the professionalization and commercialization of journalism is beyond the scope of the studies presented here.

The rise of media logic is closely related to the development from press-party parallelism (strong ties between specific political parties and media outlets) to political parallelism (no specific ties, but general political leaning in media outlets). Media logic indicates a system where the structure and norms of media rather than those of politics determine what becomes news (Altheide, 2013; Altheide & Snow, 1979; Brants & Van Praag, 2006). As a result, reporting is more professional and less partisan, and journalists increasingly rely on the use of shared journalistic norms (Esser, 2013; Esser & Umbricht, 2014; Patterson & Donsbach, 1996). Reliance on such norms has also led to an increasing importance of news values (Galtung & Ruge, 1965; Harcup & O'Neill, 2001, 2017). Attributes such as relevance and magnitude, which are inherent to an event, are evaluated for their newsworthiness by journalists. These evaluations differ between journalists and outlets, leading to news that is more similar in its adherence to professional norms, but does not result in outlets producing the exact same content. Or as Hallin & Mancini (2004, p. 26) formulate it: “no serious media analyst would argue that journalism anywhere in the world is literally neutral”. Rather, journalists consciously explain and interpret events, leading to interpretive journalism (Esser & Umbricht, 2014; Soontjens, 2019). Through these different interpretations of events, and different evaluations of news values, different representations emerge of what the news is (Hagar et al., 2021; Patterson & Donsbach, 1996).

Newspapers are of particular interest when considering the professionalization of journalism, because of their persistence throughout time. Many argue that the “golden age” of newspapers is long past, and that their importance (readership) has greatly diminished in both Liberal (UK: Lewis et al., 2008; US: Curran, 2010) and Democratic-Corporatist (Scandinavia: Allern & Blach-Ørsten, 2011; the Netherlands: Brants & Van Praag, 2006; Germany: Brüggemann et al., 2012; Switzerland: Vogler et al., 2020) media systems. This does however not exclusively occur with newspapers. In general, besides a shift from political to media logic, there has also been a shift in media markets from supply-driven to demand-driven (Brants & van Praag, 2017). Additionally, the merging of newsrooms and convergence in journalistic processes (Menke et al., 2018, 2019; Paulussen, 2012) leads to a situation where online and print news is increasingly produced by the same newsrooms. Arguably then, the decline of newspaper readership does not necessarily indicate a decreasing relevance of what specifically newspapers write. Rather, decreasing reach and increasing audience fragmentation are felt across entire media markets, and lead to (the content of) any individual news outlet becoming less relevant. That makes Curran’s (2010) “newspaper crisis” more a “news media crisis”.

Increasing commercial pressures however do have a profound effect on newspaper content. Specifically, the increasing competition in shrinking newspaper markets (Curran, 2010), due to the audience switching to other news sources such as the Internet, have put the profitability of newspapers across Europe under pressure (Brüggemann et al., 2012; Lewis et al., 2008; Vogler et al., 2020). Decreasing profitability has resulted in staff cuts (Curran, 2010) and mergers between publishing houses and newsrooms (Picard, 2014). Both developments have increased the workload of the remaining journalists, who are expected to produce content for both online and print outlets. As a result, more "desk work" and less time to go out and gather own information (Paulussen, 2012) leads to an increased reliance on content produced by external sources ("information subsidies": Gandy, 1980), such as PR material, press releases and content produced by news agencies (Boumans et al., 2018; Vogler et al., 2020). But journalists also increasingly rely on the products of other journalists (Boczkowski, 2009).

2.5 Conflicting expectations

The theories on the professionalization and commercialization of journalism lead to conflicting assumptions regarding news diversity and political parallelism. From a professionalization perspective, shared norms and the use of news values lead to the expectation of a decrease in both news diversity and parallelism. From a commercialization perspective, things are a bit more nuanced. While external sources provide newsrooms with a cost-effective option to produce news, the effects on news diversity and parallelism depend on which sources are used. And the pressure to retain audience share drives publishing houses and newsrooms to develop content aimed at a specific part of the audience. This can take form through the events are covered (news diversity) or from which political perspective (parallelism) they are interpreted. Seen from this perspective is the rise of interpretive journalism (Esser & Umbricht, 2014) perhaps not so much a trend in the professionalization of journalism as it is in the commercialization of journalism.

Empirical findings support these conflicting theoretical expectations towards news diversity and parallelism. Several studies show that commercialization decreases news diversity through newsroom mergers (Beckers et al., 2019; Dailey et al., 2005; Hendrickx & Ranaivoson, 2019) and increases the reliance on external sources and other journalists (e.g. Boumans et al., 2018). These findings are corroborated by longitudinal studies in Argentina (Boczkowski & Santos, 2007) and Switzerland (Vogler et al., 2020). Other studies however find an increase of news diversity under competitive market conditions as outlets start to invest in their content and quality (Lacy & Simon, 1993) in order to "capture certain subsets of news readers" (Hagar et al., 2021, p. 4). And in a recent Belgian study (Beckers et al., 2019) no decline in news diversity is found.

With regards to political parallelism, various studies have found copious but differing amounts of parallelism and partisan bias in European media outlets (Popescu et al., 2011), such as in Spanish (Baumgartner & Bonafont, 2015) and British (Brandenburg, 2006) newspapers. Clear longitudinal findings are however lacking. Arguably, both the increased reliance on external (political) sources and the struggle to retain audience share leads newspapers to develop a profile that emphasizes a specific political preference. This would be consistent with the cross-sectional findings in Spain and the UK. On the other hand, shared professional norms (notably factual and objective reporting) and the use of news values might decrease political parallelism over time.

Looking more broadly, professionalization decreases both news diversity and political parallelism, while commercialization has the potential to counteract this decrease. Newspapers have a commercial incentive to differentiate themselves from each other, but their ability to do so depends on the financial resources and staff available to them. And while the various theories regarding professionalization and commercialization are not tested explicitly in this thesis, investigating trends in news diversity and political parallelism can shed some light on which of the theoretical assumptions seem most likely. Because the abundance of conflicting theoretical expectations and empirical findings begs the question how news diversity and political parallelism have actually developed in European newspapers during the first two decades of the 21st century, and to what extent this can be attributed to structural developments in newspaper markets and journalism more broadly in various European democracies.

3 Design & Data

In this chapter, a concise philosophical discussion is presented on (post-)positivism and the importance of hermeneutics and interpretation. Following this discussion, section 3.2 elaborates on the combined research design of the three studies (Table 1), while section 3.3 describes the sample selection process of both the countries and newspapers included in the studies.

3.1 Research paradigm

Investigating the theoretical concepts of this thesis, news diversity and political parallelism, through computational methods implies a positivist ontological perspective. By comparing the content of newspaper articles to determine the diversity between them, there is the implicit assumption that there exists an objective social reality, that can be observed in and represented by news. More generally, diversity can only be analyzed in a quantitative fashion if there is a confined reference frame (objective social reality) in which to evaluate the level of diversity. This is also represented in the theories on news production, the functioning of media markets and political parallelism used in this thesis, which all presume an objectively observable social reality. But while the ontological perspective is quite clear, the epistemological perspective is less well-defined and more relevant to discuss. As the units of analysis are literally texts (newspaper articles), the concept of hermeneutics applies in particular. Taylor (1971) argues that hermeneutics can be used to study objects that have meaning independent of the subject constructing (the author) or observing (the reader) the object. The question then is *where* this meaning resides.

Considering newspaper articles, or any text containing natural language, meaning resides at a number of different levels. Individual words contain meaning, while the sentences that they form contain additional meaning. Paragraphs constitute another level of meaning, with the final level in this thesis being the full text. To observe the meaning of an entire article, the hermeneutic cycle consists of evaluating the meaning contained in each of these levels independently, and re-evaluate them given the interpretations at the other levels. In that sense it is ironic to combine Taylor's (1971) hermeneutic cycle with computational text analysis, as he explicitly argues against empiricism and the use of "brute data", while computational text analysis can in many ways be seen as the pinnacle of both. Others warn against brute data in different ways. Computational analysis methods can for example give a false impression of objective observation, free from human error (Sætra, 2018). Free from error does however not imply free from values. And even though computational analysis methods are often value-laden, they are so complex that human interpretation of the outcomes (and the process that created them) is often near-impossible. Incidentally, interpretation is also where "humans do things computers can't" (Sætra, 2018, p. 520).

This is illustrated by the fact that interpretation is conditional on a separation between meaning and its expression. For an object to have meaning, it needs a subject to have meaning *for* (Taylor, 1971, p. 4). Meaning then resides in the subject rather than the message (expression), and the subject needs to have some form of understanding of the context of a "text" to ascribe it meaning. If interpretation depends on understanding, the question arises if a computer can have the capacity to understand (and interpret) the data it is processing. The Chinese room thought experiment (Searle, 1980) applies specifically to this question, as it is about the difference between applying formal rules versus understanding. A person not understanding Chinese is capable of translating Chinese questions and providing Chinese answers to them as long as they have rules to translate. However, these rules do not guarantee that a sensible answer to a question is produced. For that, an understanding of the question (and the answer) beyond the formal rules is required. In the case of computers, "they have only a syntax but no semantics" (Searle, 1980, p. 422). Thus computers are able to deal with expressions, but not with meaning. Algorithms then can classify newspaper articles based on the expressions contained within, but not based on the meaning of the article. Essentially, the "brute data" is transformed from one format (word counts) to another (classifications of an article). In this process a lot of the original data is reduced to a much smaller amount. This is desirable, as the original data cannot be analyzed manually. But it also has drawbacks, as large amounts of information and context are lost in the transformation.

Ultimately then, due to the non-interpretation of algorithms, the burden of interpretation falls on the researcher analyzing the output of the computer. As Doshi-Velez & Kim (2017) argue, interpretability is an essential aspect of model evaluation besides traditional performance measures such as model accuracy. The need for such interpretability arises from the fact that, regardless of the model, there will always be an "*incompleteness* in the problem formalization" (Doshi-Velez & Kim, 2017, p. 3). In essence it is the same problem that arises in the Chinese room experiment: When the formal rules are not exhaustive there can be no true interpretation. Hence any computational text analysis model should provide information to explain its output in addition to the output itself. If these explanations can be used without any additional information by humans to come to the same conclusions as the model, the model provides adequate interpretability. That makes interpretability a concept that is hard, if not impossible to quantify, and rather something that "you'll know when you see it" (Doshi-Velez & Kim, 2017, p. 1).

The use of computational text analysis methods is always a tradeoff between reducing the amount of data to a more manageable size and retaining as much interpretability as possible. Specifically, information that is relevant for the research question being investigated should remain interpretable to the largest extent possible. This interpretability is relevant both directly for answering the question and indirectly as a means to evaluate the validity and reliability of the used method(s). Algorithms should therefore provide ample room for interpretation of both their functioning and output. However, a computer does not have a mind, it does not think (Searle, 1980). As such, it cannot in itself pose a threat to meaningful science. The only threat to that are we ourselves, the researchers that choose to focus our attention either on the brute (and meaningless) side of the data, or on the more substantive and qualitative interpretation of it.

Considering the radical importance of interpretation when using computational text analysis methods, the ontological perspective in this thesis might be better described as realist post-positivist rather than purely positivist. Because in the acknowledgement of the importance of interpretation is an implicit assumption that "our tools (human understanding and interpretation) are inevitably value-laden, theory-laden and context-dependent" (Fox, 2008, p. 7).

3.2 Research design

The strong emphasis on computational methods in the research paradigm section combined with a focus on news content logically leads to research design using computational content analysis methods. The choice between a manual or computational content analysis is one of preference and practical considerations. When it comes to practical considerations, computational methods enable the analysis of larger amounts of data than manual analysis (Boumans & Trilling, 2016; Grimmer & Stewart, 2013). The choice for a longitudinal (20 years of newspaper articles) and comparative (involving four distinct countries and languages) study involves large amounts of text data. In such cases in particular, computational methods provide a suitable means of analysis. But besides offering a suitable means, the use of computational methods also provides the opportunity to investigate and improve the application of such methods, both in general and specifically for the analysis of news diversity and political parallelism. The comparative aspect of the studies additionally provides an opportunity for evaluating computational methods in the context of multilingual text data. While computational methods have some disadvantages, as described in the previous section, they also provide a number of advantages over manual content analysis: 1) They remove the need for sampling of individual units/articles, which can facilitate 2) very detailed longitudinal research and 3) leads to a data set that can be utilized to investigate many different research questions. In general, computational text analysis methods provide very detailed descriptive information. With such information existing theoretical assumptions can be tested in a different way from what would be possible with manual content analysis (Boumans & Trilling, 2016). In this thesis computational methods for example facilitate the analysis of trends over longer time periods (20 years), outlets (3) and countries (4) than would under normal circumstances be possible with a manual quantitative content analysis. Computational analyses are thus not necessarily “better” or “worse”, but rather provide a different kind of knowledge focused strongly on empirical observation. Considering the examples mentioned before, the research design of this thesis can best be formulated as a longitudinal, time-series cross-sectional, comparative computational content analysis.

3.3 Data

The selection of countries used in this thesis is primarily of relevance because of their differing media and political systems. With regards to Hallin and Mancini's (2004) media systems the major difference is between the Liberal system in the UK and the Democratic Corporatist system in the other three countries. Most importantly the UK system has had an early start when it comes to the professionalization and commercialization of the press, as illustrated by the early rise and continuing presence of commercially oriented tabloid journalism (Esser, 1999). In the other countries the press has historically been strongly linked to the party system, with individual newspapers being mouthpieces of their respective parties. Professionalization and commercialization started later than in the UK, but these differences are becoming smaller over time (Denmark: Esmark & Ørsten, 2008; Netherlands: Brants & Van Praag, 2006; Norway: Østbye & Aalberg, 2008). As for the political systems, there is again a difference between the majoritarian/"Westminster" model in the UK and the consensus model in the other countries (Lijphart, 2012). Essentially, the UK has a two-party system where either of the parties has an absolute majority, where in the other countries coalitions between parties are often required to form a government.

Table 2: Newspaper sample

	Left-wing	**Right-wing**	**Tabloid**	**Total articles**[1]
Denmark	Politiken	Jyllands-Posten	Ekstra Bladet	2.10
Netherlands[2]	De Volkskrant	NRC Handelsblad	De Telegraaf	2.18
Norway	[3]	Aftenposten	VG/ Dagbladet	2.28
United Kingdom[2]	The Guardian	The Daily Telegraph[4]	The Sun	5.12

Note: [1]In millions; [2]Until December 2018; [3]Due to lack of suitable data, Dagbladet as substitution; [4]From January 2001

The static concept of systems does however not cover the dynamics over 20 years of country-specific developments. Different trends and contexts in each country are likely linked to differences in terms of news diversity and parallelism. Even though Denmark, Norway and The Netherlands can be grouped together based on their political and media system when comparing them to the UK, there are also relevant differences between them. The Netherlands in particular has experienced a trend towards an increasingly fragmented parliamentary landscape, with new (mostly populist/right-wing) parties forming throughout the period. In both Norway and Denmark these trends are also visible, but not as pronounced as in The Netherlands. These dynamics, along with the political structures and historical ties between parties and newspapers are of particular relevance when considering political parallelism. In addition, the Brexit campaign and implementation of the results are likely to have had substantial impact on news content in the UK. Table 2[1] shows the selection of

[1] The number of articles in the “Sentiment is in the Details” paper is incorrect.

newspapers for each country.

The newspapers included in this thesis are chosen with the goal of constructing a sample that is representative of the mainstream newspaper market in the four countries, and allows for cross-national comparisons. To this end, three newspapers are selected in each country, a left-leaning broadsheet newspaper, a right-leaning broadsheet newspaper, and a tabloid newspaper. These are selected based on the selection procedure used by de Vreese et al. (2016), with the exception of Dagbladet in Norway due to data availability issues for Dagsavisen[2]. This exception in Norway also leads to the exclusion of Norway in the political parallelism study, as there is no data available from a national left-wing broadsheet.

In the sample of newspapers, the most relevant difference between the countries is the ownership structure. In Denmark, all of the newspapers are owned by the same publishing house (JP/Politikens Hus) since 2003. In both Norway (VG, Aftenposten) and The Netherlands (Telegraaf, NRC, since 2015) the right-wing broadsheet and tabloid newspapers are owned by the same publishing house. In the UK, ownership of the three newspapers is totally separate. As argued in chapter 2, ownership concentration can in particular influence news diversity when content is shared between different outlets. Hence, while ownership structures are not explicitly investigated in this thesis, they are a likely explanatory factor of news diversity.

The data used in this thesis (see Table 2) is collected from national newspaper archives in each country. Generally speaking, this data is not ready to be used in computational analyses straight away. Issues that in this case arose are 1) the presence of duplicate articles due to various newspaper editions being included in the data, 2) the presence of articles that are merely excerpts, referring to the full article somewhere else in the newspaper, 3) non-natural lan-

[2] The data from Dagsavisen is formatted page by page, instead of article by article, making it impossible to analyze individual articles.

guage text, such as chess results with coordinates, weather reports, television guides and answers to crossword puzzles and 4) articles about culture/sports/entertainment that are irrelevant from a political/democratic perspective. In all cases, the solution is to remove this noise from the data.

4 Methods

This chapter provides an overview of the methods used in each of the studies described in Table 1. Section 4.1 starts with a discussion of the general concepts of computational text analysis, such as bag-of-words, vector space models, data cleaning and feature weighting. The steps used for cleaning the text data are described in section 4.2, while in section 4.3 the dictionary expansion method used for the sentiment analysis is described. Sections 4.4 and 4.5 describe the construction of the news diversity and political parallelism measures respectively.

4.1 General concepts

Computationally analyzing texts comes down to reducing natural language to numbers. The most straightforward way of doing so is by counting the words occurring in a text. In this so-called "bag-of-words" approach word order and syntax are discarded and the word counts are used as the input for various methods (see Boumans & Trilling, 2016 for an overview). When word order and syntax are relevant, a vector space or "word embedding" model can be used in combination with the raw word counts (Mikolov et al., 2013; see Almeida & Xexéo, 2019 for an overview). These models are based on the assumption formulated by Firth (1957) that "a word is known for the company it keeps", and constitute a multi-dimensional vector space in which each word is positioned based on its co-occurrences with other words. The assumption is that (combinations of) the dimensions in this model represent different latent aspects of meaning (Mikolov et al., 2013), implying that words that are closer together on these dimensions share some of their meaning with neighboring words.

Word embeddings thus provide a way to account for context, by modelling words as related entities. However, the context of words is also dependent on the units of text that are analyzed. This strongly relates to the concept of hermeneutics (different levels of meaning) as described in section 3.1. When considering news diversity and parallelism, the lowest unit of analysis that provides meaningful information about these concepts is words in sentences. While it is generally possible to aggregate information that is computationally derived from a unit of text to a higher level (e.g. paragraph or article), it is not possible to disaggregate to a lower level. Using the lowest relevant level of meaning as the base unit is therefore important when conducting computational text analysis. In this thesis, expression and arguably meaning are derived computationally at both the word level (word embeddings) and the sentence level. Manually, meaning is derived and inferred at the sentence level through human validation of the computational methods. Because this validation is conducted at sentence level, sentences are used as the base unit of analysis throughout the thesis.

Regardless of how texts are converted into numbers (bag-of-words, word embeddings), they can be cleaned using various processes to increase information density and remove unwanted noise from the data. While there are many possible ways to achieve this (see Denny & Spirling, 2018), they generally serve either the purpose of 1) reducing the amount of noise/uninformative text or 2) grouping together inflected word forms. The former can be achieved using term frequency-inverse document frequency (tf-idf) weighting of the word counts. This statistical procedure developed by Jones (1972) weighs word counts by their relative occurrence in all texts combined (the corpus). Hence, values increase when a word occurs frequently in a specific document but infrequently in the corpus, while it decreases when a word occurs infrequently in a document and frequently in the corpus.

To group inflected words together, either stemming (see Denny &

Spirling, 2018) or lemmatization can be used. Stemming is based on general heuristic rules that cut off the end of an inflected word to reduce it to its “stem”. As no proper linguistic argumentation underlies this process, the success depends on the specificity of the rules applied as well as the structure of the language it is applied to. Lemmatization on the other hand is a procedure that is part of Natural Language Processing (NLP), and reduces words to their dictionary lemmas using computational models that do account for linguistic syntax/rules. That makes NLP preferable over stemming, but it is also a computationally more demanding and complex procedure. NLP however also performs different tasks, such as identifying sentence borders, labeling individual words in a sentence through Universal Part-Of-Speech (UPOS) tags and identifying dependency relations between words in a sentence. Specifically the UPOS tags are relevant because they allow disambiguation of words that are spelled exactly the same, but have a different meaning (e.g. “evening” or “entrance” as either a noun or a verb). To perform NLP, the UDPipe package (Straka & Straková, 2017) is used in combination with version 2.3 of the Danish DDT, Dutch Alpino, English EWT and Norwegian Bokmål Universal Dependencies Models (Nivre et al., 2018).

4.2 Data cleaning

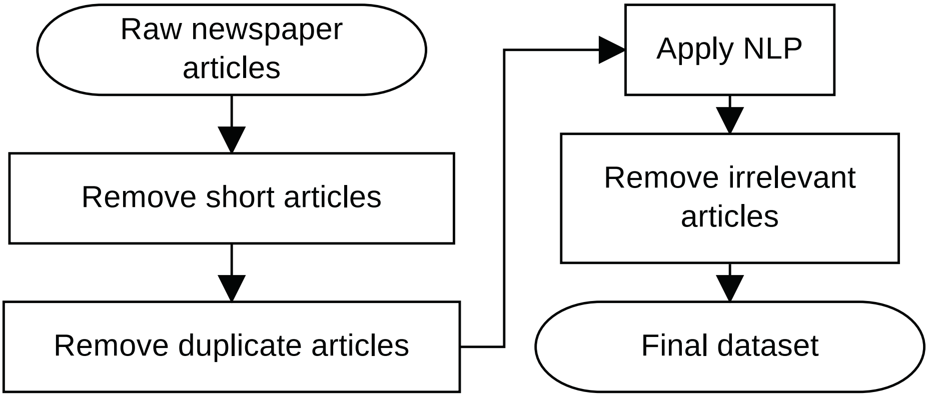


Figure 1: Data cleaning flowchart

As with any statistical procedure, results in computational text analysis depend on the quality of input data used. Because the quality of the content provided by newspaper archives is low (see section 3.3), the data is cleaned according to the steps presented in Figure 1. As a first step, all articles with a length of 30 words or less are filtered out, as they are too short to contain much, if any, meaningful content. This happens before the articles are imported into Elastic (formerly known as ElasticSearch), which is open-source software that makes the articles searchable, and in this case is also used as the database to store the articles. After adding the articles to Elastic, duplicate articles are filtered out by comparing articles published on the same day and in the same newspaper, based on the first 300 words of each article. When the cosine similarity between articles is .85 or higher, only one of these articles is kept in the database.

The remaining articles are then processed using NLP, followed by the removal of irrelevant articles that do not contain any politically relevant news content. These articles are filtered out using a Multinomial Naive Bayes model trained on articles coded by student assistants as either relevant or irrelevant. Intercoder reliability for the human coders ranges between a Krippendorff's alpha of .81 and .86 (for full details, see Appendix 8.4, Table 2.2). The input used for the Naive Bayes model is a combination of the lemmas and UPOS tags produced by the NLP procedure, with the counts weighted using tf-idf. The best performing model for each country is selected using a 3 by 5 nested cross-validation procedure, where one part of the data is for either optimization or evaluation of the model (hence nested) while the other parts are used for model training. The final models achieve a precision of between 0.87 (DK) and 0.94 (UK), with the level of precision being an indicator to what extent *only* irrelevant articles are removed.

4.3 Sentiment analysis

News tone, or sentiment, is measured using a method specifically developed for this study, based on an approach developed by Rheault et al. (2016). This method uses GloVe (Pennington et al., 2014) word embedding (WE) models, one per language, to construct custom sentiment dictionaries (lists of words). I expand upon this method by using a more sophisticated way of selecting the words to include in the dictionaries, and by applying them to individual sentences, rather than articles as a whole.

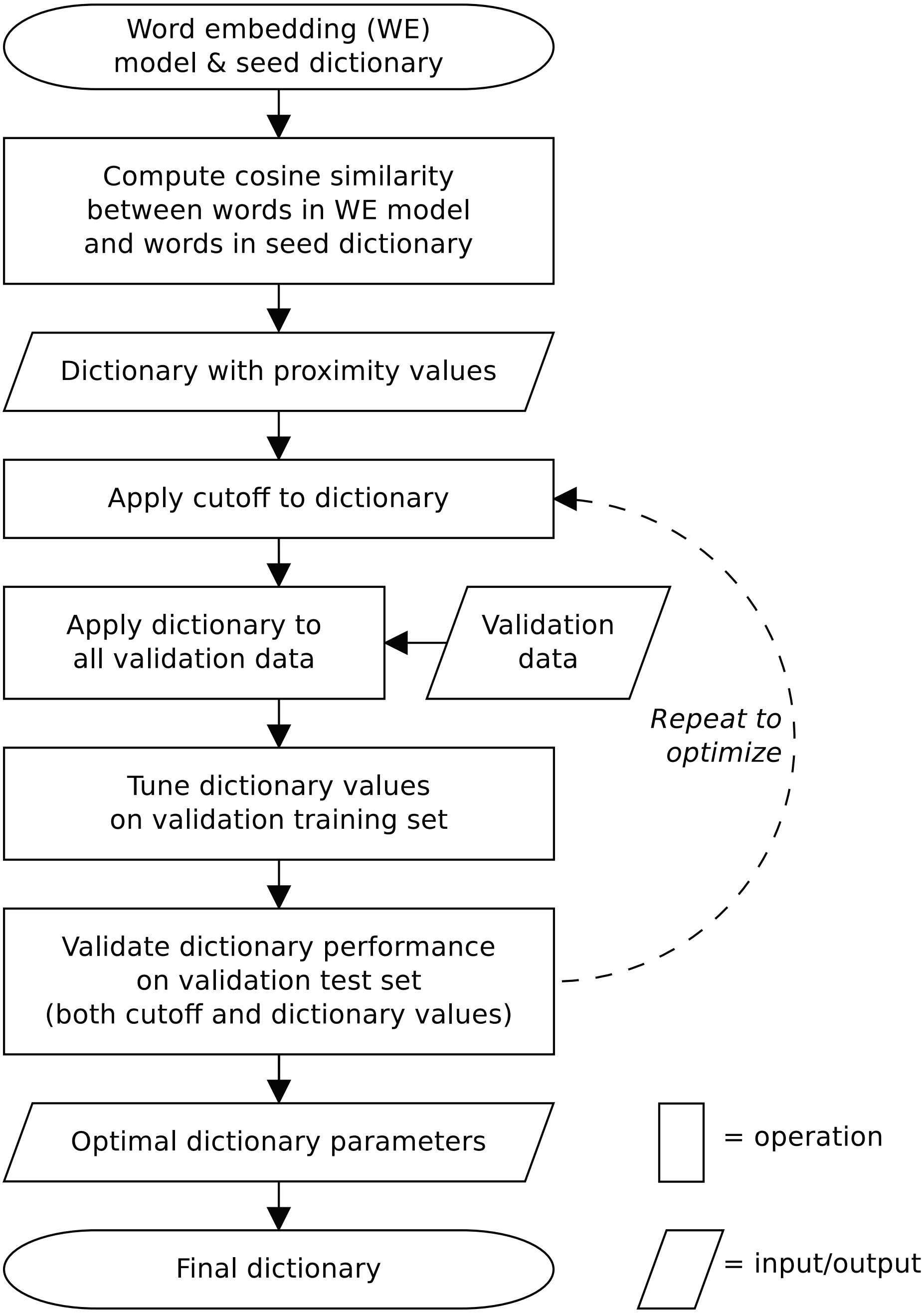

Word embedding (WE) model & seed dictionary
Compute cosine similarity between words in WE model and words in seed dictionary
Dictionary with proximity values
Apply cutoff to dictionary
Apply dictionary to all validation data
Validation data
Repeat to optimize
Tune dictionary values on validation training set
Validate dictionary performance on validation test set (both cutoff and dictionary values)
Optimal dictionary parameters
Final dictionary
= operation
= input/output

Figure 2: Sentiment dictionary construction flowchart

The steps involved in creating the sentiment dictionary are visualized in Figure 2, and described here[3]. A WE model is constructed for each language, using the articles (lemma-UPOS pairs) that remain after the data cleaning steps. The models are then combined with a list of unambiguously positive and negative seed words, the same as those used by Rheault et al. (2016). For languages other than English, these word lists are literally translated. Using the WE model, the initial list of positive/negative seed words is expanded by selecting words that are close to the seed words in the vector space (i.e. words that have similar values on the different dimensions in the model). The assumption is that these words should carry a positive or negative meaning, like the words in the seed dictionary. This meaning is assumed to be stronger when the average distance (in terms of cosine similarity) between a word and the positive or negative seed words decreases.

Cutoff values, based on human-coded validation sentences, are used to select the words that are included in the final sentiment dictionary. These sentences are coded as either positive, neutral or negative, based solely on the connotation of the words in the sentence. The hand-coded sentences are also used to fine-tune the conversion of the (continuous) sentiment scores for each sentence to ordinal (positive, neutral, negative) values. Using a 5-fold cross-validation approach, the performance of the WE dictionaries is estimated, relative to the human coding.

To get an indication of the relative performance of the WE dictionaries, the same procedure to convert continuous scores to categories and estimate performance is also applied to the Polyglot sentiment dictionaries (Al-Rfou et al., 2013; Chen & Skiena, 2014). These dictionaries are chosen as a comparison because they are available in

[3] The description of the sentiment analysis method provided here is a concise summary of the process that omits some of the nuances that are present in the original paper.

many languages and have been demonstrated to perform well when analyzing media content (Boukes et al., 2020).

As a final test, the sentiment scores from the WE dictionaries are aggregated to construct article-level sentiment scores. These article scores are used to test the predictive validity of the dictionaries. Their capability to detect the well-documented negativity bias in political news (see Lengauer et al., 2012 for an overview) then demonstrates the predictive validity of the dictionaries.

4.4 Topic diversity

To get an indication of news diversity two aspects are measured, topic diversity and sentiment diversity. Topic diversity is operationalized through a combination of cosine similarity with tf-idf weighted features. Boumans (2016) uses the same procedure for a similar task, detecting similarity between newspaper articles, press releases and PR material. Relative to other similarity metrics, such as Jaccard similarity in combination with tri-grams (counting groups of three words instead of individual words) (Vogler et al., 2020) the use of cosine similarity with tf-idf is less dichotomous, making it more a measure of similarity rather than an indicator of article pairs that are (near-) identical in their content.

The cosine similarity metric is computed for all combinations of newspaper articles published within a country on the same day, discarding comparisons of articles within the same newspaper. What remains are the similarity scores of one article in newspaper A with all articles in newspapers B and C. The maximum scores for B and C indicate the most similar article pairs (i.e. the articles in B and C that are most like the article in A). Inversion of those scores then results in a diversity measure. Using cutoff values derived through human validation, the scores are converted to a binary value indicating whether or not two articles are about the same topic. The weighted percentage of article pairs that have the same topic is then used as the final indicator of news diversity.

The sentiment diversity measure is based on the sentiment metric described in the previous section, and only used for article pairs that are about the same topic. The sentiment score of each article is subtracted from the other, after which the absolute result is used as diversity indicator. In this way, the general sentiment of the two articles is compared to one another. This provides information on whether that topic is discussed in a similar way in the two articles, as these articles are presumably about the same topic. While such a measure is unable to account for nuances relating to the specific source and/or target of the tone, the degree of congruence in coverage across outlets provides information about similarity, and thus of the level of external diversity. When the sentiment differs between two articles about the same topic, this is an indicator of the diversity of the context within which information on this topic is provided. More specifically, words with different connotations imply that articles with the same topical focus still say something different, and thus add different information and interpretation to the news supply.

4.5 Political parallelism

To investigate political parallelism in newspapers the amount and sentiment of attention for political parties is measured and combined with a party-newspaper alignment variable. Case-sensitive queries for either the full party name or the most commonly used party abbreviations are used to measure the attention for political parties in news articles. Where necessary special characters like opening and closing brackets for the abbreviations (con) and (lab) in the UK are also taken into account. In Norway, several of the major political parties have single letter abbreviations. In these specific cases regular expression filters are used to filter out common mistakes, such as the letter V being used both as a roman numeral and as abbreviation for the left-wing party Venstre. Besides party names, party attention is also measured through the attention for individual politicians (ministers, party leaders and MPs). In this case queries consist of the (first) given name and surname of politicians in close proximity (5 words). For cabinet members the queries are extended by including formal titles as alternative for given names (e.g. Secretary Johnson and Boris Johnson).

Party attention (both individuals and party names) is measured per sentence. Each sentence can contain a reference to a political party only once, in order to avoid double counts when a politician's name is mentioned along with a party abbreviation. The sentence counts are aggregated by party and newspaper to construct a relative indicator of the monthly attention each party receives in a newspaper. Sentiment is aggregated in a similar way - for all sentences mentioning a party - to construct an indicator of the sentiment context in which a political party is mentioned.

The presence of parallelism is evaluated based on the effect of overlap in political leaning between a newspaper and a party on the amount and tone of coverage that party receives. Hence, politicial parallelism is operationally defined as the effect of party-newspaper alignment on the amount of political bias (in amount and valence of attention) towards a specific party. The ties between parties and newspapers are constructed using a proxy based on the political leaning of each newspaper and the left-right placement of parties (based on the Chapel Hill Expert Survey, see Bakker et al., 2015).

The categorization of party-newspaper pairs as presented in Table 3 is based on two assumptions. First, newspaper leanings are interpreted as moderate, because of their mainstream nature. Second, alignment is only assumed for the traditionally dominant mainstream left/right party. This results in social-democratic/labour and liberal/conservative party-newspaper pairs, which are comparable across countries.[4]

Table 3: Party-newspaper alignment

	Left-wing alignment	**Right-wing alignment**
Denmark	Politiken *Social Democrats (S)*	Jyllands-Posten *Liberal Party (V)*
Netherlands	de Volkskrant *Labour Party (PvdA)*	NRC Handelsblad *People's Party For Freedom and Democracy (VVD)*
United Kingdom	The Guardian *Labour Party (Lab)*	The Daily Telegraph *Conservative Party (Con)*

In addition to attention, s, and party-newspaper alignment, several variables from the ParlGov data set (Döring & Manow, 2021) are used as control variables. These variables are included to account for aspects that influence the inherent newsworthiness of a party (i.e. news values). Specifically, party size (as vote share) abd dummies for government parties and the prime minister party are used as indicators of the news values “power elite” and “relevance” (Harcup & O’Neill, 2017). A control variable for the news values “surprise” and “conflict” is also included, using the ParlGov ideology scales to construct an indicator of ideological party extremity. To test developments over time, a running counter of months is used, interacted with the alignment variable. Because the temporal trend is not necessarily linear, both a first and second degree polynomial function of the month counter are included. The data are modeled using multi-

[4] Norway is excluded from this study because of data availability issues with the left-leaning broadsheet, see section 3.3.

level regression with random intercepts across parties, to deal with the correlated errors that are produced by the clustered data (Gill & Womack, 2013).

5 Results

In this chapter the empirical findings of the three studies (see Table 1) are presented. In section 5.1 the predictive and concurrent validity of the sentiment analysis method are discussed, as well as the concurrent validity of the news diversity measure. The relative presence of and trends in news diversity are discussed in section 5.2, while the same is discussed for political parallelism in section 5.3.

5.1 Computational text analysis

Three tests are conducted to determine the validity of the sentiment analysis method. The concurrent validity is tested by comparing the computationally generated sentiment scores to human classification of the same sentences. Intercoder reliability (Krippendorff's alpha) for the human-coded sentences ranges between .71 and .84, based on an intercoder reliablity test of 50 sentences coded by both the principal researcher in a country and their student assistant(s). The predictive performance is illustrated by detecting the negativity bias in news, while relative performance is assessed by comparison to another sentiment dictionary (Polyglot). Table 4 shows the performance of both the Polyglot and WE dictionaries when compared to human coding. The averages shown are weighted based on the relative occurrence of each of the three categories (negative, neutral, positive) in the hand-coded data set. Two things are visible in these results. Firstly, there is a clear balance between precision and recall. Which means that it is as good in predicting a sentence as positive/negative/neutral *only* when it should as it is in *always* predicting a sentence correctly as positive/negative/neutral. Secondly, the relative performance of the WE dictionaries is high, as they consistently outperform the Polyglot dictionaries.

Table 4: Weighted average performance for Polyglot and word embedding dictionaries

	F1		**Precision**		**Recall**	
	P [1]	*WE* [2]	*P* [1]	*WE* [2]	*P* [1]	*WE* [2]
Danish	0.48	0.62	0.54	0.62	0.46	0.63
Dutch	0.34	0.64	0.53	0.63	0.32	0.66
English	0.56	0.61	0.58	0.62	0.55	0.61
Norwegian	0.48	0.61	0.52	0.62	0.47	0.63

Note: [1]Polyglot; [2]Word Embedding

The same conclusion can be drawn based on the results presented in Figure 3. This figure illustrates for each language the distribution of errors, where the values indicate the difference from the true value. Hence, +2 indicates a positive prediction while the sentence is negative, while -1 indicates a negative prediction while the sentence is neutral and the 0 category shows the accuracy. What can clearly be seen is the higher accuracy of the WE dictionaries compared to Polyglot, but also the more uniform distribution of errors. Even with a modest absolute performance of the WE dictionaries (all metrics around .6), the normally distributed errors cancel each other out when aggregating. The absolute performance at sentence level is thus a conservative estimate when analyzing the sentiment of multiple sentences, or entire articles. So especially when aggregating, the WE dictionaries provide a valid way of measuring sentiment.

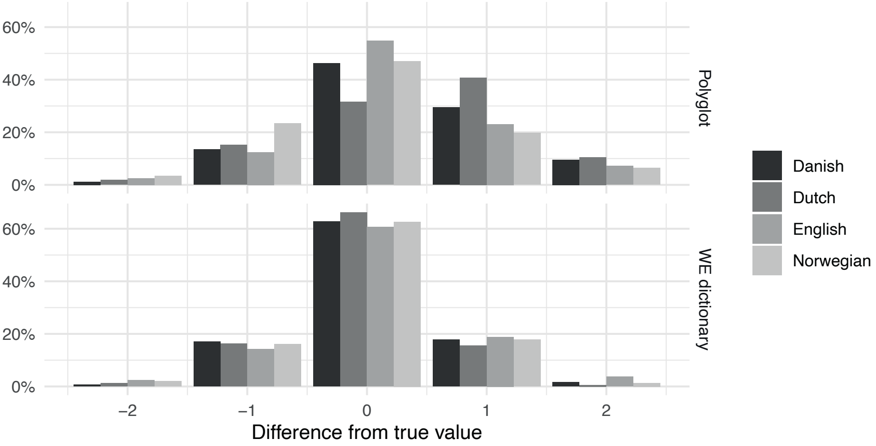


Figure 3: Difference between predicted and true sentiment category

The adequate performance of the WE dictionaries at the aggregate level is also illustrated in Figure 4, illustrating the presence of a negativity bias in political news. In this figure trends lines are shown for the sentiment in articles from broadsheet and tabloid newspapers respectively. The negativity bias itself is evident from the fact that all sentiment values (on the y-axis) are negative. At the same time, the tabloids are in general significantly more negative than the broadsheets. In the UK the difference between the tabloid and the broadsheet newspapers is particularly pronounced, which corresponds to the strong tabloid profile of The Sun when compared to the tabloids in the other countries. The findings of a general negativity bias and stronger negativity in tabloids versus broadsheets corresponds to theoretical expectations, lending further support to the validity of the WE dictionaries.

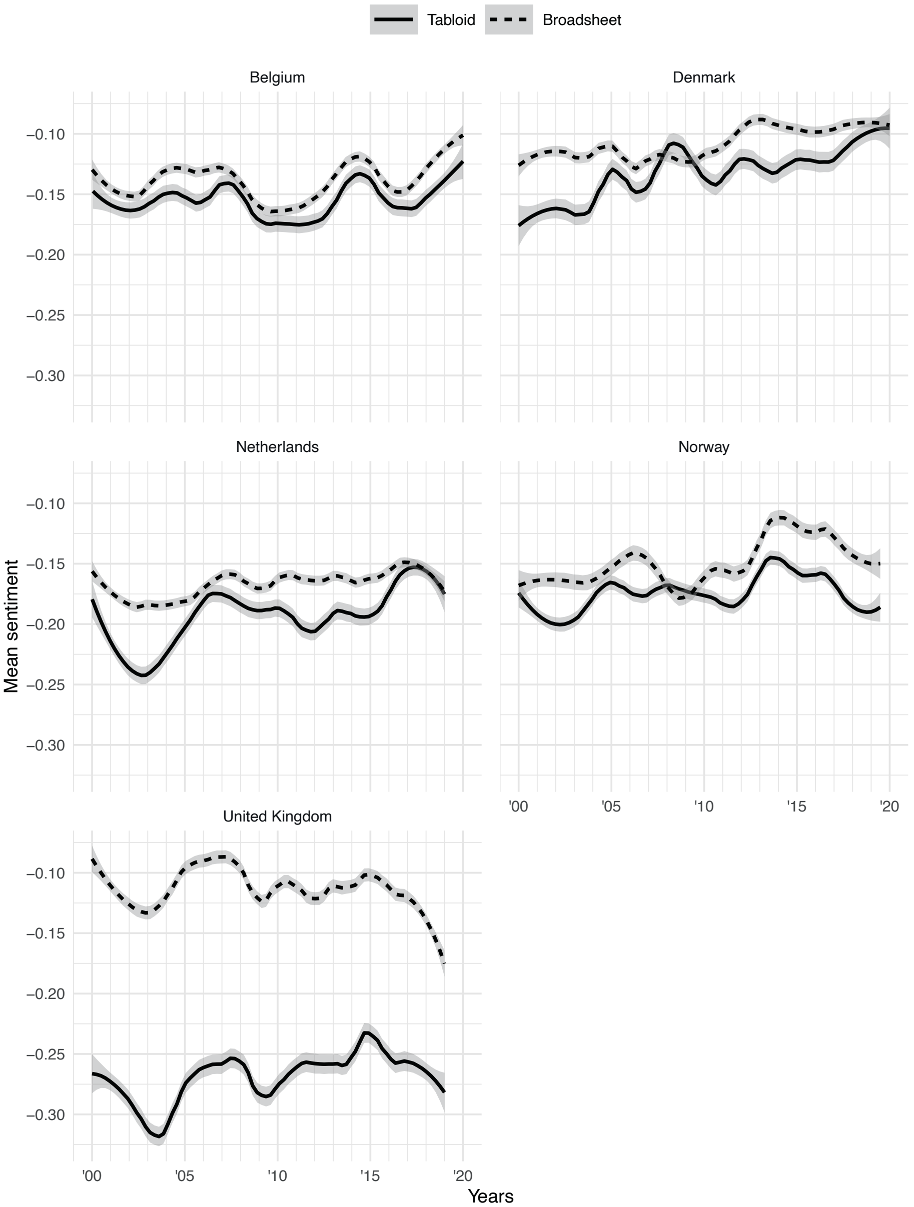


Figure 4: Sentiment of political news in tabloid and broadsheet newspapers by country, 2000-2019

The validity of the topic diversity measure is evaluated using a set of 200 Norwegian and 200 English human-coded article pairs. These article pairs have been sampled in a stratified way, so that each level of the (continuous) content diversity measure is represented equally in the human-coded sample. From the UK dataset, a random subsample of 20 article-pairs is used to test the intercoder reliability between two coders (the main author and a student assistant), resulting in a sufficient Krippendorff's alpha of .88. The final validation results show a strong correlation between the hand-coded and computed diversity measure in both the UK ($r(198)$ = -.79, $p <$ .001) and Norway ($r(198)$ = -.73, $p <$ .001). When visualized as a box plot (Figure 5), it is clear that manually coded article pairs that are about the same topic generally have a topic diversity below .6, while pairs that are not about the same topic are generally above .6. Thus a topic diversity value of below .6 (or a cosine similarity value above .4) is used as a cut-off point to determine which article pairs are about the same topic.

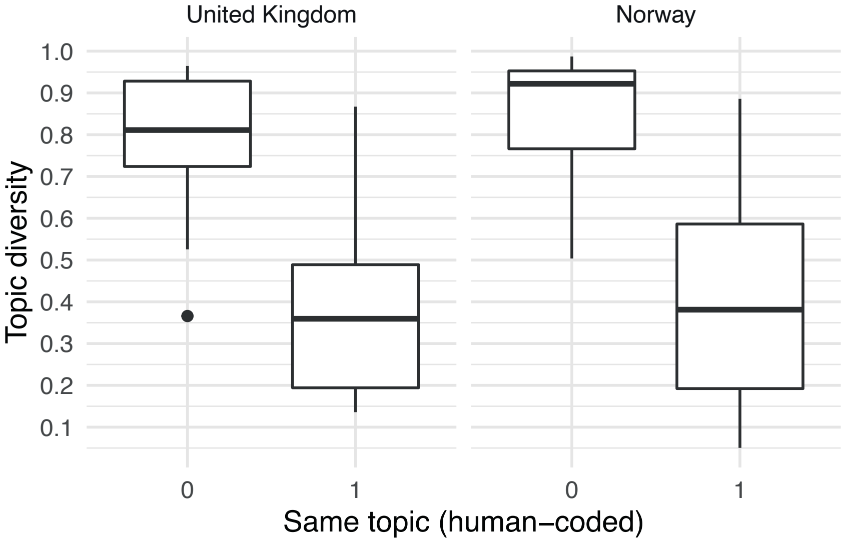


Figure 5: Topic diversity for articles that are (1) or are not (0) about the same topic

5.2 News diversity

The tables and figures in this section present the levels of and trends in news diversity in Denmark, Norway, The Netherlands and the UK. Figure 6 shows the percentage of article pairs with a diversity of .6 or lower (i.e. article pairs that are about the same topic) for each possible newspaper pair (rows) in each country (columns), with the bottom row showing the average trend. In this figure no clear trends are visible, except for the UK. Even so, the regression results in Table 5 show highly significant and negative results on the time variable, indicating a slight downward trend in the percentage of article pairs about the same topic. News diversity in terms of the discussed topics thus slightly increases rather than decreases over time.

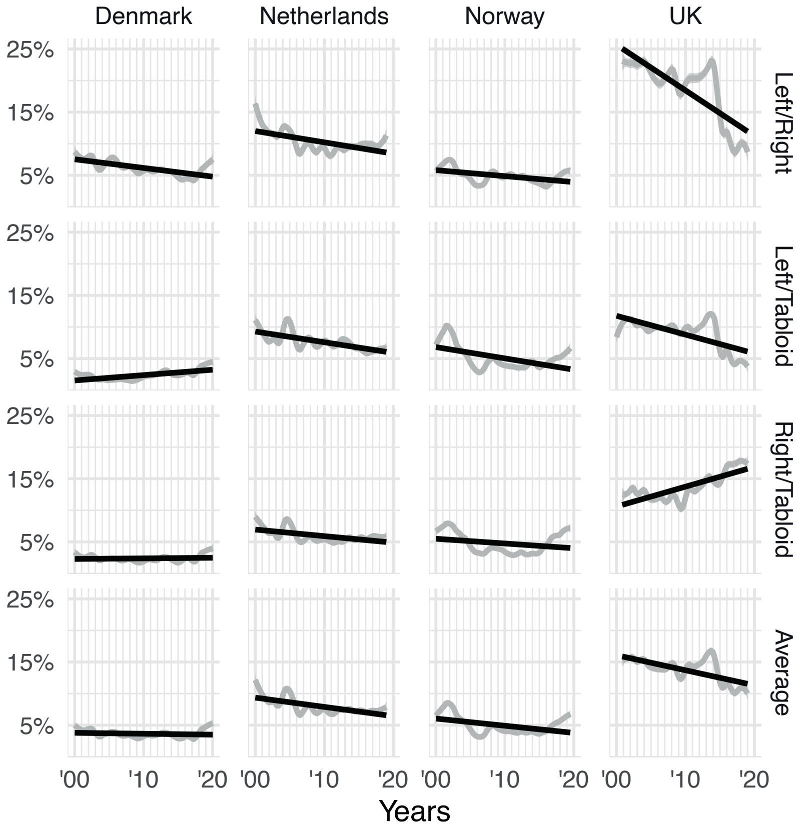


Figure 6: Percentage of article pairs that are about the same topic, 2000-2019

There are however notable variations between the different countries. In all countries except Norway, the coefficients for the newspaper pair dummies are highly significant and negative, indicating that the left- and right-wing broadsheets in each country are more similar to each other than they are to the tabloid newspaper. This is most strongly visible in Denmark, where both broadsheets are equally less similar to the tabloid than to each other. The effect of time in Denmark is also an order of magnitude weaker than in the other countries. In The Netherlands the left-wing broadsheet is clearly closer to the tabloid than the right-wing broadsheet, though both remain more similar to each other. In the UK this is exactly opposite, with the right-wing broadsheet being closer to the tabloid than the left-wing broadsheet. But what is even more noteworthy is that in the right-most column of Figure 6 the right-wing and tabloid newspapers in the UK actually become more alike over time, while the similarity between the left- and right-wing newspapers strongly decreases.

Table 5: Regression models, dependent variable: topic diversity (% of article pairs about the same topic). Denmark, The Netherlands, Norway and United Kingdom, 2000-2019.

	Dependent variable:			
	Topic diversity			
	DK	NL	NO	UK
Topic diversity	.148***	.110***	.250***	.343***
(lagged)	(.007)	(.008)	(.007)	(.007)
Time (in years)	−.018***	−.146***	−.114***	−.113***
	(.006)	(.007)	(.007)	(.006)
Left-wing/Tabloid	−.821***	−.469***	.037**	−.855***
	(.016)	(.017)	(.017)	(.017)
Right-wing/Tabloid	−.820***	−.772***	−.021	−.430***
	(.016)	(.018)	(.016)	(.016)
Constant	.547***	.411***	−.005	.445***
	(.011)	(.012)	(.012)	(.012)
Observations	21,686	17,291	20,097	16,688
R^2	.224	.166	.085	.396
Adjusted R^2	.224	.166	.085	.396
Residual Std. Error	.881	.913	.957	.777
F Statistic	1,564.952***	861.050***	468.263***	2,734.278***

Note: *p<0.1; **p<0.05; ***p<0.01

The findings presented in Table 6 show a highly significant difference in the level of sentiment diversity between either the left- or right-wing broadsheet and the tabloid, except in Norway. So when broadsheets discuss the same topic, they tend to do so with more similar sentiment compared to the tabloid. When combining this finding with the negativity bias results presented in Figure 4, it seems likely that the sentiment diversity between broadsheets and tabloids stems from a difference in the level of negativity with which they report events. This difference is relatively stable in Denmark, The Netherlands and Norway, while a significant trend is observed in the UK. In the latter case, the tone with which newspapers discuss the same topics becomes more dissimilar, but the broadsheets remain more similar to each other than to the tabloid. Also, the amount of explained variance in each model is very low, indicating that alternative aspects that are omitted from these models are important when explaining tone and the difference between newspaper articles. In none of the countries there is any indication that news diversity in terms of tone is either decreasing or increasing.

Table 6: Regression models, dependent variable: sentiment diversity. Denmark, The Netherlands, Norway and United Kingdom, 2000-2019.

	Dependent variable:			
	Sentiment diversity			
	DK	NL	NO	UK
Sentiment diversity	.006	.008	.001	−.009
(lagged)	(.008)	(.008)	(.008)	(.008)
Time (in years)	.012	−.013*	−.010	.121***
	(.008)	(.008)	(.008)	(.008)
Left-wing/Tabloid	.139***	.190***	.001	.244***
	(.019)	(.019)	(.019)	(.019)
Right-wing/Tabloid	.171***	.207***	.005	.232***
	(.020)	(.019)	(.018)	(.019)
Constant	−.091***	−.127***	−.004	−.163***
	(.012)	(.013)	(.013)	(.014)
Observations	14,328	16,566	17,050	16,477
R^2	.007	.009	0.000	.027
Adjusted R^2	.007	.009	−0.000	.027
Residual Std. Error	.962	.994	.989	.978
F Statistic	24.851***	39.193***	.498	113.992***
Note:		*p<0.1; **p<0.05; ***p<0.01		

5.3 Political parallelism

The tables and figures in this section present the levels of and trends in political parallelism in Denmark, The Netherlands and the UK. In Table 7 the multilevel regression results show the effect of party-newspaper alignment on the amount of attention for political parties. In all countries, alignment between a newspaper and a political party (see Table 3) has a highly significant and positive effect on the amount of coverage a party receives, indicating the persistence of political parallelism in terms of party attention. Results on party sentiment do however not show any significant effect of party-newspaper alignment (Table 8). Looking at the interaction models in Table 9, there is also a highly significant interaction between alignment and time for both The Netherlands and the United Kingdom. For Denmark, this interaction is not significant, but the interaction with the squared time variable is. For sentiment the results are again non-significant (Table 10). The marginal effects plots for the party attention models are presented in Figure 7. These plots show the marginal effect of time, for aligned and unaligned parties, on attention (on the left) and the marginal effect of alignment on attention (on the right).

The plots show a decidedly linear and negative trend in the marginal effect of alignment on attention in both The Netherlands and the United Kingdom. At the same time, the marginal effect of time is also negative in these countries, but only for aligned parties. Thus, the effect of parallelism on attention decreases in these countries over time. In Denmark, the results are quite different. Here there is a non-linear trend in the marginal effect of alignment on attention. This is also reflected in the non-significant difference in the marginal effect of time for aligned and unaligned parties. So in Denmark, rather than decrease, the effect of parallelism on attention is strong at the start of the period, drops towards the middle of the period, then rises again at the end of the period. Consistent with the regression results, there are no such trends to be identified in the marginal effects plots for sentiment (Figure 8). In these plots, the results do not differ significantly between aligned and non-aligned parties, nor do they differ significantly from zero over time.

Table 7: Multilevel regression models, dependent variable: party attention in newspaper. Denmark, The Netherlands and United Kingdom, 2000-2019.

	Dependent variable:		
	Attention		
	DK	NL	UK
Vote share	.597***	.435***	.165***
	(.018)	(.016)	(.016)
Party extremity	.027	-.061	-.207
	(.069)	(.067)	(.146)
Cabinet party	.420***	.744***	.193***
	(.021)	(.021)	(.022)
Prime minister party	.321***	.187***	.897***
	(.032)	(.032)	(.024)
Time (in months)	-.859**	-1.148**	-.300
	(.388)	(.472)	(.188)
Time, squared	.194	-.975**	.259
	(.364)	(.447)	(.189)
Alignment	.059***	.100***	.329***
	(.023)	(.028)	(.012)
Constant	-.199***	-.218***	-.138
	(.072)	(.062)	(.146)
No. of groups	13	14	12
SD(group)	0.255	0.227	0.505
Observations	4,570	4,252	4,870
Log Likelihood	-1,779.312	-2,361.188	1,256.359
Akaike Inf. Crit.	3,578.624	4,742.375	-2,492.719
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 8: Multilevel regression models, dependent variable: party sentiment in newspaper. Denmark, The Netherlands and United Kingdom, 2000-2019

	Dependent variable:		
	Sentiment		
	DK	NL	UK
Vote share	-.036	.053*	-.028
	(.035)	(.032)	(.062)
Party extremity	-.118***	-.033	.009
	(.039)	(.082)	(.071)
Cabinet party	.023	-.077*	.243**
	(.055)	(.043)	(.112)
Prime minister party	-.069	-.061	-.100
	(.085)	(.065)	(.125)
Time (in months)	5.300***	12.814***	-2.272**
	(1.051)	(.968)	(1.045)
Time, squared	.524	-1.540*	-2.889***
	(1.006)	(.920)	(1.049)
Alignment	.072	.032	.116*
	(.062)	(.057)	(.065)
Constant	.008	-.018	-.049
	(.042)	(.077)	(.068)
No. of groups	13	14	12
SD(group)	0.127	0.272	0.225
Observations	4,520	4,248	4,417
Log Likelihood	-6,353.556	-5,419.711	-6,207.277
Akaike Inf. Crit.	12,727.110	10,859.420	12,434.550
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 9: Multilevel regression models with interaction, dependent variable: party attention in newspaper. Denmark, The Netherlands and United Kingdom, 2000-2019.

	Dependent variable:		
		Attention	
	DK	NL	UK
Vote share	.594***	.425***	.154***
	(.019)	(.016)	(.017)
Party extremity	.029	-.060	-.213
	(.069)	(.068)	(.149)
Cabinet party	.421***	.746***	.186***
	(.021)	(.021)	(.021)
Prime minister party	.320***	.207***	.903***
	(.032)	(.033)	(.023)
Time (in months)	-.898**	-.360	.495**
	(.410)	(.500)	(.195)
Time, squared	-.650*	-1.349***	.076
	(.384)	(.475)	(.194)
Alignment	.064***	.087***	.322***
	(.023)	(.028)	(.012)
Alignment * Time	.435	-6.440***	-8.003***
	(1.227)	(1.469)	(.620)
Alignment * Time, squared	7.754***	2.564*	1.383**
	(1.164)	(1.415)	(.644)
Constant	-.197***	-.222***	-.138
	(.072)	(.063)	(.149)
No. of groups	13	14	12
SD(group)	0.255	0.227	0.505
Observations	4,570	4,252	4,870
Log Likelihood	-1,754.947	-2,345.402	1,342.416
Akaike Inf. Crit.	3,533.895	4,714.805	-2,660.831
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 10: Multilevel regression models with interaction, dependent variable: party sentiment in newspaper. Denmark, The Netherlands and United Kingdom, 2000-2019.

	Dependent variable:		
		Sentiment	
	DK	NL	UK
Vote share	-.038	.042	-.027
	(.036)	(.033)	(.063)
Party extremity	-.118***	-.031	.009
	(.039)	(.082)	(.071)
Cabinet party	.023	-.073*	.242**
	(.055)	(.043)	(.112)
Prime minister party	-.068	-.031	-.099
	(.085)	(.067)	(.125)
Time (in months)	5.331***	13.558***	-2.140*
	(1.115)	(1.027)	(1.107)
Time, squared	.203	-1.581	-2.874***
	(1.065)	(.979)	(1.108)
Alignment	.074	.021	.116*
	(.062)	(.057)	(.065)
Alignment * Time	-.219	-6.490**	-1.230
	(3.356)	(3.031)	(3.370)
Alignment * Time, squared	2.955	-.333	-.233
	(3.246)	(2.922)	(3.440)
Constant	.008	-.024	-.049
	(.042)	(.076)	(.068)
No. of groups	13	14	12
SD(group)	0.127	0.272	0.225
Observations	4,520	4,248	4,417
Log Likelihood	-6,348.914	-5,413.374	-6,202.921
Akaike Inf. Crit.	12,721.830	10,850.750	12,429.840
Note:			*p<0.1; **p<0.05; ***p<0.01

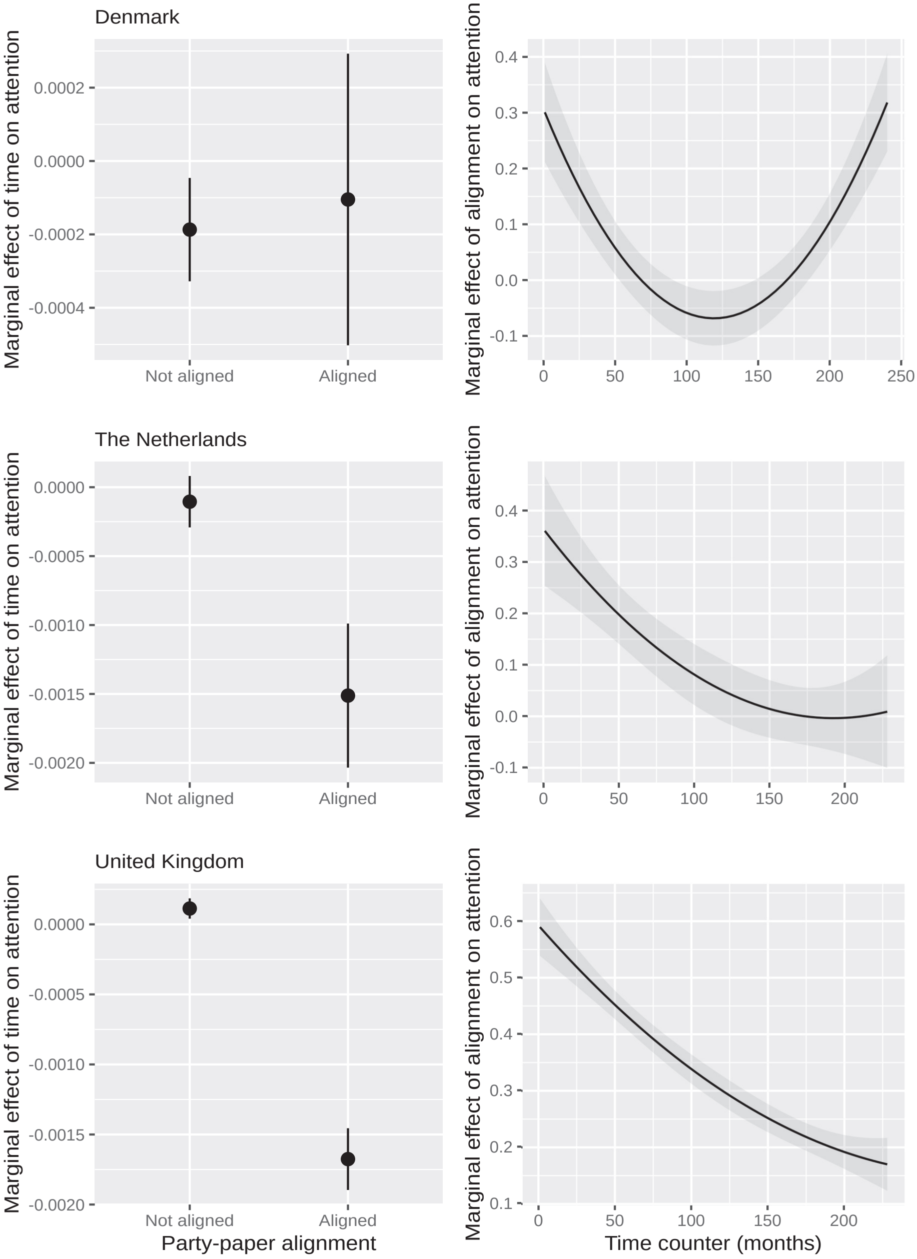


Figure 7: Conditional marginal effects of time (across alignment) and alignment (across time) on party attention

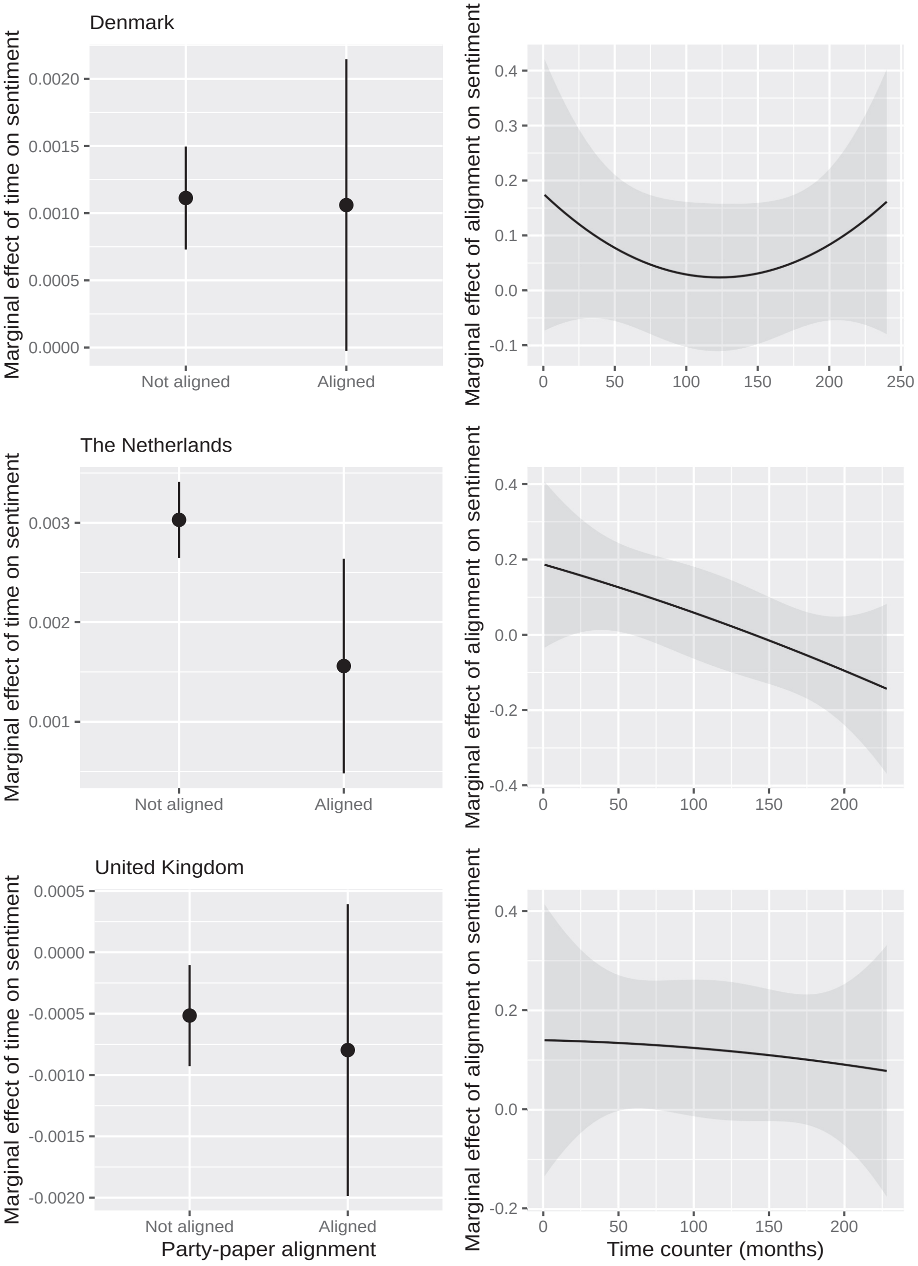


Figure 8: Conditional marginal effects of time (across alignment) and alignment (across time) on party sentiment

6 Discussion & Conclusion

In this final section answers are presented to the research questions formulated in Chapter 1. The first research question, relating to the use of computational methods for the measurement of topic diversity and sentiment, is discussed in section 6.1. Research questions 2 and 3, relating to the levels of news diversity and political parallelism and their development over time, are discussed in section 6.2. The limitations in the answers and opportunities for further research are discussed in section 6.3, while the final chapter is concluded in section 6.4 with a concise summary and concrete answers to each research question.

6.1 Computational text analysis

The validation results of both the sentiment dictionaries and the topic diversity measure (RQ1) illustrate how human labor can to a substantial extent be replaced by computational approaches while continuing to produce valid results. When comparing the performance of the WE dictionaries to the results found by Boukes et al. (2020), the WE dictionaries outperform all of the manually created dictionaries. The WE dictionary method is however not entirely without human input, as manually coded sentences are used to optimize the dictionaries. The same is true for the topic diversity measure, where human input is used both to validate the computational measure and to derive the optimal cutoff point between article pairs that are (not) about the same topic. Such "supervised" computational methods (see e.g. Boumans & Trilling, 2016) have clear advantages when it comes to increasing the validity of the resulting data. The results presented here emphasize that point. Combining a limited amount of human effort with computational methods yields useful and valid results.

As illustrated by Denny & Spirling (2018), data validity can also be profoundly affected by data cleaning and pre-processing steps. In this regard, using NLP to extract lemma-UPOS pairs for use in further methods seems to account for most of the issues with pre-processing that Denny & Spirling (2018) raise. Although not evaluated explicitly in this thesis, it is very likely that thorough data cleaning and pre-processing have significantly contributed to the performance of both the WE sentiment dictionaries and the topic diversity measure.

While the sentiment dictionaries are extensively validated, it is important to note that they are validated on their capacity to detect valence. Besides valence, evaluation (the source/target of valence) is a relevant aspect of sentiment, especially in the context of news diversity and political parallelism. This aspect is partially addressed by using sentences as the unit of analysis, based on the premise that proximity between valence words and a political party strongly increases the likelihood that those words provide an evaluation of the party. Similarly, with sentiment diversity the articles are about the same topic, supporting the general statement that these articles discuss the same topic with more (dis)similar sentiment. Both these examples illustrate how the validity and relevance of the sentiment analysis method can be improved by accounting for context, even without the computationally complex measurement of evaluations.

Regardless, the question remains to what extent the absence of sentiment findings on both diversity and parallelism can be attributed to the method. By their very nature, dictionaries can only be used to detect valence, and the addition of fine-grained analyses can only partly mitigate this limitation. Despite that fact, I would however argue that the findings do bear substantive relevance. The absence of findings in both cases cannot be used to definitely say that there are no effects of evaluation. For that a more suitable method is necessary, one that accounts for evaluation. However, the WE dictionaries are valid and effective in detecting valence, and the absence of valence effects is thus a valid result. Broadsheet newspapers do not differentiate themselves based on valence/negativity bias when talking about the same topic, nor do they structurally make use of more positive/negative valence when mentioning (un)aligned political parties. Tabloid newspapers are however significantly more negative than broadsheets in the way they discuss topics. These conclusions are, while constrained, substantively relevant.

The example with sentiment and evaluation is illustrative of a more general pitfall in the use of computational text analysis methods. Besides the necessity of good data quality (garbage in, garbage out, as with any method), computational methods that might seem to work fine (or even great) in one scenario do not perform adequately in another. With sentiment analysis specifically, this is illustrated by the large body of research into product reviews, often stating that sentiment analysis methods provide excellent results on product reviews. Product reviews are however not newspaper articles, and the study of economic news (see Boukes et al., 2020) is not the study of other news categories, or news in general. If this thesis shows anything with regards to computational methods, it is that these methods should be very closely scrutinized to determine whether or not they are adequate for the task at hand. Invariably, methods that are more closely intertwined with the data to which they are applied are capable of producing more valid results. The findings with sentiment dictionaries illustrate this, as the word embeddings used to create these dictionaries are based directly on the data to which they are applied. Manual evaluation is however essential, both for the validation and fine-tuning of the method.

Arguments can be made about the limits of computational analyses and the interpretability of results, such as Taylor's (1971) "brute data" argument. However, I argue that computational analyses are not inferior to analyses based directly on human observation.

Rather, they are different, and can be used to assess different aspects of social reality. Knowing how often political parties are mentioned by newspapers (or the sentiment of the context) is not the same as manually interpreting and evaluating those contexts. However, how often a party gets mentioned is in itself already informative, even without knowing the exact context of each mention. Similarly, knowing the lexical diversity between two newspaper articles is informative even without knowing the actual substantive content of those two articles. The results presented here provide a large-scale overview of the trends and developments in news diversity in the first two decades of the 21st century. The fact that these results are highly empirical and descriptive illustrates a crucial contribution that computational methods can make to the study of political communication. Before drawing conclusions on *why* trends occur, such methods are very well suited to evaluate in first instance *if* theorized trends really occur. In a scientific field that is overrun by an exponentially increasing amount of information to study, such methods provide a way to strengthen investigations, by providing stronger descriptive inferences as a starting point for causal explorations.

The above is not to say that context is irrelevant. Context is key in further investigation of the results presented here. Without knowing further (contextual) details about news diversity, it is not possible to evaluate its role in the functioning of democracy. Neither is it enough to simply observe political parallelism by counting the number of times parties are being mentioned. The (absence of) results relating to both general sentiment diversity in news articles and the sentiment context in which parties are mentioned illustrates that computational analyses are not perfect. It is very likely that the nuances in context are not detected by the methods employed, further illustrating the need and relevance of manually evaluating context in texts. Computational analysis methods are an additional and potentially very useful tool, but they are not capable of interpretation and therefore not likely to replace human evaluation anytime soon.

6.2 News diversity & Political parallelism

As argued previously, various theories on the economic and professional developments of newspaper markets implicitly provide conflicting expectations about the possible trends in news diversity. One pervasive argument, the economically challenging conditions in shrinking newspaper markets (Curran, 2010), does however not seem to have the expected effect of a downward trend in news diversity. Rather, the results presented in this thesis show a decidedly different trend. The levels of and trends in news diversity (RQ2) are generally high, and show a slightly increasing rather than decreasing trend. Variations in the levels of news diversity between newspaper pairs are however visible. Particularly the diversity in topics and sentiment between either of the broadsheets and the tabloid is higher than between the broadsheets themselves in all investigated countries.

As for the level of and trends in political parallelism (RQ3), the results indicate first and foremost that parallelism can still be found, and is more pronounced in UK newspapers than in Dutch and Danish ones. A probable cause for this difference is the majoritarian model of democracy in the UK versus the consensus model in The Netherlands and Denmark (Lijphart, 2012). This difference likely influences the way in which news media report on politics, more focused on conflict and partisanship in the majoritarian model and less so in the consensus model. The focus on partisanship is also reflected in the persistence of partisan endorsements in the British press, as well as the more pronounced trends in topic and sentiment diversity. Important news values such as relevance and power are however still the predominant drivers of news attention in all three countries, as party size and incumbency explain the main share of variation in news attention to parties. The results nevertheless suggest that there is a case to be made that partisan perspectives systematically interfere with non-partisan news values in news production.

Additionally, the trends in The Netherlands and the UK suggest that parallelism in terms of attention is decreasing over time. And while Denmark shows an unexpected pattern, a decrease in parallelism followed by an increase, the trends in the other two countries might cautiously be interpreted as an indication that parallelism is decreasing over time. That being said, there are many circumstances that may influence the level of parallelism over time. In the Netherlands, for example, the dominant mainstream left party (PvdA) has suffered major electoral losses during the end of the investigated period. In the UK the Brexit campaign, vote and negotiations have dominated the political debate, and through that domination reinforced the party allegiances of newspapers. In Denmark there has been substantial variation in the support for mainstream parties during the investigated period, partially related to the extreme ups and downs of the radical right Danish People's Party.

While substantive explanations for the trends in news diversity and parallelism are not explicitly tested, the direction of the findings does provide an indication of which theories provide more likely explanations. Increasing professionalization (e.g. Deuze, 2005) and through that the increasing importance of news values (Galtung & Ruge, 1965; Harcup & O'Neill, 2001, 2017) are likely explanations of the downward trend in political parallelism, as shared norms and a more objective/factual style of reporting would result in less parallelism. However, they do not explain the slightly increasing levels of news diversity, which would be expected to decrease because of the same reasons as for parallelism. The concept of interpretive journalism provides a plausible alternative explanation, allowing for stable or even increasing news diversity. Journalistic role perceptions have changed in recent years (Mellado et al., 2017) due to the decreasing importance of traditional media when it comes to “bringing the news (first)” in comparison to online and social media. This applies in particular to printed newspapers, which increasingly provide interpretation, analysis and opinions instead (Esser & Umbricht, 2014; Soontjens, 2019; Strömbäck & Aalberg, 2008). When such content is used by newspapers to differentiate themselves from competitors, news diversity increases, or when accounting for economic pressures (e.g. Curran, 2010), at least prevent a decrease. Such interpretive journalism seems however not to depend on partisan bias, as the findings show a decrease of political parallelism over time, at least in the UK and The Netherlands. Despite the economic concerns described by Curran (2010), newspapers have not decreased in their (external) diversity. At the same time, it also seems unlikely that economic pressures have resulted in more politically biased news content as a means to ensure profitability, because parallelism has either remained stable or decreased.

6.3 Limitations and further research

The design of the studies in this thesis imposes some limitations with regards to the interpretation of the results. While incorporating all articles published by the selected newspapers in each country, the sample of both countries and newspapers limits the generalizability of the results. Examples are the lack of a left-wing broadsheet in Norway due to data availability and little variation in the country selection in terms of political and media systems. Besides these general notes, the methods also impose limitations. The most prominent example is the lack of evaluation in the sentiment analysis method, as already discussed extensively in section 6.1. This makes the current investigation of parallelism incomplete, as it only accounts for discrepancies in the amount of valence in proximity to a political party, while political parallelism can also take the form of implicit and explicit party evaluations. The absence of evaluation in relation to valence points to the general limitation that computational text analysis methods are very suitable for empirically describing large amounts of data (valence) but not as much for interpreting complex linguistic relationships (evaluations).

Some critics may argue that because of this limit, the results are simply not valid. It is however the generalizability rather than the validity of the results that is at stake here. The parallelism findings are largely in line with theoretical expectations, and while acknowledging that party attention is only part of political parallelism, the disproportionality in the amount of attention that parties receive is a relevant factor. As for news diversity, there is a clear discrepancy between the theoretical assumptions and the current findings, which might be explained by methodological differences between the studies. Manual evaluations of news diversity often either focus on the systemic level (i.e. publishing houses and ownership structures) or on categorizing individual news items based on topics, to see the variation in the attention for those topics between different outlets. Both approaches are substantively different from the method employed here. It might be that the variation in topics between news outlets has declined during the last two decades (e.g. Boczkowski & Santos, 2007; Vogler et al., 2020). However, this thesis shows there is no decline in the variety of word use, nor an increase in the amount of article pairs that concern the same concrete (not general) topic. So while others might find a decrease in news diversity where I do not, this is most likely caused by different operationalizations of news diversity (see also Joris et al., 2020). There is however a clear argument for using computational methods, as they remove the need to group the content of articles into categories.

As the findings in this thesis are primarily descriptive, their main function is to inform further theoretical development and provide information regarding the possible directions for future research. In this specific context, it seems that more in-depth investigation of where news diversity originates is warranted, rather than whether or not it is declining. Interpretive journalism (Esser & Umbricht, 2014) provides a framework for investigating this, although one needs to be careful not to confuse professional journalistic practice (e.g. Patterson & Donsbach, 1996) with the commercial logic to capture and retain audiences (Hagar et al., 2021). With regards to parallelism, the deviating curvilinear trend in Denmark seems to suggest that parallelism is not necessarily tightly linked to a newspaper's profile. Further research might investigate and explain possible causes for such a trend. These might be found in context-specific developments in party strength and general developments in the media landscape.

6.4 Conclusion

The findings presented in this thesis show that despite the heralded downfall of newspapers and the diversity of their content, they seem to have dealt with the challenges of an increasingly competitive 24-hour news landscape in a way that does not decrease news diversity, nor increase political parallelism. As described in Chapter 2, there might have been cause for alarm regarding the impact of news media on the functioning of democracy if large developments towards either more (polarization) or less (homogenization) news diversity would be found. In concrete answer to RQ2, the level of news diversity during the first two decades of the 21st century is however high, as illustrated by the low percentage of article pairs with the same topic. At the same time there are tentative trends towards an increase in news diversity. The level of political parallelism (RQ3) remains substantial in all investigated countries, though a decreasing trend is visible in The Netherlands and the UK. Considering these findings, the information-supplying role of newspapers in democracy does not seem to be under any immediate threat.

In answer to the question how computational text analysis methods might be employed in the analysis of news sentiment and topic diversity (RQ1), the findings suggest that these methods have the potential to test long-standing theoretical assumption on a larger scale than what is possible with manual analysis. In doing so, the strengths and weaknesses of computational analyses are revealed. While it is difficult to provide substantive explanations on why certain trends are (not) observed, the empirical contributions serve to further the academic debates on news diversity, political parallelism and the use of computational methods in the fields of journalism studies and political communication. The findings show that theoretical assumptions that seem logical and plausible do not always properly reflect the reality that they are based on. Thus, computational analyses can provide a new and possibly surprising perspective on such theories and assumptions.

7 References

Aldayel, A., & Magdy, W. (2021). Stance detection on social media: State of the art and trends. *Information Processing & Management*, *58*(4), 102597. https://doi.org/10.1016/j.ipm.2021.102597

Allern, S., & Blach-Ørsten, M. (2011). The News Media as a Political Institution. *Journalism Studies*, *12*(1), 92–105. https://doi.org/10.1080/1461670X.2010.511958

Almeida, F., & Xexéo, G. (2019). Word Embeddings: A Survey. *arXiv:1901.09069 [Cs, Stat]*. https://arxiv.org/abs/1901.09069

Al-Rfou, R., Perozzi, B., & Skiena, S. (2013). Polyglot: Distributed Word Representations for Multilingual NLP. *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, 183–192.

Altheide, D. L. (2013). Media Logic, Social Control, and Fear. *Communication Theory*, *23*(3), 223–238. https://doi.org/10.1111/comt.12017

Altheide, D. L., & Snow, R. P. (1979). *Media Logic*. Sage Publications.

Bakker, R., de Vries, C., Edwards, E., Hooghe, L., Jolly, S., Marks, G., Polk, J., Rovny, J., Steenbergen, M., & Vachudova, M. A. (2015). Measuring party positions in Europe: The Chapel Hill expert survey trend file, 1999–2010. *Party Politics*, *21*(1), 143–152. https://doi.org/10.1177/1354068812462931

Baumgartner, F. R., & Bonafont, L. C. (2015). All News is Bad News: Newspaper Coverage of Political Parties in Spain. *Political Communication*, *32*(2), 268–291. https://doi.org/10.1080/10584609.2014.919974

Beckers, K., Masini, A., Sevenans, J., van der Burg, M., De Smedt, J., Van den Bulck, H., & Walgrave, S. (2019). Are newspapers' news stories becoming more alike? Media content diversity in Belgium, 1983–2013. *Journalism*, *20*(12), 1665–1683. https://doi.org/10.1177/1464884917706860

Beckers, K., Walgrave, S., Wolf, H. V., Lamot, K., & van Aelst, P. (2021). Right-wing Bias in Journalists' Perceptions of Public Opinion. *Journalism Practice*, *15*(2), 243–258. https://doi.org/10.1080/17512786.2019.1703788

Boczkowski, P. J. (2009). Materiality and Mimicry in the Journalism Field. In B. Zelizer (Ed.), *The Changing Faces of Journalism : Tabloidization, Technology and Truthiness* (pp. 56–67). Routledge.

Boczkowski, P. J., & Santos, M. de. (2007). When More Media Equals Less News: Patterns of Content Homogenization in Argentina's Leading Print and Online Newspapers. *Political Communication*, *24*(2), 167–180. https://doi.org/10.1080/10584600701313025

Boukes, M., van de Velde, B., Araujo, T., & Vliegenthart, R. (2020). What's the Tone? Easy Doesn't Do It: Analyzing Performance and Agreement Between Off-the-Shelf Sentiment Analysis Tools. *Communication Methods and Measures*, *14*(2), 83–104. https://doi.org/10.1080/19312458.2019.1671966

Boumans, J. W. (2016). *Outsourcing the news? An empirical assessment of the role of sources and news agencies in the contemporary news landscape.*

Boumans, J. W., & Trilling, D. (2016). Taking Stock of the Toolkit. *Digital Journalism*, *4*(1), 8–23. https://doi.org/10.1080/21670811.2015.1096598

Boumans, J. W., Trilling, D., Vliegenthart, R., & Boomgaarden, H. (2018). The Agency Makes the (Online) News World go Round: The Impact of News Agency Content on Print and Online News. *International Journal of Communication*, *12*(0), 22.

Brandenburg, H. (2006). Party Strategy and Media Bias: A Quantitative Analysis of the 2005 UK Election Campaign. *Journal of Elections, Public Opinion and Parties*, *16*(2), 157–178. https://doi.org/10.1080/13689880600716027

Brants, K., & Van Praag, P. (2006). Signs of Media Logic Half a

Century of Political Communication in the Netherlands. *Javnost - The Public*, *13*(1), 25–40. https://doi.org/10.1080/13183222.2006.11008905

Brants, K., & van Praag, P. (2017). Beyond Media Logic. *Journalism Studies*, *18*(4), 395–408. https://doi.org/10.1080/1461670X.2015.1065200

Brüggemann, M., Esser, F., & Humprecht, E. (2012). The Strategic Repertoire of Publishers in the Media Crisis. *Journalism Studies*, *13*(5-6), 742–752. https://doi.org/10.1080/1461670X.2012.664336

Bullock, J. G. (2011). Elite Influence on Public Opinion in an Informed Electorate. *American Political Science Review*, *105*(3), 496–515. https://doi.org/10.1017/S0003055411000165

Bullock, J. G. (2020). Party Cues. In E. Suhay, B. Grofman, & A. H. Trechsel (Eds.), *The Oxford Handbook of Electoral Persuasion* (pp. 128–150). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190860806.013.2

Chen, Y., & Skiena, S. (2014). Building Sentiment Lexicons for All Major Languages. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 383–389. https://doi.org/10.3115/v1/P14-2063

Curran, J. (2010). The Future of Journalism. *Journalism Studies*, *11*(4), 464–476. https://doi.org/10.1080/14616701003722444

Dailey, L., Demo, L., & Spillman, M. (2005). The Convergence Continuum: A Model for Studying Collaboration Between Media Newsrooms. *Atlantic Journal of Communication*, *13*(3), 150–168. https://doi.org/10.1207/s15456889ajc1303_2

de Vreese, C. H., Boomgaarden, H. G., & Semetko, H. A. (2011). (In)direct Framing Effects: The Effects of News Media Framing on Public Support for Turkish Membership in the European Union. *Communication Research*, *38*(2), 179–205. https://doi.org/10.1177/0093650210384934

de Vreese, C. H., Esser, F., & Hopmann, D. N. (2016). *Compar-*

ing Political Journalism. Routledge. https://doi.org/10.4324/9781315622286

Denny, M. J., & Spirling, A. (2018). Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. *Political Analysis*, *26*(2), 168–189. https://doi.org/10.1017/pan.2017.44

Deuze, M. (2005). What is journalism?: Professional identity and ideology of journalists reconsidered. *Journalism*, *6*(4), 442–464. https://doi.org/10.1177/1464884905056815

Döring, H., & Manow, P. (2021). *Parliaments and governments database (ParlGov): Information on parties, elections and cabinets in modern democracies. Development version.* http://www.parlgov.org/.

Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. *arXiv:1702.08608 [Cs, Stat].* https://arxiv.org/abs/1702.08608

Esmark, A., & Ørsten, M. (2008). Media and Politics in Denmark. In *Communicating Politics : Political Communication in the Nordic Countries.* Nordicom, University of Gothenburg.

Esser, F. (1999). "Tabloidization" of News: A Comparative Analysis of Anglo-American and German Press Journalism. *European Journal of Communication*, *14*(3), 291–324. https://doi.org/10.1177/0267323199014003001

Esser, F. (2013). Mediatization as a Challenge: Media Logic Versus Political Logic. In H. Kriesi, S. Lavenex, F. Esser, J. Matthes, M. Bühlmann, & D. Bochsler (Eds.), *Democracy in the Age of Globalization and Mediatization* (pp. 155–176). Palgrave Macmillan. https://doi.org/10.1057/9781137299871_7

Esser, F., & Umbricht, A. (2014). The Evolution of Objective and Interpretative Journalism in the Western Press: Comparing Six News Systems since the 1960s. *Journalism & Mass Communication Quarterly*, *91*(2), 229–249. https://doi.org/10.1177/1077699014527459

Firth, J. R. (1957). A Synopsis of Linguistic Theory, 1930-1955. In *Studies in Linguistic Analysis* (pp. 10–32). Blackwell.

Fox, N. J. (2008). Post-positivsm. In L. M. Given (Ed.), *The SAGE Encyclopaedia of Qualitative Research Methods* (Vol. 2, pp. 659–664). Sage.

Galtung, J., & Ruge, M. H. (1965). The Structure of Foreign News. *Journal of Peace Research*, *2*(1), 64–91. https://www.jstor.org/stable/423011

Gandy, O. H. (1980). Information in health: Subsidised news. *Media, Culture & Society*, *2*(2), 103–115. https://doi.org/10.1177/016344378000200201

Gill, J., & Womack, A. J. (2013). The Multilevel Model Framework. In M. A. Scott, J. S. Simonoff, & B. D. Marx (Eds.), *The SAGE Handbook of Multilevel Modeling* (pp. 3–20). Sage.

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Hagar, N., Wachs, J., & Horvát, E.-Á. (2021). Publishing patterns reflect political polarization in news media. *arXiv:2101.05044 [Cs]*. https://arxiv.org/abs/2101.05044

Hallin, D. C., & Mancini, P. (2004). *Comparing Media Systems: Three Models of Media and Politics*. Cambridge University Press. https://doi.org/10.1017/CBO9780511790867

Harcup, T., & O'Neill, D. (2001). What Is News? Galtung and Ruge revisited. *Journalism Studies*, *2*(2), 261–280. https://doi.org/10.1080/14616700118449

Harcup, T., & O'Neill, D. (2017). What is News? *Journalism Studies*, *18*(12), 1470–1488. https://doi.org/10.1080/1461670X.2016.1150193

Haselmayer, M., Wagner, M., & Meyer, T. M. (2017). Partisan Bias in Message Selection: Media Gatekeeping of Party Press Releases. *Political Communication*, *34*(3), 367–384. https://doi.

org/10.1080/10584609.2016.1265619

Hendrickx, J., & Ranaivoson, H. (2019). Why and how higher media concentration equals lower news diversity – The Mediahuis case. *Journalism*, *22*(11). https://doi.org/10.1177/1464884919894138

Hopmann, D. N., van Aelst, P., & Legnante, G. (2012). Political balance in the news: A review of concepts, operationalizations and key findings. *Journalism*, *13*(2), 240–257. https://doi.org/10.1177/1464884911427804

Hopmann, D. N., Vliegenthart, R., de Vreese, C. H., & Albæk, E. (2010). Effects of Election News Coverage: How Visibility and Tone Influence Party Choice. *Political Communication*, *27*(4), 389–405. https://doi.org/10.1080/10584609.2010.516798

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*(1), 11–21. https://doi.org/10.1108/eb026526

Joris, G., Grove, F. D., Damme, K. V., & Marez, L. D. (2020). News Diversity Reconsidered: A Systematic Literature Review Unraveling the Diversity in Conceptualizations. *Journalism Studies*, *21*(13), 1893–1912. https://doi.org/10.1080/1461670X.2020.1797527

Kahn, K. F., & Kenney, P. J. (2002). The Slant of the News: How Editorial Endorsements Influence Campaign Coverage and Citizens' Views of Candidates. *American Political Science Review*, *96*(2), 381–394. https://doi.org/10.1017/S0003055402000230

Lacy, S., & Simon, T. F. (1993). *The economics and regulation of United States newspapers*. Ablex Publishing Corporation.

Larcinese, V., Puglisi, R., & Snyder, J. M. (2011). Partisan bias in economic news: Evidence on the agenda-setting behavior of U.S. newspapers. *Journal of Public Economics*, *95*(9), 1178–1189. https://doi.org/10.1016/j.jpubeco.2011.04.006

Lelkes, Y. (2016). Winners, Losers, and the Press: The Relationship Between Political Parallelism and the Legitimacy Gap. *Political Communication*, *33*(4), 523–543. https://doi.org/10.1080/

10584609.2015.1117031

Lengauer, G., Esser, F., & Berganza, R. (2012). Negativity in political news: A review of concepts, operationalizations and key findings. *Journalism*, *13*(2), 179–202. https://doi.org/10.1177/1464884911427800

Lewis, J., Williams, A., & Franklin, B. (2008). Four Rumours and an Explanation. *Journalism Practice*, *2*(1), 27–45. https://doi.org/10.1080/17512780701768493

Lijphart, A. (2012). *Patterns of democracy: Government forms and performance in thirty-six countries* (2nd ed). Yale University Press.

McCombs, M. E., & Shaw, D. L. (1972). The Agenda-Setting Function of Mass Media. *The Public Opinion Quarterly*, *36*(2), 176–187. https://www.jstor.org/stable/2747787

Mellado, C., Hellmueller, L., Márquez-Ramírez, M., Humanes, M. L., Sparks, C., Stepinska, A., Pasti, S., Schielicke, A.-M., Tandoc, E., & Wang, H. (2017). The Hybridization of Journalistic Cultures: A Comparative Study of Journalistic Role Performance. *Journal of Communication*, *67*(6), 944–967. https://doi.org/10.1111/jcom.12339

Menke, M., Kinnebrock, S., Kretzschmar, S., Aichberger, I., Broersma, M., Hummel, R., Kirchhoff, S., Prandner, D., Ribeiro, N., & Salaverría, R. (2018). Convergence Culture in European Newsrooms. *Journalism Studies*, *19*(6), 881–904. https://doi.org/10.1080/1461670X.2016.1232175

Menke, M., Kinnebrock, S., Kretzschmar, S., Aichberger, I., Broersma, M., Hummel, R., Kirchhoff, S., Prandner, D., Ribeiro, N., & Salaverría, R. (2019). Insights from a Comparative Study into Convergence Culture in European Newsrooms. *Journalism Practice*, *13*(8), 946–950. https://doi.org/10.1080/17512786.2019.1642133

Mikolov, T., Yih, W., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. *Proceedings of the 2013*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 746–751.

Mohammad, S. M. (2016). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In H. L. Meiselman (Ed.), *Emotion measurement*. Elsevier.

Muddiman, A., McGregor, S. C., & Stroud, N. J. (2019). (Re)Claiming Our Expertise: Parsing Large Text Corpora With Manually Validated and Organic Dictionaries. *Political Communication*, *36*(2), 214–226. https://doi.org/10.1080/10584609.2018.1517843

Napoli, P. M. (1999). Deconstructing the diversity principle. *Journal of Communication*, *49*(4), 7–34. https://doi.org/10.1111/j.1460-2466.1999.tb02815.x

Nivre, J., Abrams, M., Agić, Ž., Ahrenberg, L., Antonsen, L., Aplonova, K., Aranzabe, M. J., Arutie, G., Asahara, M., Ateyah, L., Attia, M., Atutxa, A., Augustinus, L., Badmaeva, E., Ballesteros, M., Banerjee, E., Bank, S., Barbu Mititelu, V., Basmov, V., . . . Zhu, H. (2018). *Universal Dependencies 2.3.* LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Nordø, Å. D. (2021). Do Voters Follow? The Effect of Party Cues on Public Opinion During a Process of Policy Change. *Scandinavian Political Studies*, *44*(1), 45–66. https://doi.org/10.1111/1467-9477.12187

Østbye, H., & Aalberg, T. (2008). Media and Politics in Norway. In *Communicating Politics : Political Communication in the Nordic Countries* (pp. 83–102). Nordicom, University of Gothenburg.

Patterson, T. E., & Donsbach, W. (1996). News decisions: Journalists as partisan actors. *Political Communication*, *13*(4), 455–468. https://doi.org/10.1080/10584609.1996.9963131

Paulussen, S. (2012). Technology and the Transformation of News

Work: Are Labor Conditions in (Online) Journalism Changing? In E. Siapera & A. Veglis (Eds.), *The Handbook of Global Online Journalism* (pp. 192–208). Wiley-Blackwell. https://doi.org/10.1002/9781118313978.ch11

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Picard, R. G. (2014). Twilight or New Dawn of Journalism? *Journalism Practice*, *8*(5), 488–498. https://doi.org/10.1080/17512786.2014.905338

Popescu, M., Toka, G., Gosselin, T., & Pereira, J. S. (2011). *European Media Systems Survey: Results and Documentation.* University of Essex.

Rheault, L., Beelen, K., Cochrane, C., & Hirst, G. (2016). Measuring Emotion in Parliamentary Debates with Automated Textual Analysis. *PLOS ONE*, *11*(12), 1–18. https://doi.org/10.1371/journal.pone.0168843

Roessler, P. (2007). Media Content Diversity: Conceptual Issues and Future Directions for Communication Research. *Annals of the International Communication Association*, *31*(1), 464–520. https://doi.org/10.1080/23808985.2007.11679073

Sætra, H. S. (2018). Science as a Vocation in the Era of Big Data: The Philosophy of Science behind Big Data and humanity's Continued Part in Science. *Integrative Psychological and Behavioral Science*, *52*(4), 508–522. https://doi.org/10.1007/s12124-018-9447-5

Schudson, M. (2014). How to Think Normatively About News and Democracy. In K. Kenski & K. H. Jamieson (Eds.), *The Oxford Handbook of Political Communication* (pp. 95–106). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199793471.013.73

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral* *and Brain Sciences*, *3*(3), 417–424. https://doi.org/10.1017/S0140525X00005756

Seymour-Ure, C. (1974). *The political impact of mass media*. Constable.

Siles, I., & Boczkowski, P. J. (2012). Making sense of the newspaper crisis: A critical assessment of existing research and an agenda for future work. *New Media & Society*, *14*(8), 1375–1394. https://doi.org/10.1177/1461444812455148

Sjøvaag, H. (2016). Media diversity and the global superplayers: Operationalising pluralism for a digital media market. *Journal of Media Business Studies*, *13*(3), 170–186. https://doi.org/10.1080/16522354.2016.1210435

Soontjens, K. (2019). The Rise of Interpretive Journalism. *Journalism Studies*, *20*(7), 952–971. https://doi.org/10.1080/1461670X.2018.1467783

Soroka, S., & McAdams, S. (2015). News, Politics, and Negativity. *Political Communication*, *32*(1), 1–22. https://doi.org/10.1080/10584609.2014.881942

Straka, M., & Straková, J. (2017). Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 88–99.

Strömbäck, J., & Aalberg, T. (2008). Election News Coverage in Democratic Corporatist Countries: A Comparative Study of Sweden and Norway. *Scandinavian Political Studies*, *31*(1), 91–106. https://doi.org/10.1111/j.1467-9477.2008.00197.x

Taylor, C. (1971). Interpretation and the Sciences of Man. *The Review of Metaphysics*, *25*(1), 3–51. https://www.jstor.org/stable/20125928

Trenz, H.-J. (2004). Media Coverage on European Governance: Exploring the European Public Sphere in National Quality Newspapers. *European Journal of Communication*, *19*(3), 291–319. https://doi.org/10.1177/0267323104045257 van Aelst, P., & Walgrave, S. (2016). Information and Arena: The Dual Function of the News Media for Political Elites. *Journal of Communication*, *66*(3), 496–518. https://doi.org/10.1111/jcom.12229

van Atteveldt, W., Sheafer, T., Shenhav, S. R., & Fogel-Dror, Y. (2017). Clause Analysis: Using Syntactic Information to Automatically Extract Source, Subject, and Predicate from Texts with an Application to the 2008–2009 Gaza War. *Political Analysis*, *25*(2), 207–222. https://doi.org/10.1017/pan.2016.12

Van Cuilenburg, J. (1999). On Competition, Access and Diversity in Media, Old and New: Some Remarks for Communications Policy in the Information Age. *New Media & Society*, *1*(2), 183–207. https://doi.org/10.1177/14614449922225555

Van Cuilenburg, J. (2000). On Measuring Media Competition and Media Diversity: Concepts, Theories and Methods. In R. G. Picard (Ed.), *Measuring media content, quality, and diversity: Approaches and issues in content research*. Turku School of Economics and Business Administration.

van Dalen, A., Albæk, E., & de Vreese, C. H. (2011). Suspicious minds: Explaining political cynicism among political journalists in Europe. *European Journal of Communication*, *26*(2), 147–162. https://doi.org/10.1177/0267323111404841

van der Pas, D. J., van der Brug, W., & Vliegenthart, R. (2017). Political Parallelism in Media and Political Agenda-Setting. *Political Communication*, *34*(4), 491–510. https://doi.org/10.1080/10584609.2016.1271374

van Kempen, H. (2007). Media-Party Parallelism and Its Effects: A Cross-National Comparative Study. *Political Communication*, *24*(3), 303–320. https://doi.org/10.1080/10584600701471674

Vogler, D., Udris, L., & Eisenegger, M. (2020). Measuring Media Content Concentration at a Large Scale Using Automated Text Comparisons. *Journalism Studies*, *21*(11), 1459–1478. https://doi.org/10.1080/1461670X.2020.1761865

Young, L., & Soroka, S. (2012). Affective News: The Automated Coding of Sentiment in Political Texts. *Political Communication*, *29*(2), 205–231. https://doi.org/10.1080/10584609.2012.671234

Amsterdam
University
Press

COMPUTATIONAL COMMUNICATION RESEARCH 4.2 (2022) 424-462
HTTPS://DOI.ORG/10.5117/CCR2022.2.003.VRIE

The Sentiment is in the Details

A Language-agnostic Approach to Dictionary Expansion and Sentence-level Sentiment Analysis in News Media

Erik de Vries
Department of Media and Social Sciences, University of Stavanger
erikdevries@uis.no

Abstract

Determining the sentiment in the individual sentences of a newspaper article in an automated fashion is a major challenge. Manually created sentiment dictionaries often fail to meet the required standards. And while computer-generated dictionaries show promise, they are often limited by the availability of suitable linguistic resources. I propose and test a novel, language-agnostic and resource-efficient way of constructing sentiment dictionaries, based on word embedding models. The dictionaries are constructed and evaluated based on four corpora containing two decades of Danish, Dutch (Flanders and the Netherlands), English, and Norwegian newspaper articles, which are cleaned and parsed using Natural Language Processing. Concurrent validity is evaluated using a dataset of human-coded newspaper sentences, and compared to the performance of the Polyglot sentiment dictionaries. Predictive validity is tested through two long-standing hypotheses on the negativity bias in political news. Results show that both the concurrent validity and predictive validity is good. The dictionaries outperform their Polyglot counterparts, and are able to correctly detect a negativity bias, which is stronger for tabloids. The method is resource-efficient in terms of manual labor when compared to manually constructed dictionaries, and requires a limited amount of computational power.

Introduction

The availability of large amounts of text data, cheap computing power, and an abundance of analytical tools have all led to a rising interest in automated text analysis methods. Despite this interest, the development of automated methods is far from finished. Using automated methods to determine the positive or negative tone (sentiment) in the individual sentences of a newspaper article remains a major challenge. Manually created sentiment dictionaries (e.g. Soroka et al., 2015; Young & Soroka, 2012) are often used, but have been shown to perform badly when compared to the gold standard of human coding (Boukes et al., 2020). These dictionaries all consist of words that are manually selected based on human expertise, which is a time-consuming process. Computer-generated dictionaries are more time-efficient, and seem to work slightly better than manually constructed ones. Evaluation is however conducted based on simplified tasks such as distinguishing 'clearly positive [. . .] and clearly negative headlines' (Khoo & Johnkhan, 2018, p. 505). Even when using supervised machine learning to classify levels of negativity in parliamentary speeches, performance does not reach higher than .61 (F_1-score) (Rudkowsky et al., 2018). These examples show there is still a lot of room for improvement when it comes to analyzing sentiment in media texts.

Rheault et al. (2016) provide a method for such an improvement, with their application of a word embedding model in combination with a small dictionary of positive and negative 'seed words'. However, they do not apply their method to political news, and validate the performance of their method in a different domain (movie reviews) than the domain that is of substantive interest to them (political speeches). The main question is therefore if their method can be successfully applied to the domain of political news in multiple languages, and at the level of individual sentences instead of documents/newspaper articles.[1]

To answer this question, a dataset containing two decades of newspaper articles in four languages (Danish, Dutch, English, Norwegian) is used. These articles are taken from three different newspapers for each language (six for Dutch/Flemish). Word embedding models are constructed for each of these languages and used to generate sentiment dictionaries based on a small list of positive and negative seed words, replicating to a large extent the method described by Rheault et al. (2016). Concurrent validity is evaluated by comparing the dictionary-based classification to the gold standard of human-coded sentiment. Predictive validity is evaluated by testing two long-standing hypotheses concerning the negativity bias in political news. Finally, the performance of the method is compared to the performance of the Polyglot sentiment dictionaries (Chen & Skiena, 2014), which is one of the best performing dictionaries in the comparison by Boukes et al. (2020).

Sentiment Analysis

While there are many aspects to sentiment, it can generally be described as the 'attitude towards a particular target or topic' (Mohammad, 2016, p. 201). These attitudes are either evaluative or emotional in nature. Evaluative attitudes are based on a simple one-dimensional scale for judging whether something is 'good' or 'bad'. Emotion, on the other hand, is a multidimensional concept, making it much harder to measure using automated methods than one-dimensional evaluative attitudes. Even so, the evaluative aspect of attitudes still remains hard to analyze in an automated fashion.

For one, it is hard to determine the source and target of an evaluation. Semantic role labeling provides a possible solution for this issue, by aiming to extract source-subject-predicate structures from a sentence. One way in which this can be done is by using the syntactic dependencies between words (Shi et al.,2020), as van Atteveldt et al. (2017) successfully do. When the source, subject and predicate in a sentence are known, this information can be used to conduct stance detection. The goal of stance detection is to determine the evaluative stance of the source towards the subject, based on the predicate. However, the stance of a source cannot be directly derived from the words that are used. A negative statement might still contain a positive evaluation, such as in the sentence 'I am sad that Hillary lost this presidential race' (example from Aldayel & Magdy, 2021, p. 5). While this statement is negative, the implicit evaluation of the target (Hillary) by the source (I) is positive.

The above relates closely to the difference between evaluation and valence, or between 'good' and 'bad' versus 'positive' and 'negative'. The former depends on perspective, what is good for somebody can be bad for somebody else. The latter disregards perspective, and is solely based on the inherent positive or negative connotation present in a word or sentence. As this paper is concerned with the creation of sentiment dictionaries, the only aspect of sentiment that can be investigated is valence. And while valence can be combined with semantic role labeling, the election example from Aldayel & Magdy (2021) illustrates that even then it is not always possible to reliably determine stance. Thus, the operational definition of sentiment in this paper is limited to the sum of the positive and negative connotations of words in a sentence.

When using a dictionary for sentiment analysis, there are two main aspects to consider: 1) the construction and content of the dictionary, and 2) the specific domain to which it is going to be applied. Constructing a suitable sentiment dictionary for a specific task is complex, as words have different meanings in different domains. Thus, a sentiment dictionary needs to be domain-specific to some extent (Young & Soroka, 2012). Muddiman et al. (2019) show that manually constructed dictionaries work quite well when they are applied in a very specific domain. Boukes et al. (2020) however show that when using sentiment dictionaries in a more general way (i.e. applied to multiple newspapers, on a general (economy) topic), none of the tested dictionaries perform particularly well. This illustrates the tradeoff in dictionary construction between specificity and general applicability.

Manually constructing dictionaries is a time-consuming process because of this tradeoff, and automating the process of dictionary creation can save valuable time. An additional advantage of automation is that the dictionary can be based on the corpus to which it will be applied, ensuring at least some level of balance between applicability and domain-specificity. One way to construct a computer-generated dictionary is by expanding a short list of positive and negative seed words to a full dictionary through a word embedding (WE) model (Rheault et al., 2016). WE models (Mikolov et al., 2013; see Almeida & Xexéo, 2019 for an overview) make use of the distributional hypothesis, 'a word is known for the company it keeps' (Firth, 1957), to construct a multi-dimensional vector space in which each word is positioned based on its co-occurrences with other words. The assumption is that the dimensions in this vector space represent different latent aspects of meaning (Mikolov et al., 2013), implying that words that are close together in one or more of these dimensions share to a larger or smaller degree their semantic meaning with neighbouring words.

Considering that words with similar meaning occur closely together, it is possible to construct a sentiment dictionary using the words that are closest to the positive and negative words in the seed dictionary (Rheault et al., 2016). Of course, the words in the seed dictionary need to be positive or negative in all possible semantic contexts. Otherwise, the words most closely associated with an ambiguous seed word will also contain ambiguous meaning/sentiment, and thus cause a bias in the final dictionary.[2] Assuming bias is absent from the seed dictionary, the procedure described here allows for the creation of domain-specific sentiment dictionaries as defined by Young & Soroka (2012) from any large corpus of documents.

Assuming that a sentiment dictionary is domain-specific to the data it is applied to, the next question is to which unit of text it should be applied. Ideally, the units of text being analyzed contain only information needed to answer the research question, without any noise. This is of course not realistic, especially when considering that a single newspaper article generally contains references to multiple topics, events, and/or actors. Because each of these subjects are associated with their own sentiment, it makes sense to only analyze those parts of an article that actually relate to the subject of interest. This trend is also visible in media studies, shifting from documents as the unit of analysis (e.g. Bleich & van der Veen, 2018; Young & Soroka, 2012) to smaller units, such as sentences or headlines (Boukes et al., 2020; Khoo & Johnkhan, 2018; Rudkowsky et al., 2018; van Atteveldt et al., 2021). These smaller units are less likely to contain multiple subjects, and make it possible to determine only the sentiment in close proximity to the subject(s) of interest. If document-level metrics are required for further analyses, the scores of individual sentences can be aggregated into sentence groups, providing a sentiment score at the document level. As such, there are no theoretical downsides to analyzing sentiment at the sentence level. From a methodological perspective there is the downside of increasing the complexity of the analyses. But considering that more precise measures are generally preferred, this is an acceptable tradeoff. The question that remains is how well a WE sentiment dictionary applied to newspaper sentences works when compared to human coding.

> **RQ$_1$:** *How well do sentiment dictionaries based on word embeddings and seed dictionaries perform, when compared to human expert-coding?*

Another question is to what extent the proposed method outperforms other dictionary approaches. To test this, the performance of the WE dictionaries is compared to that of the Polyglot (Al-Rfou et al., 2013) sentiment dictionaries in Dutch, Danish, English and Norwegian (Chen & Skiena, 2014). These dictionaries, like the whole Polyglot project, are based on the most frequently used words in Wikipedia articles from a specific language. These words are used to construct a huge network of one- and bi-directional semantic links between words. By propagating from selected seed words (much as in the approach above), the final sentiment dictionary in each language is constructed. Boukes et al. (2020) show that the Polyglot dictionary is one of the best performing dictionaries for detecting positive and negative sentiment in Dutch economic news headlines.

> **RQ$_2$:** *How well do sentiment dictionaries based on word embeddings and seed dictionaries perform, when compared to the Polyglot sentiment dictionaries?*

Negativity bias

In addition to evaluating the concurrent validity of the WE dictionaries through the research questions formulated above, the concept of negativity bias is used to assess the predictive validity of the method. Negativity is a predominant feature of political news (see Lengauer et al., 2012 for an overview), which should be easily detected by the dictionaries. Furthermore, theories on hard versus soft news provide a clear expectation regarding the amount of negativity present in tabloid and broadsheet newspapers, as soft news is a hallmark of tabloid journalism (Otto et al., 2017). It is characterized by a focus on author opinion (Glogger, 2019) and emotion (Reinemann et al., 2012). Combined with the negativity bias in political news, it is therefore likely that tabloid newspapers are more negative in their coverage of political news than broadsheet newspapers.

> **H_1**: *Sentiment will be more negative than positive in political news coverage*

> **H_2**: *Sentiment will be more negative in tabloid newspapers than in broadsheet newspapers*

Data & Methods

In table 1, an overview is presented of the newspaper data used for each language, which runs from January 2000 until December 2019 unless otherwise noted. The division between left-wing, right-wing and tabloid newspapers is based on De Vreese et al. (2016).

In order to get a usable sentiment dictionary, the raw data is processed in five steps: 1) The raw newspaper articles are pre-processed, 2) the processed

Table 1. Newspaper sample

	Left-wing	Right-wing	Tabloid/Popular	Total articles[1]
Danish	Politiken	Jyllands-Posten	Ekstra Bladet	2.08
Dutch(NL)[2]	de Volkskrant	NRC Handelsblad	de Telegraaf	2.16
Dutch(BE)	de Morgen	de Standaard	Het Laatste Nieuws	2.18
English[2]	The Guardian	The Daily Telegraph[3]	The Sun	5.12
Norwegian	[4]	Aftenposten	VG / Dagbladet	2.28

Note:
[1] In millions
[2] Until December 2018
[3] From January 2001
[4] Due to lack of suitable data, Dagbladet as substitution

articles are used to create a (GloVe) word embedding model, 3) from the raw articles, sentences for validation are extracted and manually coded, 4) the word embedding model is combined with the seed word dictionary to create an expanded sentiment dictionary, 5) the validation data and expanded sentiment dictionary are combined to optimize the dictionary through feature selection and tuning the interpretation of the raw sentiment scores. These five steps are elaborately described below, followed by a short summary. The general steps involved in the dictionary expansion process (steps 4 and 5) are visualized in figure 1.

Pre-processing

In the first step, the raw newspaper articles are pre-processed for use in a word embedding model. The complexity of the articles is reduced by using Natural Language Processing (NLP) to convert inflected word forms to their dictionary lemmas. UPOS (Universal Part-Of-Speech) tags are extracted in this process to allow disambiguation between lemmas that are spelled the same way, but have different meanings (such as 'evening' or 'entrance' as either a noun or a verb). NLP also allows for more accurate identification of sentence borders. For example, periods in abbreviations and initials are not treated as sentence borders. NLP is conducted using the R package UDPipe (Straka & Straková, 2017), in combination with version 2.3 of the Danish DDT, Dutch Alpino, English EWT and Norwegian Bokmål Universal Dependencies Models (Nivre et al., 2018).

After NLP parsing, pre-processing continues with the removal of irrelevant articles, such as articles about sports and cultural events, weather forecasts, etc.[3].] The reason for removing these articles is that they often contain nonnatural language (e.g. solutions to crossword puzzles, sports results and weather forecasts), which can interfere with the construction of a word embedding model. The resulting set of processed articles is used in two ways: 1) to construct a word embedding model that in turn is used to create the sentiment dictionary, and 2) to extract sentences to validate and optimize the final sentiment dictionary.

Creating the word embedding model

In the second step, the lemmas and UPOS tags from the pre-processed articles are used to create GloVe word embedding models (Pennington et al., 2014) for each of the languages. The parameters used to generate these models are kept the same as the ones used by Rheault et al. (2016), because the goal of the study is to replicate their approach in a different domain and in different languages. Another reason for not optimizing the model parameters further is because

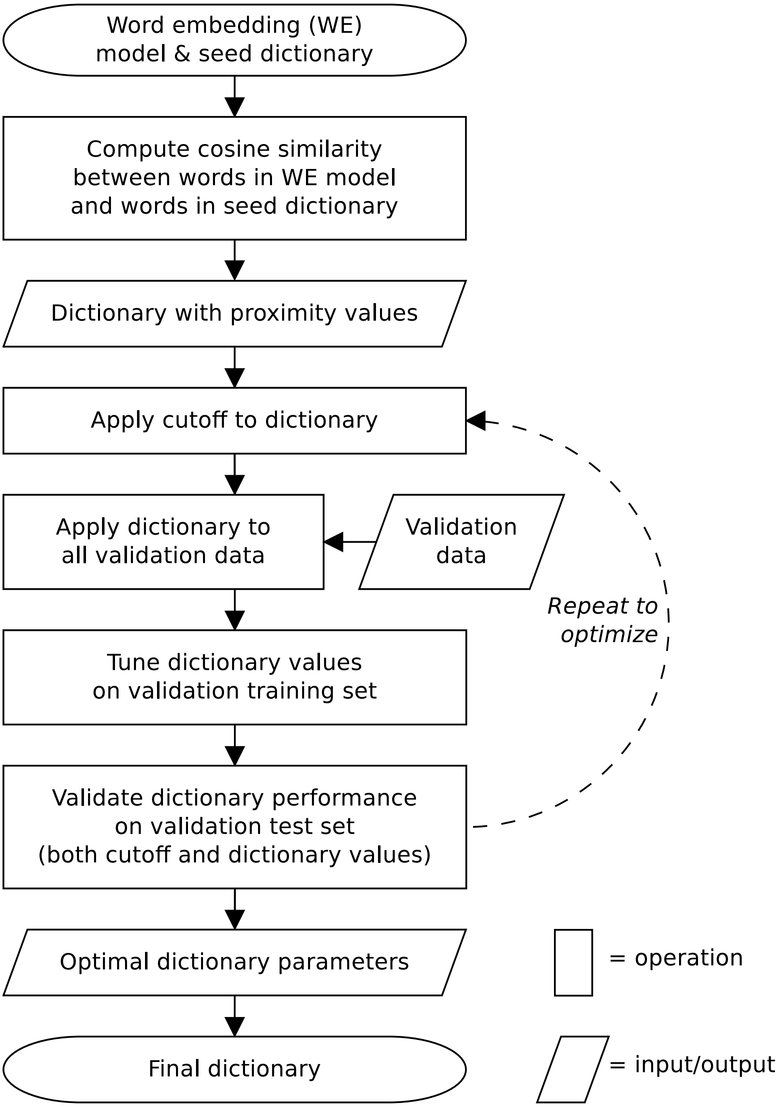

Word embedding (WE) model & seed dictionary
Compute cosine similarity between words in WE model and words in seed dictionary
Dictionary with proximity values
Apply cutoff to dictionary
Apply dictionary to all validation data
Validation data
Repeat to optimize
Tune dictionary values on validation training set
Validate dictionary performance on validation test set (both cutoff and dictionary values)
Optimal dictionary parameters
= operation
Final dictionary
= input/output

Figure 1. Data flow chart

of the computational complexity of estimating these models, especially in four languages. The parameters used are as follows: 1) tokens (a group of characters separated by whitespaces) that occur less than 5 times in the corpus are filtered out, 2) the symmetric token window size is set to 7 tokens,

Table 2. Corpora details

	Danish	Dutch[1]	English	Norwegian
Documents (x million)	1.15	2.39	2.25	1.31
Tokens (x million)	463.08	891.75	896.39	442.35
Vocabulary (x 100,000)	799.78	1176.31	615.23	763.58

Note:
[1] Both Dutch and Flemish data

meaning that the 7 tokens preceding and the 7 tokens following a token are considered as co-occurrences, 3) tokens are positioned in a 300-dimensional vector-space. Based on these model parameters, the word embedding model is estimated over 100 iterations, after which tokens occurring less than 20 times in the corpus are removed from the models. Corpus stastistics (with all tokens included) for the different languages are presented in table 2.

Creating validation data

In the third step, the articles that remain after pre-processing are filtered. Only political news articles are kept to answer the research questions and test the hypotheses, by removing any articles that do not mention at least one political actor. This is done by querying the articles at sentence level for the presence of both parties and/or individual actors (MPs, party leaders and (prime) ministers). The queries are date-limited, so that actors are only included on dates that they were actually active. Details relating to the political actor queries, and how they were constructed and executed, can be found in the appendix. All queries combined result in a total of 264,141 (BE), 309,701 (DK), 247,702 (NL), 237,244 (NO) or 512,180 (UK) newspaper articles in which one or more political actors are mentioned.

A random subset of these articles is sampled for individual sentences containing one or more political actors. These sentences are manually coded to construct a validation dataset for the sentiment dictionary.[4] They are coded by student assistants based on the following question: 'How would you describe the overall tone expressed in this sentence?' The answer options are negative, neutral/absent and positive. Training and intercoder reliability testing is done by the principal researcher in each country/language, and one (or two, in Denmark) student coders per language. Note that the coders are explicitly instructed to evaluate the valence of the sentence (its connotation), rather than the stance held towards or by the actor. During the final intercoder reliability test, 50 sentences are coded by both the researcher and the student assistant(s), after which the student codes the remaining

Table 3. Validation details

	Danish	Dutch[1]	English	Norwegian
Hand-coded sentences	3187	3538	4569	3933
Coding time[2]	7	7	14	8
Intercoder reliability[3]	0.75	0.84	0.71	0.79

Note:
[1] Both Dutch and Flemish data
[2] Median time per sentence, in seconds
[3] Using Krippendorff's alpha

sentences. English is an exception, as in contrast to the other languages sentences are not coded by native speakers, but rather Norwegian coders that are fluent in English. In table 3 the number of coded sentences, median coding time per sentence, and the intercoder reliablity are shown for each language. All student coders have been paid for their work.

Expanding the dictionary

In the fourth step, a measure is constructed to indicate the proximity of all words in the word embedding model to the seed dictionary. The seed dictionary is taken directly from Rheault et al. (2016), to replicate their method in the domain of political news. As the original seed dictionary is only in English, it is manually translated to the other three languages. The main goal during this translation process is to stay as close as possible to the original literal meaning of the English seed words as possible. While this ensures that the seed dictionaries in different languages are as similar as possible in a literal sense, it also opens up room for differences in the semantic meaning of the translations. This tradeoff is considered worthwhile, as the current process requires comparatively little human labor. In addition, recent work by Proksch et al. (2019) shows that automatic (Google Translate) translations of sentiment dictionaries perform remarkably well, illustrating the limited impact of literal translations. The full seed dictionaries for all four languages can be found in the appendix.[5]

In figure 1, the (translated) seed dictionaries and word embedding models are used as input to cacluclate the proximity of all corpus words (including the seed words themselves) to the words in the seed dictionary. Proximity is determined for each possible pair of corpus and seed words individually, by computing the cosine similarity of all pairs based on their values on the 300 dimensions of the word embedding model. By subtracting the sum of cosine similarity with the negative seed words from the sum of similarity with the positive seed words a measure is constructed indicating the relative proximity of each word to the positive and negative words in the seed dictionary. These raw values are scaled, but in a slightly different way than Rheault et al. (2016) propose. Rather than scaling the positive and negative values separately, which disregards the relative proximity of positive and negative words to the seed words, all values are scaled by dividing by the maximum absolute value among those values. This results in a dictionary of all words in the corpus with their proximity to the seed dictionary (third step in figure 1). These proximity values are operationalized as sentiment scores, as higher positive values indicate closer proximity to the positive seed words, and higher negative values indicate closer proximity to the negative seed words. Because the values are based on proximity, they also take into account context-related issues, such as negation (i.e. a positive word that is often negated will have a lower proximity to the positive seed words than a positive word that is hardly ever negated).

Creating and validating the final dictionary

In step five, the final sentiment dictionary is created by selecting words from the expanded dictionary created in the previous step, and tuning the interpretation of the proximity values. This process corresponds to all operations following the 'Dictionary with proximity values' input/output in figure 1. Various values, ranging from .15 to .35 in steps of .05, are tested as threshold for the minimum absolute proximity above which words are included in the dictionary. Based on the resulting dictionary, sentiment scores are computed by summing the proximity values of all sentiment words present in a sentence, and dividing that by the total number of words in the sentence. Then, these values are interpreted according to an ordinal scale (negative, neutral, positive), to make them correspond to the manual coding. The cutoff points required to convert the sentence values to ordinal categories are optimized as well, by testing cutoff values between -.1 and .1 in steps of .005 for both the positive and negative cutoff. A simplified example of how the final dictionary is constructed and applied can be found in the appendix.

As both the (absolute) proximity threshold, positive cutoff and negative cutoff are tested concurrently, a total of 8405 parameter combinations is tested using a 5-fold cross-validation approach. The hand-coded sentences are split into five equal parts/folds which are each used once to test the performance of the optimal dictionary parameters, while the other four folds are used to learn the optimal parameters. Thus, the optimal parameters are determined five times on different parts of the validation data. Similarly, the performance of those different parameter sets is also tested each time, and each time on a different part of the total hand-coded dataset.

The optimal parameters for each fold are determined based on the weighted (by the proportion of manually coded sentences in each category) F_1-score. The final performance is determined by taking the mean of all performance indicators over the 5 folds, and used to answer RQ_1 and RQ_2. Then, the optimal parameters are determined based on the whole hand-coded dataset. These parameters are used to classify sentiment for all political news articles (i.e. articles that mention at least one political actor). The sentiment scores for each sentence are aggregated to the document level, and used to test H_1 and H_2.

Summary & Costs

While the method described above is quite specific and elaborate, these steps can be generalized to a substantial extent. Most importantly, the method can be used to classify different aspects of texts than sentiment (e.g. topics), simply by using seed dictionaries with different words (see also Amsler, 2020). Assuming a corpus of texts, a word embedding model (ideally constructed from the corpus), a seed dictionary and a validation dataset, there are only two steps required to construct the final dictionary (see also figure 1). 1) Determine the optimal proximity value above which words should be included in the dictionary, and 2) determine how high the sum of the word values needs to be in order to consider a concept (e.g. topic, frame, etc.) as being present in a text/document. Both should ideally be done by using a human-labeled validation set. One might notice these steps are somewhat different from the procedure above, where positive and negative sentiment is measured using a single seed dictionary. However, two separate seed dictionaries are effectively used, and the proximity to the negative seed dictionary is subtracted from the proximity to the positive seed dictionary, to construct a single measure. This can be done with any one-dimensional concept.

To estimate the word embedding models for this study, a 16-core server with 32GB of RAM was used. The models took in total around 34 hours to estimate. The costs to compute these models was around $3.40. Of course, hand-coding of sentences used for validation is significantly more expensive. The median time for student assistants to code one sentence is between 7 and 14 seconds (see table 2). Rounding the coding time per sentence up to 15 seconds, it would take ~17 hours to code 4000 sentences per country. When including an additional 10 hours per language for training the student assistants, and assuming a wage of $15 per hour, the total costs of hand-coding

Table 4. Dictionary parameters and size

	Proximity[1]	Positive sentiment[2]	Negative sentiment[2]	Negative words	Positive words	Total words
Danish	0.30	0.03	0	11352	10211	21563
Dutch	0.30	0.04	0.005	12911	13690	26601
English	0.20	0.03	0.005	12442	10458	22900
Norwegian	0.25	0.05	0.010	14149	13994	28143

Note:
Zero and values between the positive and negative sentiment cutoffs are interpreted as neutral
[1] Minimum required proximity between word and positive/negative seed dictionary
[2] Values above/below which sentiment is interpreted as positive/negative

all four languages is $1620. As is shown below, the costs can however be even lower, as the method described here also works with a smaller hand-coded dataset. And when coding in their native language student coders are almost twice as fast as assumed here.

Validating a computer-generated sentiment dictionary

To answer both research questions, the sentiment dictionaries in each language are optimized and validated through the use of a hand-coded validation dataset. In this process two sets of parameters are tuned: 1) the threshold for including words in the dictionary (i.e. the minimal proximity score a word must have to the seed words in order to be included in the dictionary), and 2) the positive and negative cutoffs for converting the sentiment scores to categories. The cutoffs for the minimum proximity, minimum positive sentiment score, and maximum negative sentiment score (excluding 0, which is always considered as no sentiment) are presented in the first three columns of table 4. Using the optimal proximity cutoff, the final three columns of table 4 show the number of positive and negative words, as well as the total size of the dictionaries.[6]

The stability of the three dictionary parameters is tested for all languages using smaller sample sizes of 100, 500, 1000 or 2000 sentences. While these samples in some cases result in slightly different optimal dictionary parameters, the general performance of the dictionaries remains stable when using 1000 sentences or more to optimize the parameters. Smaller sample sizes tend to produce diverging dictionary parameters, and with the exception of Danish an overestimation of

Table 5. Classification performance for Polyglot and word embedding dictionaries

	F1		**Precision**		**Recall**		**# of sentences**		
	Poly.	WE	Poly.	WE	Poly.	WE	Poly.	WE	Human
Danish									
Negative	0.46	0.61	0.57	0.67	0.39	0.57	743	922	1081
Neutral	0.54	0.69	0.61	0.64	0.48	0.74	1291	1901	1659
Positive	0.33	0.40	0.23	0.45	0.59	0.36	1153	364	447
Weighted average	*0.48*	*0.62*	*0.54*	*0.62*	*0.46*	*0.63*	*3187*	*3187*	*3187*
Dutch									
Negative	0.36	0.51	0.37	0.56	0.35	0.46	816	718	863
Neutral	0.35	0.77	0.66	0.71	0.24	0.83	817	2613	2246
Positive	0.24	0.24	0.15	0.37	0.65	0.18	1905	207	429
Weighted average	*0.34*	*0.64*	*0.53*	*0.63*	*0.32*	*0.66*	*3538*	*3538*	*3538*
English									
Negative	0.61	0.66	0.70	0.72	0.53	0.61	1485	1637	1948
Neutral	0.55	0.61	0.55	0.57	0.56	0.66	1828	2094	1812
Positive	0.44	0.48	0.36	0.47	0.56	0.49	1256	838	809
Weighted average	*0.56*	*0.61*	*0.58*	*0.62*	*0.55*	*0.61*	*4569*	*4569*	*4569*
Norwegian									
Negative	0.46	0.56	0.42	0.62	0.52	0.51	1507	989	1207
Neutral	0.52	0.71	0.65	0.65	0.44	0.78	1420	2526	2108
Positive	0.37	0.39	0.30	0.49	0.49	0.33	1006	418	618
Weighted average	*0.48*	*0.61*	*0.52*	*0.62*	*0.47*	*0.63*	*3933*	*3933*	*3933*

Note: Poly. and WE refer to the Polyglot and word embedding dictionaries respectively

the dictionary performance. Upsampling is explored as a method to counteract the obvious class imbalances when classifying newspaper sentiment. However, without any exceptions, the dictionaries constructed using upsampling resulted in worse average performance (weighted F1-score) than when using dictionaries based on unbalanced samples. Thus, balancing the classes in this way does not increase the performance of the method.[7]

In table 5, the mean sentiment performance metrics for the word embedding (WE) dictionaries are shown per language, alongside the performance of the respective Polyglot dictionaries.[8] As the results show, the WE dictionaries in different languages perform comparably on average, despite the differences in size and balance between the positive and negative words. The performance of individual categories in each language is clearly related to their prevalence (i.e. the most frequently occurring category performs best, the least frequently occurring category

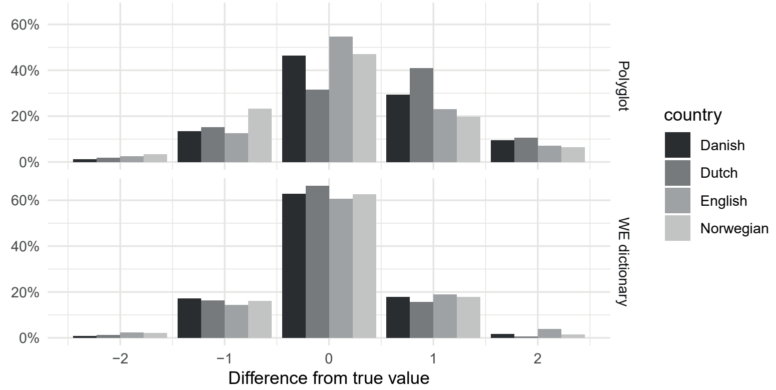


Figure 2. Difference between predicted and true sentiment category

performs worst). Generally speaking, the neutral category thus performs best, followed by the negative category, and then the positive category. The one exception in this is English, where the negative category is both slightly larger and performs slightly better than the neutral category. The weighted average F_1 scores for each of the languages, ranging between .61 and .64, are not high enough for detailed substantive research at the level of individual sentences. However, it can be argued that errors might cancel each other out when the sentiment of individual sentences is aggregated.

To test this assumption, figure 2 shows the distribution of errors between different classes for both the Polyglot and WE dictionaries[9]. The values indicate the number of steps/categories the predicted category is from the true value (e.g. +2 means the prediction is positive while the true value is negative, while -2 indicates the inverse). The 0 category thus shows the accuracy of the dictionaries. In all cases, the accuracy of the WE dictionaries is above 60%, while the Polyglot dictionaries fail to reach higher than 50% accuracy, with the exception of the English dictionary (55%). Besides making less mistakes in general, the severity of the mistakes is also lower with the WE dictionaries than with Polyglot. The vast majority of the WE errors fall within the +/-1 categories, indicating that errors between positive/negative and neutral are most common. For Polyglot, there is also a substantial amount of errors that falls in the +2 category. The distribution of errors is also less skewed for the WE dictionaries than for Polyglot, indicating that errors will cancel each other out to a larger extent in the former than in the latter. These results support the assumption that aggregation will improve the performance of the method. So while the weighted F1-scores of the WE dictionaries are too low for detailed analyses, the method is suitable for aggregate-level analyses, providing a clear answer to RQ_1.

Based on the average F_1 scores, the WE dictionaries outperform the Polyglot dictionaries by a substantial margin in all languages, except English. In the latter case, the performance advantage of the WE dictionary is still present, but less pronounced. Looking at the F_1 scores of individual categories, the same picture emerges, regardless of the language. Only in the Dutch positive category does the Polyglot dictionary perform on-par with the WE dictionary, and both perform equally bad. In general, the difference in performance between Polyglot and the WE dictionaries is smallest for the positive categories. This is caused primarily by the recall of the positive category being higher for Polyglot than the WE dictionaries in all languages, meaning that Polyglot captures a larger percentage of the human-coded positive sentences. This is however the only point where Polyglot outperforms the WE dictionaries. These results provide a clear answer to RQ_2, as the WE dictionaries perform substantially better than the alternatives provided by Polyglot. The stable performance between languages also shows that the WE approach is especially suitable for comparative research.

Investigating actor sentiment

The predictive validity of the WE sentiment dictionaries is evaluated by testing for the well-established presence of negativity bias in political news (H_1), and the hypothesis that this bias is stronger in tabloid newspapers than in broadsheet newspapers (H_2). For each country, figure 3 shows the average sentiment over time in tabloid and broadsheet newspapers (see table 1 for details). All plots are smoothed using a LOESS function with a span of .25 and the gray bands indicating the 95% confidence interval of the standard error. Descriptive statistics of the sentiment scores for the different newspapers can be found in the appendix. The sentiment shown in figure 3 is negative throughout the whole period in all countries, as is illustrated by the y-axis not reaching higher than -.1. Unsurprisingly, the mean sentiment scores per newspaper (see appendix) are also negative in all cases. Both results provide clear evidence for the presence of a negativity bias in political news, confirming H_1.

The graphs in figure 3 also show that the tabloid newspapers are generally speaking more negative in their coverage than the broadsheets. This difference is most pronounced in the UK, while it is moderate in Denmark,

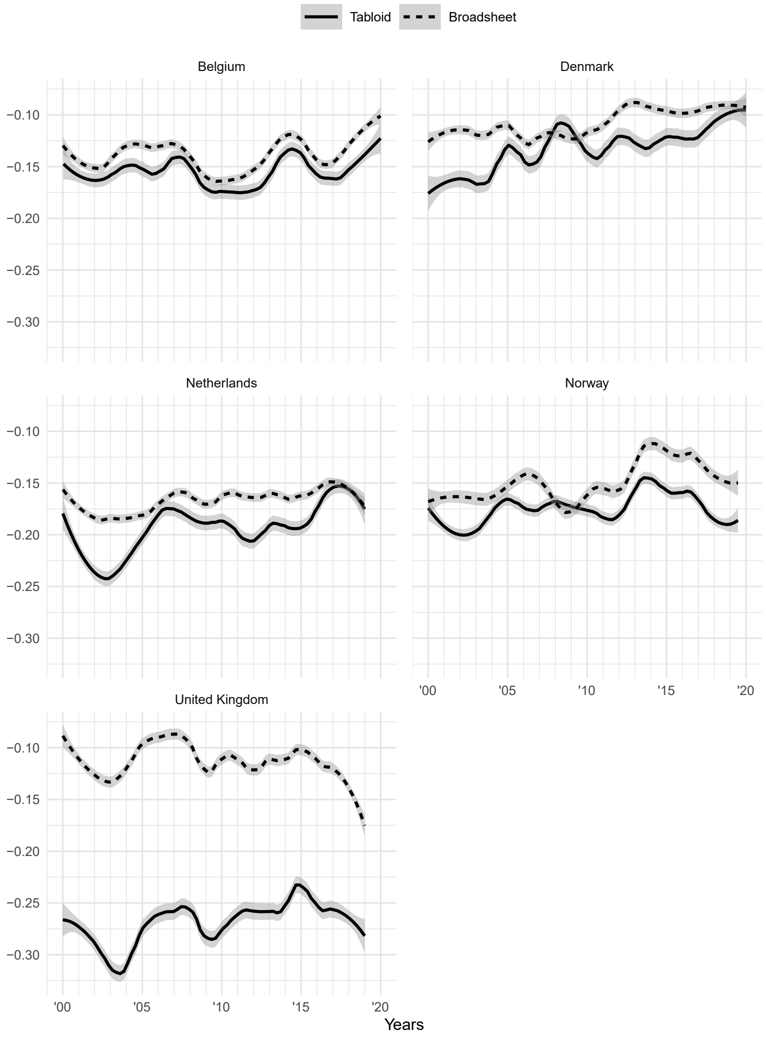


Figure 3. Sentiment by newspaper

The Netherlands and Norway. In Belgium, the difference in sentiment between the tabloid and broadsheets is still significant, but limited in size. These interpretations of the graphs are supported by the unstandardized regression results in table 6. Time is included in these models as a control variable to account for over-time developments, and while it is a significant predictor, the effect sizes are negligible in all countries. The tabloid dummy is also highly significant, and its effect sizes are at least an order of magnitude larger than those of time. Even so, the effect sizes are marginal at best, with the exception of the UK, where a stronger effect is visible. This does however not impede the testing of H_2, as it only concerns the direction (and not the strength) of the effect. Therefore the significance and negative value of the tabloid variable provides enough support to confirm H_2.

That being said, the fluctuations in sentiment that are visible in figure 3 do not really align between tabloids and broadsheets, except in Belgium and, to a lesser extent, the UK. The absence of a clear relation between tabloid and broadsheet sentiment illustrates that tabloids offer substantially different content, with substantially different sentiment, than broadsheet newspapers. The stronger relationship between tabloid and broadsheet sentiment in Belgium can be explained by the highly concentrated ownership in the Flemish newspaper market. For the UK, the explanation is not as apparent, but a tentative explanation might be that in the more professionalized and market-driven media system of the UK (see Hallin & Mancini, 2004), newspapers in general follow the sentiment of the general public, with the only difference being that tabloid newspapers are more expressive/sensational in their sentiment than broadsheets. The difference in media system also provides a tentative explanation for the relatively large difference in negativity between broadsheets and tabloids in the UK, when compared to the other countries.

Regardless of these differences, the impact of the start of the economic crisis around 2008 is clearly visible in all countries, although in Denmark the impact is most visible in tabloid coverage, while in Norway it is more pronounced in the broadsheet newspaper. Besides the onset of the economic crisis, no other trends are clearly shared between the countries. At the national level, however, the most striking trend is the increase in negativity following the Brexit referendum in the UK. The fact that both the economic crisis and Brexit referendum are clearly reflected in the sentiment of newspapers provides additional support for the validity of the sentiment measure.

Conclusion

The results presented in this paper illustrate the advantages of generating a custom sentiment dictionary based on a word embedding model and a

Table 6. Regression results by country

	Dependent variable:				
			Sentiment		
	Belgium	Denmark	Netherlands	Norway	UK
Tabloid (dummy)	−.0166***	−.0255***	−.0255***	−.0241***	−.1498***
	(.0010)	(.0012)	(.0009)	(.0011)	(.0011)
Time (in years)	.0007***	.0020***	.0020***	.0016***	−.0004***
	(.0001)	(.0001)	(.0001)	(.0001)	(.0001)
Constant	−.1432***	−.1256***	−.1833***	−.1644***	−.1084***
	(.0009)	(.0010)	(.0009)	(.0012)	(.0011)
Observations	264,141	309,701	247,702	237,244	512,180
R2	.0013	.0027	.0047	.0029	.0331
Adjusted R2	.0013	.0027	.0047	.0029	.0331
Residual Std. Error	.2249	.2816	.2179	.2703	.3643
F Statistic	172.4721***	417.3835***	588.7979***	342.2341***	8,780.0050***

Note: *p<0.1; **p<0.05; ***p<0.01

limited set of seed words. As shown in the validation section, WE dictionaries perform adequately when classifying sentiment in individual sentences, when compared to human coding. This human coding is also leveraged to improve dictionary performance by optimizing the selection of words included in the dictionary, and by tuning the interpretation of the raw sentiment scores when converting them into categories. Comparing the performance of the WE dictionaries to the well-established (see e.g. Boukes et al., 2020; van Atteveldt et al., 2021) Polyglot sentiment dictionaries, it is clear that the method described in this study provides a substantial improvement, especially in languages other than English. A likely explanation for this difference can be found in the data both methods are based on. The WE dictionaries are created specifically from the data (newspaper articles) to which they are applied, while the Polyglot dictionaries are based on the more formal language of Wikipedia articles. Another advantage of the WE method is the relatively stable performance across different languages, making it especially suitable for comparative research. This conclusion is further reinforced by the correct detection of a negativity bias in political news in all five countries, which is stronger in tabloids than in broadsheets.

That being said, the performance of the custom dictionaries when compared to human coding still leaves room for improvement. Specifically machine/deep learning methods seem to be capable of outperforming the

WE dictionaries (e.g. van Atteveldt et al., 2021, p. 128, Table 2). The use of sentences as unit of analysis and the essentially random classification errors from the WE dictionaries however make it likely that the performance of these dictionaries will be higher on the document level than on the sentence level. So even though the performance might not yet be high enough for valid sentence-level analyses, the method is performing well enough when analyzing aggregated data. There are also ample options for improvement of the method. For example, using more advanced sampling techniques to deal with the inherent class imbalances in the sentiment of political news. Or using separate cutoffs for the inclusion of positive and negative dictionary words, optimizing the seed dictionary further, and investigating which words in the dictionary most often cause errors in classification. On a more fundamental level, and assuming the availability of sufficient computing power, the method can also be further optimized by explicitly validating different sets of parameters used for generating word embedding models.

While this study presents a single, weakly supervised approach to extend a (sentiment) seed dictionary, there are many related ways to expand seed dictionaries. For example, the doctoral dissertation of Michael Amsler (2020) describes a similar but far more elaborate algorithm than the one used here. Notable differences are an iterative approach to dictionary expansion, and an extensive evaluation of the cosine similarity relationships between newly suggested words and words that are already part of the dictionary. As a result, the algorithm uses the cosine similarity between individual words, rather than the similarity to the entire pool of words in a seed dictionary. What lacks in this approach, is a point where human input can be effectively leveraged. Other studies (e.g. Alba et al., 2018; Makki et al., 2014) do make use of human input to expand their dictionaries. For example by determining the words that are most similar to a seed dictionary in a word embedding model, let humans evaluate which of those most similar words are most suitable to be included in the seed dictionary, expanding the seed dictionary and then repeat the process (Alba et al., 2018). This approach differs in its application of human labor from the current study, as it directly evaluates words, rather than providing labeled examples. This has the upside of directly (instead of indirectly) evaluating dictionary words, but also comes with the downside that the approach can become quite labor-intensive when a large vocabulary needs to be evaluated. Yet another approach is suggested by Alhothali & Hoey (2017), who combine word embeddings with pre-existing semantic resources, such as WordNet. The rationale here is that datasets containing synonyms and/or antonyms can by themselves be used for dictionary expansion, and that the semantic proximity of words in a vector space can be leveraged to improve this rule-based dictionary expansion method. An upside of this approach is that it is unsupervised (like the one described by Amsler, 2020), while the reliance on external linguistic resources limits its application to languages for which such resources are available.

Although there is room for improvement, the results presented here illustrate three main points. Firstly, it is possible to analyze sentiment at sentence (instead of document) level with reasonable accuracy, illustrating the opportunities for creating more fine-grained sentiment analysis methods in the future. Secondly, the costs of the WE dictionary approach are relatively low. The costs for constructing and optimizing a dictionary for a single country remains well below $500, with the amount of required hand-coding being limited to around 2000 sentences. In addition, the computational requirements are modest, when using corpora of sizes similar to the ones used here. And while there are substantial costs associated with the construction of the corpora used in this study, those costs do not relate exclusively to the method described here. Cleaning and NLP parsing of a corpus is a worthwhile investment for all kinds of automated text analysis methods. Finally, and perhaps most importantly, the results show there is still room for dictionary-based approaches in automated sentiment analysis, and there is no longer a need to manually create such dictionaries when working with sufficiently large data sets.

Supplementary Materials

Irrelevant article coding procedure

To classify irrelevant articles, around 12,000 news articles have been hand-coded in English, and between 6,000 and 7,000 in Danish, Dutch and Norwegian. The reason for the difference between English and the other languages is because similar classification performance for all countries needs to be obtained, and this required more data in English than the other languages. Student assistants have classified these articles based on the categories “Culture/art events and entertainment,” “Sporting events and athletes” and “Miscellaneous.” If articles fall into any of these three categories, they are considered irrelevant, if not, they are relevant. The miscellaneous category contains all articles that cannot be classified in any of the other categories in the codebook. The hand-coded articles are then used as input for a multinomial Naive Bayes classifier. The input features for this model are the tf-idf weighted lemmas and UPOS tags generated in the

NLP procedure described in the paper. The “format” of each word/feature in an article becomes lemma_UPOS. For getting the best-performing model for each country, a 3 by 5 nested cross-validation procedure is used, with the 3 outer folds being used for performance estimation of the final model, and the 5 inner folds of each outer fold being used for parameter optimization. In this case, parameter optimization consists of only a single parameter, for feature selection. Features are selected based on the chi2 measure to determine which features are most and least strongly associated with the “irrelevant” topic. Using the absolute chi2 values, the top x-th percentile of features are kept to construct a model.

Through the nested cross-validation procedure described above, the optimum cutoff values for feature selection are determined as follows: 0.99 (BE), 0.995 (DK), 0.996 (NL), 0.994 (NO), 0.994 (UK). Using these parameters, the final models achieve a precision of between 0.87 (DK) and 0.94 (UK). Precision is used as optimization measure to avoid as much as possible that relevant articles are classified as irrelevant, allowing for some relevant articles to remain in the relevant articles category. Other performance measures can be found in table 2.

Actor query construction and execution

Data for political parties is collected using case-sensitive queries on either the full party name, or the most commonly used party abbreviations. When necessary, special characters like opening and closing brackets for the abbreviations (con) and (lab) in the UK, are also taken into account. In Norway, several of the major political parties have single letter abbreviations. In these specific cases, regular expression filters are used to filter out common mistakes, like V (the abbreviation for the left-wing party Venstre) as a roman number 5 in the names of monarchs.

Queries for individual politicians (ministers, party leaders and MPs), are constructed by looking for the combination of the (first) given name and surname within 5 words of each other. A larger distance between the two would result in too many false positives, and a smaller distance in too many false negatives. The queries are also limited to articles published during the time the politician was in office. For ministers the queries include their formal title as an alternative for their given name (e.g. both Secretary Johnson and Boris Johnson are valid hits).

Tables

Table 1. Sentiment descriptives by newspaper

Newspaper	Mean	SD	Median	N
The Daily Telegraph	-.127	.365	-.124	165829
The Guardian	-.101	.352	-.100	202538
The Sun	-.263	.380	-.285	143813
Aftenposten	-.150	.270	-.137	104679
Dagbladet	-.159	.272	-.144	65046
VG	-.187	.269	-.179	67519
De Morgen	-.142	.217	-.131	91661
De Standaard	-.132	.220	-.120	103484
Het Laatste Nieuws	-.153	.242	-.138	68996
Ekstra Bladet	-.131	.291	-.120	67308
Jyllands-Posten	-.106	.285	-.099	133985
Politiken	-.110	.272	-.101	108408
NRC Handelsblad	-.167	.201	-.155	87718
De Telegraaf	-.188	.247	-.183	78311
De Volkskrant	-.163	.207	-.150	81673

Table 2. Irrelevant articles classification performance

	English	Norwegian	Danish	Dutch (BE)	Dutch (NL)
Accuracy	0.873	0.859	0.843	0.865	0.866
Kappa	0.737	0.715	0.685	0.730	0.731
Sensitivity	0.853	0.767	0.801	0.813	0.796
Specificity	0.906	0.944	0.883	0.917	0.934
Pos Pred Value	0.937	0.926	0.865	0.909	0.923
Neg Pred Value	0.788	0.814	0.825	0.828	0.822
Precision	0.937	0.926	0.865	0.909	0.923
Recall	0.853	0.767	0.801	0.813	0.796
F1	0.893	0.839	0.832	0.859	0.855
Prevalence	0.623	0.480	0.484	0.505	0.498
Detection Rate	0.531	0.368	0.387	0.411	0.397
Detection Prevalence	0.567	0.397	0.448	0.452	0.430
Balanced Accuracy	0.879	0.855	0.842	0.865	0.865

Table 3. Optimal dictionary parameters with various hand-coded sample sizes (Norway)

n	Dictionary threshold	Positive cutoff	Negative cutoff	Weighted F1
100	0.20	0.030	0.005	0.6611382
500	0.25	0.035	0.010	0.6338293
1000	0.25	0.050	0.010	0.6212897
2000	0.25	0.050	0.010	0.6259237
3933	0.25	0.050	0.010	0.6141338

Table 4. Optimal dictionary parameters with various hand-coded sample sizes (UK)

n	Dictionary threshold	Positive cutoff	Negative cutoff	Weighted F1
100	0.15	0.045	-0.010	0.6531431
500	0.20	0.035	-0.005	0.6319481
1000	0.25	0.030	0.005	0.6132555
2000	0.20	0.030	0.005	0.6172885
4569	0.20	0.030	0.005	0.6087228

Table 5. Optimal dictionary parameters with various hand-coded sample sizes (DK)

n	Dictionary threshold	Positive cutoff	Negative cutoff	Weighted F1
100	0.30	0.015	0.00	0.6096293
500	0.30	0.030	0.00	0.6175872
1000	0.25	0.035	0.00	0.6154109
2000	0.25	0.050	0.01	0.6181187
3187	0.30	0.030	0.00	0.6222762

Table 6. Optimal dictionary parameters with various hand-coded sample sizes (NL)

n	Dictionary threshold	Positive cutoff	Negative cutoff	Weighted F1
100	0.25	0.045	-0.005	0.7619048
500	0.25	0.060	-0.010	0.6247226
1000	0.30	0.030	0.005	0.6284429
2000	0.30	0.030	-0.005	0.6346448
3538	0.30	0.040	0.005	0.6406772

Table 7. Confusion matrix (WE) with optimal dictionary parame- ters, predictions in rows (Norwegian)

	-1	0	1
-1	613	298	78
0	541	1648	337
1	53	162	203

Table 8. Confusion matrix (Polyglot), predictions in rows (Norwe- gian)

	-1	0	1
-1	630	740	137
0	325	917	178
1	252	451	303

Table 9. Confusion matrix (WE) with optimal dictionary parame- ters, predictions in rows (English)

	-1	0	1
-1	1182	346	109
0	592	1196	306
1	174	270	394

Table 10. Confusion matrix (Polyglot), predictions in rows (English)

	-1	0	1
-1	1040	330	115
0	580	1008	240
1	328	474	454

Table 11. Confusion matrix (WE) with optimal dictionary parame- ters, predictions in rows (Danish)

	-1	**0**	**1**
-1	615	284	23
0	415	1224	262
1	51	151	162

Table 12. Confusion matrix (Polyglot), predictions in rows (Danish)

	-1	**0**	**1**
-1	420	285	38
0	356	791	144
1	305	583	265

Table 13. Confusion matrix (WE) with optimal dictionary parame- ters, predictions in rows (Dutch)

	-1	**0**	**1**
-1	401	271	46
0	441	1866	306
1	21	109	77

Table 14. Confusion matrix (Polyglot), predictions in rows (Dutch)

	-1	**0**	**1**
-1	301	450	65
0	190	541	86
1	372	1255	278

Table 15. Sentiment classification performance (Norwegian, *n* = 100)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.53	0.48	0.60	25	20
Neutral	0.77	0.77	0.77	57	57
Positive	0.49	0.56	0.43	18	23
Combined	0.66	0.66	0.67	100	100

Table 16. Sentiment classification performance (English, *n* = 100)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.64	0.53	0.81	32	21
Neutral	0.72	0.83	0.63	46	60
Positive	0.54	0.50	0.58	22	19
Combined	0.65	0.66	0.68	100	100

Table 17. Sentiment classification performance (Danish, *n* = 100)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.58	0.51	0.68	41	31
Neutral	0.67	0.72	0.63	46	52
Positive	0.47	0.54	0.41	13	17
Combined	0.61	0.61	0.62	100	100

Table 18. Sentiment classification performance (Dutch, *n* = 100)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.57	0.50	0.67	8	6
Neutral	0.86	0.87	0.84	31	32
Positive	0.29	0.33	0.25	3	4
Combined	0.76	0.76	0.77	42	42

Table 19. Sentiment classification performance (Norwegian, *n* = 500)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.56	0.51	0.62	141	117
Neutral	0.74	0.78	0.69	274	311
Positive	0.43	0.40	0.47	85	72
Combined	0.63	0.64	0.63	500	500

Table 20. Sentiment classification performance (English, *n* = 500)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.65	0.57	0.76	194	146
Neutral	0.66	0.73	0.60	212	258
Positive	0.53	0.53	0.52	94	96
Combined	0.63	0.63	0.65	500	500

Table 21. Sentiment classification performance (Danish, *n* = 500)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.60	0.54	0.68	181	145
Neutral	0.68	0.73	0.64	251	285
Positive	0.42	0.43	0.41	68	70
Combined	0.62	0.62	0.62	500	500

Table 22. Sentiment classification performance (Dutch, *n* = 500)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.54	0.49	0.60	145	119
Neutral	0.74	0.82	0.68	294	357
Positive	0.26	0.18	0.46	61	24
Combined	0.62	0.65	0.63	500	500

Table 23. Sentiment classification performance (Norwegian, *n* = 1000)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.55	0.50	0.61	287	236
Neutral	0.73	0.81	0.67	546	664
Positive	0.37	0.30	0.50	167	100
Combined	0.62	0.64	0.62	1000	1000

Table 24. Sentiment classification performance (English, *n* = 1000)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.63	0.57	0.72	418	331
Neutral	0.63	0.68	0.59	411	476
Positive	0.52	0.55	0.49	171	193
Combined	0.61	0.61	0.63	1000	1000

Table 25. Sentiment classification performance (Danish, *n* = 1000)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.57	0.50	0.67	337	253
Neutral	0.70	0.75	0.65	534	620
Positive	0.38	0.38	0.39	129	127
Combined	0.62	0.62	0.62	1000	1000

Table 26. Sentiment classification performance (Dutch, *n* = 1000)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.53	0.51	0.57	281	250
Neutral	0.73	0.77	0.69	590	663
Positive	0.37	0.31	0.46	129	87
Combined	0.63	0.64	0.63	1000	1000

Table 27. Sentiment classification performance (Norwegian, *n* = 2000)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.56	0.51	0.62	594	492
Neutral	0.73	0.80	0.67	1109	1313
Positive	0.37	0.30	0.46	297	195
Combined	0.63	0.64	0.63	2000	2000

Table 28. Sentiment classification performance (English, *n* = 2000)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.65	0.59	0.71	817	677
Neutral	0.63	0.69	0.58	822	970
Positive	0.51	0.51	0.52	361	353
Combined	0.62	0.62	0.63	2000	2000

Table 29. Sentiment classification performance (Danish, *n* = 2000)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.61	0.61	0.62	660	649
Neutral	0.69	0.73	0.65	1061	1192
Positive	0.35	0.28	0.48	279	159
Combined	0.62	0.63	0.62	2000	2000

Table 30. Sentiment classification performance (Dutch, *n* = 2000)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.52	0.48	0.56	520	445
Neutral	0.75	0.80	0.70	1233	1412
Positive	0.32	0.26	0.44	247	143
Combined	0.63	0.65	0.63	2000	2000

Table 31. Sentiment classification performance (Norwegian, with upsampling to largest category)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.55	0.52	0.60	1207	1048
Neutral	0.65	0.61	0.69	2108	1868
Positive	0.42	0.55	0.34	618	1017
Combined	0.58	0.57	0.61	3933	3933

Table 32. Sentiment classification performance (English, with up- sampling to largest category)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.66	0.61	0.72	1948	1657
Neutral	0.59	0.58	0.60	1812	1737
Positive	0.49	0.60	0.41	809	1175
Combined	0.60	0.60	0.62	4569	4569

Table 33. Sentiment classification performance (Danish, with up- sampling to largest category)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.62	0.61	0.64	1081	1027
Neutral	0.62	0.57	0.68	1659	1396
Positive	0.43	0.58	0.34	447	764
Combined	0.59	0.58	0.62	3187	3187

Table 34. Sentiment classification performance (Dutch, with up- sampling to largest category)

	F1	Recall	Precision	*n* (human coding)	*n* (predicted)
Negative	0.48	0.43	0.55	863	670
Neutral	0.69	0.65	0.73	2246	1999
Positive	0.34	0.51	0.25	429	869
Combined	0.60	0.58	0.63	3538	3538

Table 35. Norwegian seed dictionary

Positive seed words		**Negative seed words**	
dyktig_ADJ	glede_NOUN	misbruk_NOUN	fryktelig_ADJ
beundringsverdig_ADJ	vennligst_ADJ	redd_ADJ	såre_VERB
verdsette_VERB	elske_VERB	sinne_NOUN	uvel_ADJ
hensiktsmessig_ADJ	herlig_ADJ	sint_ADJ	mangelfull_ADJ
vakker_ADJ	kjærlig_ADJ	angst_NOUN	utilstrekkelig_ADJ
beste_ADJ	glimrende_ADJ	bekymre_ADJ	mindreverdig_ADJ
bedre_VERB	fordel_NOUN	dårlig_ADJ	urettferdighet_NOUN
klok_ADJ	snill_ADJ	brudd_NOUN	irrelevant_ADJ
støtte_NOUN	perfekt_ADJ	brutal_ADJ	miste_VERB
komfortabel_ADJ	perfeksjon_NOUN	byrde_NOUN	tap_NOUN
sikker_ADJ	behagelig_ADJ	uforsiktig_ADJ	elendig_ADJ
kreativ_ADJ	ros_NOUN	klage_VERB	tabbe_NOUN
fryd_NOUN	skikkelig_ADJ	klage_NOUN	forsømme_VERB
hyggelig_ADJ	velstand_NOUN	forvirring_NOUN	tull_NOUN
ønskelig_ADJ	beskytte_VERB	forakt_NOUN	smerte_NOUN
verdighet_NOUN	fornuftig_ADJ	korrupt_ADJ	smertefull_ADJ
virkningsfull_ADJ	pålitelig_ADJ	korrupsjon_NOUN	dårlig_PROPN
effektivitet_NOUN	respekt_NOUN	kritikk_NOUN	fordom_NOUN
effektiv_ADJ	respektere_VERB	skade_NOUN	problem_NOUN
oppmuntre_VERB	trygg_ADJ	fare_NOUN	beklagelse_NOUN
nyte_VERB	tilfredshet_NOUN	farlig_ADJ	innskrenke_VERB
utmerket_ADJ	tilfredsstille_ADJ	død_NOUN	restriksjon_NOUN
rettferdig_ADJ	tilfredsstille_VERB	ødelegge_VERB	latterlig_ADJ
åpen_ADJ	sikre_VERB	vanskelig_ADJ	risiko_NOUN
gunstig_ADJ	betydningsfull_ADJ	vanskelighet_NOUN	trist_ADJ
heldigvis_ADV	oppriktig_ADJ	ulempe_NOUN	skam_NOUN
frihet_NOUN	smart_ADJ	skuffelse_NOUN	syk_ADJ
vennlig_ADJ	løsning_NOUN	ulykke_NOUN	dum_ADJ
vennskap_NOUN	flott_ADJ	katastrofal_ADJ	lide_VERB
oppnå_VERB	styrke_NOUN	ubehag_NOUN	forferdelig_ADJ
sjenerøs_ADJ	forsterke_VERB	nød_NOUN	trussel_NOUN
ekte_ADJ	sterk_ADJ	fiende_NOUN	tragedie_NOUN

fornøyd_ADJ	lykkes_VERB	feil_NOUN	tragisk_ADJ
vidunderlig_ADJ	suksess_NOUN	ond_ADJ	stygg_ADJ
god_ADJ	vellykket_ADJ	overdrivelse_NOUN	uønsket_ADJ
takknemlig_ADJ	suveren_ADJ	overdreven_ADJ	urimelig_ADJ
lykke_NOUN	sympatisk_ADJ	mislykkes_VERB	uheldig_ADJ
glad_ADJ	sympati_NOUN	fiasko_NOUN	dessverre_ADV
sunn_ADJ	talent_NOUN	falsk_ADJ	mislykket_ADJ
hjelpe_VERB	sann_ADJ	mangel_NOUN	urettferdig_ADJ
hjelpsom_ADJ	genuint_ADJ	frykte_NOUN	irrasjonell_ADJ
ærlig_ADJ	sannhet_NOUN	engstelig_ADJ	uakseptabel_ADJ
ære_NOUN	nyttig_ADJ	svindel_NOUN	svak_ADJ
viktighet_NOUN	verdifull_A1D2 J	skremme_VERB	svakhet_NOUN
viktig_ADJ	sprek_ADJ	ubehagelig_ADJ	hensynsløs_ADJ
forbedre_VERB	velkommen_ADJ	skade_VERB	bekymre_VERB
bedring_NOUN	bra_ADJ	skadelig_ADJ	dårligere_ADJ
integritet_NOUN	lur_ADJ	hate_VERB	dårligst_ADJ
intelligent_ADJ	fantastisk_ADJ	hat_NOUN	ynkelig_ADJ
interessant_ADJ	verdig_ADJ	håpløs_ADJ	galt_ADJ

Table 36. English seed dictionary

Positive seed words		Negative seed words	
able_ADJ	joy_NOUN	abuse_NOUN	horrible_ADJ
admirable_ADJ	kindly_ADV	afraid_ADJ	hurt_VERB
appreciate_VERB	love_VERB	anger_NOUN	ill_ADJ
appropriate_ADJ	lovely_ADJ	angry_ADJ	imperfect_ADJ
beautiful_ADJ	loving_ADJ	anxiety_NOUN	inadequate_ADJ
best_ADJ	magnificent_ADJ	anxious_ADJ	inferior_ADJ
better_ADJ	merit_NOUN	bad_ADJ	injustice_NOUN
clever_NOUN	nice_ADJ	breach_NOUN	irrelevant_ADJ
comfort_NOUN	perfect_ADJ	brutal_ADJ	lose_VERB
comfortable_ADJ	perfection_NOUN	burden_NOUN	loss_NOUN
confident_ADJ	pleasant_ADJ	careless_ADJ	miserable_ADJ
creative_ADJ	praise_NOUN	complain_VERB	mistake_NOUN
delight_NOUN	properly_ADV	complaint_NOUN	neglect_VERB
delightful_ADJ	prosperity_NOUN	confusion_NOUN	nonsense_NOUN
desirable_ADJ	protect_VERB	contempt_NOUN	pain_NOUN
dignity_NOUN	reasonable_ADJ	corrupt_ADJ	painful_ADJ
effective_ADJ	reliable_ADJ	corruption_NOUN	poorly_ADV
efficiency_NOUN	respect_NOUN	criticism_NOUN	prejudice_NOUN
efficient_ADJ	respected_ADJ	damage_NOUN	problem_NOUN
encourage_VERB	safe_ADJ	danger_NOUN	regret_NOUN
enjoy_VERB	satisfaction_NOUN	dangerous_ADJ	restrict_VERB
excellent_ADJ	satisfactory_ADJ	death_NOUN	restriction_NOUN
fair_ADJ	satisfying_ADJ	destroy_VERB	ridiculous_ADJ

Positive seed words		Negative seed words	
fairly_ADV	secure_VERB	difficult_ADJ	risk_NOUN
fortunate_ADJ	significant_ADJ	difficulty_NOUN	sad_ADJ
fortunately_ADV	sincere_NOUN	disadvantage_NOUN	shame_NOUN
freedom_NOUN	smart_ADJ	disappointment_NOUN	sick_ADJ
friendly_ADJ	solution_NOUN	disaster_NOUN	stupid_ADJ
friendship_NOUN	splendid_ADJ	disastrous_ADJ	suffer_VERB
gain_VERB	strength_NOUN	discomfort_NOUN	terrible_ADJ
generous_ADJ	strengthen_VERB	distress_NOUN	threat_NOUN
genuine_ADJ	strong_ADJ	enemy_NOUN	tragedy_NOUN
glad_ADJ	succeed_VERB	error_NOUN	tragic_ADJ
glorious_ADJ	success_NOUN	evil_ADJ	ugly_ADJ
good_ADJ	successful_ADJ	excess_NOUN	undesirable_ADJ
grateful_ADJ	superior_ADJ	excessive_ADJ	unfair_ADJ
happiness_NOUN	sympathetic_ADJ	fail_VERB	unfortunate_ADJ
happy_ADJ	sympathy_NOUN	failure_NOUN	unfortunately_ADV
healthy_ADJ	talent_NOUN	false_ADJ	unhappy_ADJ
help_VERB	true_ADJ	fault_NOUN	unjust_ADJ
helpful_ADJ	truly_ADV	fear_NOUN	unreasonable_ADJ
honest_ADJ	truth_NOUN	fearful_ADJ	unsatisfactory_ADJ
honour_NOUN	useful_ADJ	fraud_NOUN	weak_ADJ
importance_NOUN	valuable_AD1J3	frightened_ADJ	weakness_NOUN
important_ADJ	vigorous_ADJ	grim_ADJ	wicked_ADJ
improve_VERB	welcome_ADJ	harm_VERB	worry_VERB
improvement_NOUN	well_ADV	harmful_ADJ	worse_ADJ
integrity_NOUN	wise_ADJ	hate_VERB	worst_ADJ
intelligent_ADJ	wonderful_ADJ	hatred_NOUN	wretched_ADJ
interesting_ADJ	worthy_ADJ	hopeless_ADJ	wrong_ADV

Table 37. Danish seed dictionary

Positive seed words		Negative seed words	
dygtig_ADJ	glæde_NOUN	misbrug_NOUN	frygtelig_ADJ
beundringsværdig_ADJ	venligt_ADV	bange_ADJ	såre_VERB
værdsætte_VERB	elske_VERB	vrede_ADJ	usund_ADJ
passende_ADJ	dejlig_ADJ	vred_ADJ	mangelfuld_ADJ
smuk_ADJ	kærlig_ADJ	bekymring_NOUN	utilstrækkelig_ADJ
bedst_ADJ	storslået_ADJ	ængstelig_ADJ	lavere_ADJ
bedre_ADJ	fortjeneste_NOUN	dårlig_ADJ	uretfærdighed_NOUN
klog_ADJ	rar_ADJ	brud_NOUN	uvedkommende_VERB
trøst_NOUN	perfekt_ADJ	brutal_ADJ	tabe_VERB
komfortabel_ADJ	perfektion_NOUN	belastning_NOUN	tab_NOUN
fortrøstningsfuld_ADJ	behagelig_ADJ	uforsigtig_ADJ	elendig_ADJ
kreativ_ADJ	ros_NOUN	klage_VERB	fejl_NOUN

Positive seed words		**Negative seed words**	
fornøjelse_NOUN	ordentlig_ADV	klage_NOUN	forsømme_VERB
fornøjelig_ADJ	fremgang_NOUN	forvirring_NOUN	vrøvl_NOUN
attraktiv_ADJ	beskytte_VERB	foragt_NOUN	smerte_NOUN
værdighed_NOUN	fornuftig_ADJ	korrupt_ADJ	smertelig_ADJ
effektfuld_ADJ	pålidelig_ADJ	korruption_NOUN	elendigt_ADV
effektivitet_NOUN	respekt_NOUN	kritik_NOUN	fordom_NOUN
effektiv_ADJ	anerkendt_ADJ	ødelæggelse_NOUN	problem_NOUN
opmuntre_VERB	tryg_ADJ	fare_NOUN	beklagelse_NOUN
nyde_VERB	tilfredshed_NOUN	farlig_ADJ	begrænse_VERB
fremragende_ADJ	overbevisende_VERB	død_NOUN	restriktion_NOUN
rimelig_ADJ	tilfredsstillende_ADJ	ødelægge_VERB	latterlig_ADJ
ganske_ADV	sikre_VERB	svær_ADJ	risiko_NOUN
heldig_ADJ	betydningsfuld_ADJ	besvær_NOUN	trist_ADJ
heldigvis_ADV	oprigtig_ADJ	ulempe_NOUN	skam_NOUN
frihed_NOUN	smart_ADJ	skuffelse_NOUN	syg_ADJ
venlig_ADJ	løsning_NOUN	katastrofe_NOUN	dum_ADJ
venskab_NOUN	flot_ADJ	katastrofal_ADJ	lide_VERB
opnå_VERB	styrke_NOUN	ubehag_NOUN	forfærdelig_ADJ
gavmild_ADJ	forstærke_VERB	sorg_NOUN	trussel_NOUN
ægte_ADJ	stærk_ADJ	fjende_NOUN	tragedie_NOUN
glad_ADJ	lykkes_VERB	fejltagelse_NOUN	tragisk_ADJ
pragtfuld_ADJ	succes_NOUN	ond_ADJ	grim_ADJ
god_ADJ	vellykket_ADJ	overskridelse_NOUN	uønsket_ADJ
taknemmelig_ADJ	overlegenhed_NOUN	overdreven_ADJ	unfair_ADJ
lykke_NOUN	sympatisk_ADJ	mislykkes_VERB	ulykkelig_ADJ
lykkelig_ADJ	sympati_NOUN	nederlag_NOUN	uheldigvis_ADV
sund_ADJ	talent_NOUN	falsk_ADJ	utilfreds_ADJ
hjælpe_VERB	sand_ADJ	mangel_NOUN	uretfærdig_ADJ
hjælpsom_ADJ	virkelig_ADV	frygt_NOUN	urimelig_ADJ
ærlig_ADJ	sandhed_NOUN	frygtsom_ADJ	utilfredsstillende_ADJ
ære_NOUN	nyttig_ADJ	bedrageri_NOUN	svag_ADJ
betydning_NOUN	værdifuld_14ADJ	skræmt_ADJ	svaghed_NOUN
vigtig_ADJ	energisk_ADJ	barsk_ADJ	rædselsfuld_ADJ
forbedre_VERB	velkommen_ADJ	skade_VERB	bekymre_VERB
forbedring_NOUN	godt_ADV	skadelig_ADJ	værre_ADJ
integritet_NOUN	forstandig_ADJ	hade_VERB	værst_ADV
intelligent_ADJ	vidunderlig_ADJ	had_NOUN	stakkels_ADJ
interessant_ADJ	værdig_ADJ	håbløs_ADJ	forkert_ADJ

Table 38. Dutch seed dictionary

Positive seed words		Negative seed words	
capabel_ADJ	vreugde_NOUN	misbruik_NOUN	verschrikkelijk_ADJ
bewonderenswaardig_ADJ	welwillend_ADJ	bevreesd_ADJ	kwetsen_VERB
waarderen_VERB	liefhebben_VERB	woede_NOUN	kwalijk_ADJ
passend_ADJ	lief_ADJ	woedend_ADJ	imperfect_ADJ
mooi_ADJ	liefdevol_ADJ	ongerustheid_NOUN	ontoereikend_ADJ
best_ADJ	prachtig_ADJ	bezorgd_ADJ	inferieur_ADJ
beter_ADJ	verdienste_NOUN	slecht_ADJ	onrecht_NOUN
slim_NOUN	prettig_ADJ	breuk_NOUN	onbelangrijk_ADJ
comfort_NOUN	perfect_ADJ	wreed_ADJ	verliezen_VERB
comfortabel_ADJ	perfectie_NOUN	last_NOUN	verlies_NOUN
overtuigd_ADJ	aangenaam_ADJ	onzorgvuldig_ADJ	miserabel_ADJ
creatief_ADJ	lof_NOUN	klagen_VERB	vergissing_NOUN
genot_NOUN	juist_ADV	klacht_NOUN	verwaarlozen_VERB
verrukkelijk_ADJ	voorspoed_NOUN	verwarring_NOUN	nonsens_NOUN
wenselijk_ADJ	beschermen_VERB	minachting_NOUN	pijn_NOUN
waardigheid_NOUN	redelijk_ADJ	corrupt_ADJ	pijnlijk_ADJ
effectief_ADJ	betrouwbaar_ADJ	corruptie_NOUN	slecht_ADV
efficiëntie_NOUN	respect_NOUN	kritiek_NOUN	vooroordeel_NOUN
efficiënt_ADJ	geliefd_ADJ	schade_NOUN	probleem_NOUN
aanmoedigen_VERB	veilig_ADJ	gevaar_NOUN	spijt_NOUN
genieten_VERB	voldoening_NOUN	gevaarlijk_ADJ	beperken_VERB
uitstekend_ADJ	voldoende_ADJ	dood_NOUN	beperking_NOUN
eerlijk_ADJ	bevredigend_ADJ	vernietigen_VERB	belachelijk_ADJ
tamelijk_ADV	beveiligen_VERB	moeilijk_ADJ	risico_NOUN
fortuinlijk_ADJ	significant_ADJ	moeilijkheid_NOUN	verdrietig_ADJ
gelukkig_ADJ	oprecht_ADJ	nadeel_NOUN	schaamte_NOUN
vrijheid_NOUN	slim_ADJ	teleurstelling_NOUN	ziek_ADJ
vriendelijk_ADJ	oplossing_NOUN	ramp_NOUN	dom_ADJ
vriendschap_NOUN	schitterend_ADJ	rampzalig_ADJ	lijden_VERB
winnen_VERB	kracht_NOUN	ongemak_NOUN	vreselijk_ADJ
vrijgevig_ADJ	versterken_VERB	nood_NOUN	bedreiging_NOUN
authentiek_ADJ	sterk_ADJ	vijand_NOUN	tragedie_NOUN
verheugd_ADJ	slagen_VERB	fout_NOUN	tragisch_ADJ
glorieus_ADJ	succes_NOUN	onheil_ADJ	lelijk_ADJ
goed_ADJ	succesvol_ADJ	overdaad_NOUN	onwenselijk_ADJ
dankbaar_ADJ	superieur_ADJ	overdadig_ADJ	oneerlijk_ADJ
geluk_NOUN	sympathiek_ADJ	falen_VERB	onfortuinlijk_ADJ
blij_ADJ	sympathie_NOUN	mislukking_NOUN	helaas_ADV
gezond_ADJ	talent_NOUN	onjuist_ADJ	ongelukkig_ADJ
helpen_VERB	waar_ADJ	schuld_NOUN	onrechtvaardig_ADJ
behulpzaam_ADJ	werkelijk_ADJ	angst_NOUN	onredelijk_ADJ
oprecht_ADJ	waarheid_NOUN	angstig_ADJ	onbevredigend_ADJ

Positive seed words		Negative seed words	
eer_NOUN	bruikbaar_ADJ	fraude_NOUN	zwak_ADJ
belang_NOUN	waardev1o5l_ADJ	bang_ADJ	zwakte_NOUN
belangrijk_ADJ	krachtig_ADJ	grimmig_ADJ	goddeloos_ADJ
verbeteren_VERB	welkom_ADJ	schaden_VERB	piekeren_VERB
verbetering_NOUN	goed_ADV	schadelijk_ADJ	slechter_ADJ
integriteit_NOUN	wijs_ADJ	haten_VERB	slechtst_ADJ
intelligent_ADJ	geweldig_ADJ	haat_NOUN	ellendig_ADJ
interessant_ADJ	waardig_ADJ	hopeloos_ADJ	fout_ADJ

Notes

[1] An annotated reproducible example of the adapted method is provided at https://github.com/vriezer/sentiment.

[2] Even when avoiding ambiguous seed words, the final dictionary might still be biased due to the presence of bias in the (source data of the) WE model used to construct the dictionary.

[3] A full description of the irrelevant article coding procedure and its results can be found in the appendix at https://osf.io/tb3kr/

[4] The focus on political actors is due to reasons of data availability, and coders are instructed to ignore their presence when coding sentiment.

[5] Dutch and Flemish are treated as a single language (Dutch), but as geographically distinct media markets/domains.

[6] Replication materials to reproduce the results presented in this section are available as supplementary material at https://osf.io/tb3kr/.

[7] Full performance results for all languages for both the sample size and upsampling experiments can be found in the appendix.

[8] Spearman rank order correlation with human coding. WE: 0.464 (Danish), 0.347 (Dutch), 0.460 (English), 0.399 (Norwegian). Polyglot: 0.255 (Danish), 0.157 (Dutch), 0.385 (English), 0.224 (Norwegian).

[9] Confusion matrices can be found in the appendix.

References

Alba, A., Gruhl, D., Ristoski, P., & Welch, S. (2018). Interactive dictionary expansion using neural language models. *HumL@ ISWC*, 7–15.

Aldayel, A., & Magdy, W. (2021). Stance detection on social media: State of the art and trends. *Information Processing & Management*, *58* (4), 102597. https://doi.org/10.1016/j.ipm.2021.102597

Alhothali, A., & Hoey, J. (2017). Semi-Supervised Affective Meaning Lexicon Expansion Using Semantic and Distributed Word Representations. *arXiv:1703.09825* [*Cs*]. https://arxiv.org/abs/1703.09825

Almeida, F., & Xexéo, G. (2019). Word Embeddings: A Survey. *arXiv:1901.09069* [*Cs, Stat*]. https://arxiv.org/abs/1901.09069

Al-Rfou, R., Perozzi, B., & Skiena, S. (2013). *Polyglot: Distributed Word Representations for Multilingual NLP*. 10.

Amsler, M. (2020). *Using Lexical-Semantic Concepts for Fine-Grained Classification in the Embedding Space* [PhD thesis]. University of Zurich.

Bleich, E., & van der Veen, A. M. (2018). Media portrayals of Muslims: A comparative sentiment analysis of American newspapers, 1996–2015. *Politics, Groups, and Identities*, 1–20. https://doi.org/10.1080/21565503.2018.1531770

Boukes, M., van de Velde, B., Araujo, T., & Vliegenthart, R. (2020). What's the Tone? Easy Doesn't Do It: Analyzing Performance and Agreement Between Off-the-Shelf Sentiment Analysis Tools. *Communication Methods and Measures*, *14* (2), 83–104. https://doi.org/10.1080/19312458.2019.1671966

Chen, Y., & Skiena, S. (2014). Building Sentiment Lexicons for All Major Languages. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (*Volume 2: Short Papers*), 383–389. https://doi.org/10.3115/v1/P14-2063

de Vreese, C., Esser, F., & Hopmann, D. N. (2016). *Comparing Political Journalism*. Routledge. https://doi.org/10.4324/9781315622286

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis*.

Glogger, I. (2019). Soft Spot for Soft News? Influences of Journalistic Role Conceptions on Hard and Soft News Coverage. *Journalism Studies*, *20* (16), 2293–2311. https://doi.org/10.1080/1461670X.2019.1588149

Hallin, D. C., & Mancini, P. (2004). Comparing Media Systems: Three Models of Media and Politics. In *Cambridge Core*. /core/books/comparing-mediasystems/B7A12371782B7A1D62BA1A72C1395E43; Cambridge University Press. https://doi.org/10.1017/CBO9780511790867

Hlavac, M. (2018). *Stargazer: Well-Formatted Regression and Summary Statistics Tables.* https://CRAN.R-project.org/package=stargazer.

Khoo, C. S., & Johnkhan, S. B. (2018). Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science*, *44* (4), 491–511. https://doi.org/10.1177/0165551517703514

Lengauer, G., Esser, F., & Berganza, R. (2012). Negativity in political news: A review of concepts, operationalizations and key findings. *Journalism*, *13* (2), 179–202. https://doi.org/10.1177/1464884911427800

Makki, R., Brooks, S., & Milios, E. E. (2014). Context-specific sentiment lexicon expansion via minimal user interaction. *2014 International Conference on Information Visualization Theory and Applications* (*IVAPP*), 178–186.

Mikolov, T., Yih, W., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 746–751.

Mohammad, S. M. (2016). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In H. Meiselman (Ed.), *Emotion measurement*. Elsevier.

Muddiman, A., McGregor, S. C., & Stroud, N. J. (2019). (Re)Claiming Our Expertise: Parsing Large Text Corpora With Manually Validated and Organic Dictionaries. *Political Communication*, *36* (2), 214–226. https://doi.org/10.1080/10584609.2018.1517843

Nivre, J., Abrams, M., Agić, Ž., Ahrenberg, L., Antonsen, L., Aplonova, K., Aranzabe, M. J., Arutie, G., Asahara, M., Ateyah, L., Attia, M., Atutxa, A., Augustinus, L., Badmaeva, E., Ballesteros, M., Banerjee, E., Bank, S., Barbu Mititelu, V., Basmov, V., & Zhu, H. (2018). *Universal dependencies 2.3*.

Otto, L., Glogger, I., & Boukes, M. (2017). The Softening of Journalistic Political Communication: A Comprehensive Framework Model of Sensationalism, Soft News, Infotainment, and Tabloidization. *Communication Theory*, *27* (2), 136–155. https://doi.org/10.1111/comt.12102

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), 1532–1543. https://doi.org/10.3115/v1/D14-1162

Proksch, S.-O., Lowe, W., Wäckerle, J., & Soroka, S. (2019). Multilingual Sentiment Analysis: A New Approach to Measuring Conflict in Legislative Speeches. *Legislative Studies Quarterly*, *44* (1), 97–131. https://doi.org/10.1111/lsq.12218

Reinemann, C., Stanyer, J., Scherr, S., & Legnante, G. (2012). Hard and soft news: A review of concepts, operationalizations and key findings. *Journalism*, *13* (2), 221–239. https://doi.org/10.1177/1464884911427803

Rheault, L., Beelen, K., Cochrane, C., & Hirst, G. (2016). Measuring Emotion in Parliamentary Debates with Automated Textual Analysis. *PLOS ONE*, *11* (12), e0168843. https://doi.org/10.1371/journal.pone.0168843

Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š., & Sedlmair, M. (2018). More than Bags of Words: Sentiment Analysis with Word Embeddings. *Communication Methods and Measures*, *12* (2-3), 140–157. https://doi.org/10.1080/19312458.2018.1455817

Shi, T., Malioutov, I., & İrsoy, O. (2020). Semantic Role Labeling as Syntactic Dependency Parsing. *arXiv:2010.11170 [Cs]*. https://arxiv.org/abs/2010.11170

Soroka, S., Young, L., & Balmas, M. (2015). Bad News or Mad News? Sentiment Scoring of Negativity, Fear, and Anger in News Content. *The ANNALS of the American Academy of Political and Social Science*, *659* (1), 108–121. https://doi.org/10.1177/0002716215569217

Straka, M., & Straková, J. (2017). Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 88–99.

van Atteveldt, W., Sheafer, T., Shenhav, S. R., & Fogel-Dror, Y. (2017). Clause Analysis: Using Syntactic Information to Automatically Extract Source, Subject, and Predicate from Texts with an Application to the 2008–2009 Gaza War. *Political Analysis*, *25* (02), 207–222. https://doi.org/10.1017/pan.2016.12

van Atteveldt, W., van der Velden, M. A. C. G., & Boukes, M. (2021). The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms. *Communication Methods and Measures*, *15* (2), 121–140. https://doi.org/10.1080/19312458.2020.1869198

Young, L., & Soroka, S. (2012). Affective News: The Automated Coding of Sentiment in Political Texts. *Political Communication*, *29* (2), 205–231. https://doi.org/10.1080/10584609.2012.671234

Telling a Different Story

A Longitudinal Investigation of News Diversity in Four Countries

Erik de Vries
Department of Media and Social Sciences,

University of Stavanger
erik.devries@uis.no

Rens Vliegenthart
Strategic Communication Group,

Wageningen University & Research

Stefaan Walgrave
Department of Political Science,

University of Antwerp

News diversity is an important concern of journalism scholars, as its presence or absence can have a profound effect on democratic debate and the information available to citizens. Many have speculated that news diversity decreases over time, due to changing economic circumstances. This expectation especially applies to newspapers. Using nearly two decades of newspaper data from four European countries (Denmark, The Netherlands, Norway, UK), we do not find this expected decrease in news diversity. When conducting pairwise, automated comparisons between articles published on the same day in the same country, we rather find a modest over time increase in diversity between newspapers. This result suggests that newspapers differentiate rather than converge in the content they offer, shedding a more positive light on the evolution of the press in our current high-choice media environments.

Keywords: news diversity, political journalism, framing, comparative research, computational text analysis, content analysis

Introduction

During the last decades, worries have grown about a potentially increasing convergence of news coverage in traditional media. If news media indeed make increasingly similar news selection and framing choices, this could be considered worrying from a democratic perspective. External news diversity, with news outlets competing by offering different stories, is considered by many scholars to be an important feature of a healthy information environment. When different outlets highlight different stories, different elements of the same story, or evaluate the story differently, this generates a rich and more pluralist context to spark democratic debate. It also allows citizens, if they want, to be confronted with different opinions and form themselves a nuanced idea about the facts and how to evaluate them.

The reasons for widespread pessimism regarding decreasing external news diversity are strongly related to structural changes in the news business. Two complementary mechanisms are supposed to be at work. First, increasing media competition, audience fragmentation and hybridization of news lead to more pressure on journalists to produce more news stories. Without proper time to select deviant news stories and to develop their own approach, journalists produce news stories that are interchangeable and that are heavily affected by the information subsidies provided by the story’s stakeholders. Second, economic pressure on the news sector leads to a concentration of different news outlets in the hands of fewer owners who push for increased collaboration and integration of the newsrooms in their portfolio. This is likely to almost mechanically generate an increasing overlap in news stories. The two causes of the alleged decrease in diversity are well-documented; journalist surveys almost invariably point to increased productivity (Hanusch, 2015 in Australia; Jyrkiäinen & Heinonen, 2012 in Finland; Raeymaeckers et al., 2012 in Belgium) and news outlets are increasingly concentrated in conglomerates. Even so, actual empirical proof of their alleged effect on news diversity is rare, and the few studies that exist present mixed evidence.

In this paper we contribute to the ongoing debate about news diversity. Our study is empirical, not normative. We do not take a normative position and do not claim that decreasing news diversity is invariably a bad thing; in fact, some studies show that news, if diverse, is less used by the consumers (Van Aelst et al., 2017) and concentrated attention could under certain circumstances spark societal debate and put pressure on decision makers to be responsive (Walgrave et al., 2017). What we do here is empirically examining news diversity in a range of countries for a long time period. The content of individual newspaper articles published on the same day, in four countries (Denmark, The Netherlands, Norway and the UK) is analyzed and compared. This is done for three newspapers per country, over a period of twenty years (2000-2019). While newspapers all over the globe have suffered serious declines in readership figures, they are still among the most frequently used media sources in the countries under study (Newman et al., 2021) and have for example shown to exert considerable political agenda setting power (Langer and Gruber, 2021). Additionally, in times of dynamic and fastly-changing media landscapes, they have been among the few outlets that have sustained a prominent position and consequently offer ample opportunities for systematic over-time comparisons. In total, 6 million newspaper articles are examined. We rely on an automated procedure to identify for each day and each newspaper pair in a country which articles are most alike. These most similar article pairs (between newspapers) are also most likely to deal with the same topic. If the similarity of article pairs about the same topic goes up over time, we argue that there is a tendency towards less topic diversity. In addition, we leverage a computer-generated sentiment dictionary to examine the degree to which article pairs about the same topic also present these topics in a similar sentiment context. This allows us to determine if there is a trend towards less sentiment diversity as well. Our results do not support the expectation of less diversity, however. In none of the four countries we find decreasing topic diversity. The same applies to sentiment diversity: it does not decrease over time. If anything, in some countries, we see very modest signs of the exact opposite pattern, namely that newspaper content becomes more diverse over time. We discuss this contra-intuitive trend and link it to the further growth of interpretative journalism, whereby newspapers try to differentiate themselves in an increasingly competitive environment by offering distinct news facts and interpretations.

Why news diversity would be decreasing

The importance of news diversity is a topic of societal and scientific debate. Many assume that, in one way or another, diversity is relevant for the functioning of democracy (Beckers et al., 2019; Napoli, 1999; Sjøvaag, 2016; Vogler et al., 2020).

Western democracies rely on media and their supply of information to function properly. Voters need to be able to obtain diverse information in order to be able to cast an informed vote. Despite the fast-changing information environment, legacy mass media remain the most prominent source of information for the majority of citizens in most countries. At the same time, politicians too need access to media and the information media provide to communicate with voters about relevant issues (Van Aelst & Walgrave, 2016). There is debate about how large news diversity should be exactly, and some argue that too much diversity is not good either, as it leads to audience fragmentation and the public sphere falling apart (Roessler, 2007). Van Cuilenburg (1999, p. 199) states that diversity in the media cannot be evaluated in the abstract, as it "… should always be compared with relevant variations in society and social reality." In other words, media diversity should ideally be a reflection of the actual diversity of ideas and opinions in society (see also Joris et al., 2020).

Since measuring the real diversity of opinions and ideas in a society is hardly possible, it is problematic to observe the relative diversity in media markets—that is the diversity relative to real world differences. Most research therefore looks at absolute diversity, focusing in particular on over-time and cross-context variation. Some scholars have for example examined the absolute diversity of topics in news stories from different media. Do they cover the same events during a particular news cycle (see Boczkowski & Santos, 2007; Joris et al., 2020)? Topic diversity is only part of the story, though. News media that are supposed to represent the diversity of ideas and opinions in a society should also to some extent employ different interpretations when telling their stories. In fact, while investigating news topics provides an idea of what news media report about, it does not tell anything about the “interpretation, evaluation and/or solution” (Entman, 2003, p. 417) related to an event. In line with those interpretations, in this paper we consider external news diversity as the extent to which, in a given period of time, different news outlets (1) report on different topics (2), and when they report on the same topics, to what extent they use a similar tone or sentiment. Note that this definition explicitly excludes internal news diversity or the diversity of content features within a news outlet. In the remainder of this text, the term “diversity” is exclusively used in the context of external news diversity. Additionally, we constrain ourselves to topics and sentiment, and do not single out events or more elaborate frames.

Researchers have speculated that, in many countries, diversity of the news has diminished (Lee, 2007; Schudson, 2011). The dominant account holds that due to underlying structural economic evolutions, news corporations have been forced to change their strategy, with decreasing news diversity as a consequence. This so-called ‘newspaper crisis’ is characterized by high levels of competition in shrinking markets (Curran, 2010). Under these circumstances, newspapers all across Europe have been struggling to keep their businesses profitable (Brüggemann et al., 2012; Lewis et al., 2008; Vogler et al., 2020). For a large part this is due to people switching from newspapers to other news sources, especially on the Internet or social media. Newspapers have suffered from declining readership numbers, and lower advertising and subscription revenues. There exist different and contradictory accounts of how such conditions of high competition affects newspapers’ strategy. Hotelling’s Law (1929) posits that under conditions of high competition competitors generally tend to compete on price, rather than on product differentiation and quality (see also Van Cuilenburg, 1999). High competition in the news market would thus lead to cost cutbacks and, hence, to less news diversity. At the same time, some studies found the exact opposite, being that newspapers invest more in product quality, and thus product diversity, when the pressure of competition increases (Lacy & Simon, 1993).

Still, the less diversity argument is more frequently present in the literature because there is more proof of the fact that high competition has led to cost reduction and staff cuts. This is the first likely cause for a decrease in diversity: less and less journalists must produce the same amount of content. Taking some time to develop a story, and thereby make it different from that of a competing outlet, is therefore often not possible. Indeed, more resources for reporting and more specialized journalism lead to more diverse news both in terms of the events covered and the evaluation. Or, inversely, a higher workload leads to journalists increasingly relying on content—often called 'information subsidies' (Gandy, 1980) —produced by external sources, such as PR and news agencies (Boumans et al., 2018; Vogler et al., 2020). Under time pressure, journalists from different media outlets cannot but rely on the same external sources, leading to a reduction in news diversity.

Yet, even if staff were not cut, the news business has changed rapidly and news outlets such as newspapers are not only present on the print news market but have become broad news providers with elaborate news websites and a strong presence on social media. This entails a 24/7 online news presence, meaning that journalists simply have to produce more news on each day (Paulussen, 2012). The effects thereof are comparable to the effect of staff cuts and leads to a second reason for declining news diversity: less time per news item and, hence, less chance to develop a different topic and angle choice.

Besides staff cuts and increased workload, a third reason for the alleged decline in news diversity, is that publishing houses have merged as yet another consequence of the newspaper crisis (Curran, 2010). The result is a concentration of ownership, with only a few publishing houses owning the majority of national newspapers in most European countries (Picard, 2014). This mechanism has, according to some, also led to an overall decrease in news diversity. If newsrooms lose autonomy and are forced to (partly) pool resources with other newsrooms, the consequence can only be that the news choices and output of the collaborating outlets becomes more similar (Beckers et al., 2019; Dailey et al., 2005; Hendrickx & Ranaivoson, 2019). So, trends in media concentration and news outlet ownership have arguably diminished news diversity (Baker, 2007).

Topic and sentiment diversity

Notwithstanding the widely shared pessimism regarding decreasing news diversity, the number of empirical studies that demonstrates decreasing diversity over time remains very small and their findings are mixed. Joris et al. (2020) provide a systematic review of news diversity studies. These studies differ in the way the conceptualize and operationalize news diversity, as well as in scope of the investigation, and maybe not surprisingly, in the results. Regarding conceptualization, the literature overview by Joris and colleagues (2020) finds thirteen studies that have looked at topic diversity (almost none with over time comparisons) but hardly any study in their overview considers, for instance, viewpoint diversity (e.g. Day & Golan, 2005 who looked at viewpoint diversity in op-ed contributions within the same newspaper; see also Rodgers et al., 2000) or actor diversity (e.g. Masini & Van Aelst, 2017). As mentioned earlier, the variation between outlets in the interpretation and evaluation of stories may, from a democratic diversity perspective, be even more important than whether they actually cover the same events. Indeed, work in political communication has shown that how journalists discuss a topic or event matters for how news consumers digest it. Sentiment, either attributed directly towards specific objects, or at the more general level of a news item, is a frequently considered content feature (see Boukes et al., 2020). In general, it has been shown that negative coverage has a stronger effect on the audience than positive coverage (Soroka & McAdams, 2015; Vliegenthart et al., 2021) and consequently are more widely used to attract the largest audience possible (Damstra & De Swert, 2020). For example, if the economy is generally discussed in negative terms, this yields lower levels of consumer confidence, and those effects are larger than for positive coverage. Similar to topic diversity, general sentiment in issue coverage can differ in its level of diversity. If then competition for readers increases, it might well be that variety in sentiment by which topics are discussed decreases as well, following a similar logic as for topic diversity.

Only few longitudinal studies exist. They indeed show that newspaper coverage becomes less diverse (Boczkowski & Santos, 2007 in Argentina; Vogler et al., 2020 in Switzerland) and more reliant on the same external sources (Vogler et al., 2020 in Switzerland). Other studies do not find a decrease in diversity (Beckers et al., 2019 in Belgium). Yet, these few longitudinal studies all look at one specific country, they examine different time periods, and their measurements of diversity vary largely. Hence, although the underlying economic trends of competition, increasing work pressure, staff cuts and mergers are well-established, there is no firm proof for the often-assumed consequential decreasing news diversity. Further, all extant longitudinal studies only look at topic diversity and none take other aspects of diversity into account. Our study tries to improve on previous work by (1) testing the generalizability of trends by looking at newspapers in four different countries (Denmark, Norway, UK and the Netherlands) belonging to different media system types (Hallin & Mancini, 2004), (2) by covering a long time period (2000-2019), (3) by including all types of hard news content, (4) by drawing on a systematic and reliable automated approach to assess diversity (Amsalem et al., 2020; Vogler et al., 2020), and (5) by looking both at both topic diversity and sentiment diversity.

Hypotheses and research question

Although there is no compelling proof of decreasing topic diversity over time and although evidence about sentiment diversity is even entirely absent, we postulate two simple longitudinal hypotheses that follow from the literature above and that will guide our analyses in the next sections:

H_1: *Topic diversity decreased during the period 2000-2019*

H_2: *Sentiment diversity decreased during the period 2000-2019*

In addition, we specifically investigate the possible effects of differences between the four countries in terms of the political and media environment. Most notably, the country sample consists of two different media systems, Liberal in the UK, and Democratic-Corporatist in Denmark, Norway and the Netherlands (Hallin & Mancini, 2004). Additionally, the UK has a majoritarian electoral system, while the other countries have a system of proportional representation (Farrell, 2011). Also journalistic cultures and role perceptions deviate across the countries, though differences are not strongly pronounced (Hanitzsch et al., 2019). And while the “newspaper crisis” has been found in both Liberal (UK: Lewis et al. (2008); US: Curran (2010)) and Democratic-Corporatist (Germany: Brüggemann et al. (2012); Switzerland: Vogler et al. (2020)) media systems, there might be substantial differences between the UK and the other countries. With its liberal media system and majoritarian electoral system, the highest level of market pressures might be anticipated in the UK. We explore whether this also yields more substantial shifts in diversity than in the Democratic-Corporatist countries. Additionally, the British tabloid newspapers dominate the UK market, with *the Sun* as the most exemplary and outspoken example of this type of newspaper. We explore in detail the behavior of this newspaper and how it relates to the other UK newspapers. Therefore, we formulate an additional research question:

RQ1: What differences are there in the development of diversity between newspapers in the Liberal UK media system and the Democratic-Corporatist systems in the Netherlands, Denmark and Norway?

Data

Sample selection

To test our hypotheses, we use 6 million newspaper articles, representing 12 newspapers from 4 countries (3 newspapers per country). From each country a left-leaning broadsheet, right-leaning broadsheet and tabloid/popular newspaper is selected (De Vreese et al., 2016). The selection of newspapers used for each of the four countries under investigation is shown in Table 1. We use all articles from these newspapers, without further sampling.

(Table 1 around here)

Pre-processing

All newspaper articles are parsed using Natural Language Processing through the R package UDPipe (Straka & Straková, 2017), in combination with version 2.3 of the

Dutch Alpino, Danish DDT, English EWT and Norwegian Bokmål Universal Dependencies Models (Nivre et al., 2018). The relevant output of this procedure is a corpus where the original (inflected) words have been reduced to their dictionary lemmas. The advantage of using lemmas is that it reduces the number of unique words in the corpus, without losing the substantive meaning of words. In addition, UDPipe produces Universal Part-Of-Speech (UPOS) tags, which identify the grammatical function of a word. The combination of lemma_UPOS pairs is used to train a simple Naive Bayes classification model to remove any articles from the corpora that are not relevant for the current research. These are articles about sports and cultural events, weather forecasts, etc.[1] Such articles do not provide a relevant indication of content diversity from a political/democratic perspective. A full description of the irrelevant article coding procedure and its results can be found in De Vries (2022).

Methods

We use individual content units (articles) as the base unit of analysis and analyze the diversity between those individual units. By aggregating the scores, a measure is constructed, indicating the diversity between newspapers within a country on a given day. We choose to aggregate the article-level comparisons to provide a general

[1] See the appendix for a full overview.

overview of the development of content diversity in different countries, over a period of (almost) two decades. [2]

Content diversity

There are many possible dimensions of content diversity to analyze (see Joris et al. (2020) for an overview). However, because of the large amount of data, and the automated analysis methods used to analyze that data, we do not investigate highly specific dimensions of diversity, such as issue-specific framing. Rather, we focus on word usage, both overall (to what extent do two articles use the same words), and on the sentiment of the words used. The former should capture the general topic of the article, while the latter should capture the sentiment context in which the topic is presented.

To construct a content diversity measure, we compare all articles per day and country to each other, using the lemmas those articles consist of. The ways in which such a measure can be constructed vary; Boumans (2016) uses cosine similarity in combination with tf-idf feature weighting to determine the extent to which newspaper articles are based on external materials, while Vogler et al. (2020) use Jaccard similarity in combination with tri-grams to determine the amount of content sharing between

[2] The supplementary materials (scripts and data) to replicate our findings are available at https://osf.io/3hdk4/. Note that due to copyright restrictions, we are unable to share the raw data containing the full article texts, with the exception of the human validation datasets.

newspapers. The differing approaches in these two studies are in line with their respective goals. Boumans (2016) is focused on newspaper articles that are based on or resemble external materials, while Vogler et al. (2020) are searching for near-duplicate articles. In the former case, the prime goal is to detect (possibly partial) overlap between the article and external material, while in the latter the emphasis is on detecting duplicates. This difference, between overlap and duplicates is also relevant in the context of content and topic diversity.

In their very basis, automated methods for analyzing text are all based on linguistic measures, usually word counts. But through the choice of similarity measure and feature selection/weighting the substantive meaning and interpretation of such measures differs. Jaccard similarity only considers the presence or absence of specific words in a pair of texts, while cosine similarity takes into account the relative frequency of those words. Similarly, on the one hand, tri-grams (groups of three words) are dependent on word order, and as a result much more strict than regular word frequencies when it comes to measuring similarity. On the other hand, tf-idf weighting as applied by Boumans (2016) increases the weight of more informative words (Jones, 1972). Tf-idf weighting is based on the assumption that words occurring in many different articles are less informative than words occurring in only a few articles. Thus, using cosine similarity with idf-weighted word frequencies provides a substantially different measure of similarity than using Jaccard similarity with tri-grams. The former is oriented towards finding overlap, while the latter is oriented towards finding (near-)duplicates (i.e. the exact same words in the exact same order). Because we are interested in the extent to which news articles cover the same events, and not whether one article is a literal copy of another, we use cosine similarity in combination with tf-idf weighting to measure topic diversity. We also compare the sentiment between article pairs that are about the same topic to account for possible differences in the sentiment with which a topic is discussed. As our measure of topic diversity is in its basis a similarity measure, diversity is considered to be the opposite of similarity for the remainder of this paper.

Metrics

To construct the topic diversity measure, we use the functions provided by the R package Quanteda (Benoit et al., 2018). First, the raw articles, consisting of lemmas, are converted into a document-feature matrix, which represents the articles as vectors of word/feature frequencies. This is done by day and country. The feature frequencies are weighted using the idf weighting scheme, and a cosine similarity matrix is constructed. This matrix contains cosine similarity values for every possible article pair on the given day, except the comparisons of articles with themselves. We discard comparisons of articles with other articles from the same newspaper, since we are interested in diversity between rather than within newspapers—external rather than internal diversity. What remains are the similarity values for each article when compared to all articles published in the other two newspapers on the same day. These values are aggregated, so that each article in newspaper A gets a mean and maximum similarity with the articles in newspaper B (or C). These maximum values, indicating the most similar article pairs between newspapers, are used as our indicator for topic diversity.

Although the comparisons between individual articles are by definition symmetric (article A is as much like article B as article B is like article A), this symmetry disappears when using aggregate measures. In our specific application, using the most similar article pairs as indicator of diversity, it is possible that article A might be most like article B, but article B is more like article C than article A. An example of this is provided in table 2, containing fictional cosine similarity scores between three articles from newspaper A and three articles from newspaper B. This table shows that articles A3 and B3 (.8) and A2 and B1 (.9) are most alike. However, article A1 is most like B1 (.6), while article B2 is most like A1 (.4). On average, then, the similarity of newspaper A with newspaper B is $\frac{.6+.9+.8}{3} = .77$ while the similarity of newspaper B with newspaper A is $\frac{.4+.9+.8}{3} = .70$. Because of this asymmetry, newspaper comparisons for all countries are made both ways (so A-B, A-C, B-C as well as B-A, C-A, C-B).

(Table 2 around here)

To construct our final topic diversity measure, we use the inverted scores of the most similar article pairs[3]. Inversion of the scores is done to conform to the theoretical concept of diversity, with higher values indicating higher diversity rather than similarity.

[3] i.e. the articles in A that are most like the articles in B or C, and vice versa

The article pairs are then split into two categories based on their diversity scores, one with pairs that are about the same topic, and one with pairs that are not. The cutoff value for this split is based on the manual validation results (see below). Then H_1 is tested by evaluating the trend in the weighted percentage[4] of article pairs that are about the same topic.

To test H_2, an indicator for the difference in sentiment between the newspapers is constructed by comparing the sentiment of article pairs that are about the same topic (see above). Article-level sentiment scores are generated using the method described in De Vries (2022). This method consists of a computer-generated dictionary in each of the four languages, used to classify trinary (negative, neutral, positive) sentiment in each sentence of each news article. The sentence-level scores are aggregated and weighted on length to construct a sentiment score at the article level. A sentiment diversity measure is constructed by 1) subtracting the sentiment of article pairs about the same topic from each other, 2) taking the absolute value, and 3) dividing those by 2 (as the original sentiment scores range from +1 to -1). As such, we evaluate the difference in the general sentiment context used to describe a topic. Conceptually, the measure indicates the extent to which articles that already talk about the same topic differ from each other in their general presentation (in terms of positivity/negativity) of

[4] by the average article length in number of words

that topic.[5] While a similar approach could have adopted for other content features (e.g. presence of actors, or even more substantial frames), we consider general sentiment as a generic and widely studied content feature of media coverage that drives media effects on e.g. public opinion (Vliegenthart et al., 2021). Even though the source and target of the sentiment are unknown, the fact that a topic is discussed within a specific sentiment context is a relevant content feature of media coverage, in particular when we consider external diversity. The degree of congruence across coverage across outlets provides information about the similarity of and thus of the level of external diversity as we conceptualize this in our paper. When the sentiment differs between two articles about the same topic, this is an indicator of the diversity of the context within which information on this topic is provided, and more specifically, different words imply that a set of articles, even though they have the same topical focus, say something different, and thus add different information and interpretation to the news supply.

Validation

To validate the automatically generated topic diversity metric, we use a small-scale manually coded dataset consisting of 200 Norwegian and 200 UK article-pairs. These articles are coded based on the following question: Do these two newspaper articles

[5] Variations in the sentiment related to different viewpoints within a topic are not considered, due to measurement at the article level.

cover the same topic? There are two answer options (yes or no). The random sample of article-pairs to manually code is constructed in a stratified way based on the automated coding. The topic diversity scores are binned into 5 groups (0-.2, .2-.4, .4-.6, .6-.8, .8-1), and from each of these groups a sample of 40 article-pairs is drawn for a total sample of 200 article-pairs for Norway and the UK . From the UK dataset, a random subsample of 20 article-pairs is used to test the intercoder reliability between two coders (the main author and a student assistant). The result of this test is a Krippendorff's alpha of .88, which is more than sufficient. The student assistant, being proficient in both English and Norwegian, has coded the remaining UK article pairs and all 200 Norwegian article pairs.

The human-coded validation results[6] show a strong correlation with our automated topic diversity measure in both the UK (*r*(198) = -.79, *p* <.001) and Norway (*r*(198) = -.73, *p* <.001).

These correlations are visualized as a box plot (Figure 1), showing that topic diversity for article pairs that are about the same topic is generally below .6, while topic diversity for article pairs that are about different topics is generally above .6. Thus a topic diversity value of below .6 (or a cosine similarity value above .4) is used as a cut-off point to determine which article pairs are about the same topic. Extensive validation

[6] Short examples of articles at various levels of topic diversity can be found in the appendix, while more extensive examples can be found in the validation dataset that is part of the supplementary materials at https://osf.io/3hdk4/

of the sentiment analysis method used in this paper is presented in De Vries (2022). The general performance ranges from .61 to 64 (weighted F_1), indicating that in a majority of cases sentiment in sentences is classified correctly as either positive, negative or neutral. In addition, the aggregated nature of the current analyses (from sentence to article) and the near-normal distribution of errors allows a majority of the random errors to cancel each other out at the article level.

(Figure 1 around here)

In addition to the manual validations described above, we have also conducted a face validity test of the topic diversity measure, by comparing the percentage of articles pairs with a topic diversity below .6 and published on the same day to the percentage when one of the articles in a pair is published up to a week later. The reasoning behind this test is that due to the newspaper news cycle spanning a single day, article pairs should be less diverse when they are both published on the same day, than when one is published on a subsequent day. The linear regression results (with standard errors) in Figure 2 test this assumption at the country level. Topic diversity between article pairs published on the same day is indeed lower than when one of the articles is published on a subsequent day, providing additional validity to the topic diversity measure.

(Figure 2 around here)

Results

In all plots presented below, the black lines show trends, while the grey lines show LOESS-smoothed observations, using a rolling window over 15% of the data. Descriptive statistics of the variables used in the figures and regression models can be found in the appendix. The results in Figure 3 show the percentage of article pairs that are about the same topic for each newspaper pair, as well as the average per country (rows) by country (columns). On average, there are 116 article pairs per newspaper pair and day, with a standard deviation of 88.

(Figure 3 around here)

Topic diversity

As shown in the bottom row of figure 3, there is for all countries except Denmark a modest decrease in the percentage of article pairs that are about the same topic. In Norway and the Netherlands these trends are quite comparable, while the trend is somewhat more pronounced in the UK and entirely absent in Denmark. In general, no increase in the percentage of article pairs about the same topic is found in any of the countries. Thus the assumption that diversity that decreases over time, as formulated in $H_{1,}$ is not supported.

(Table 3 around here)

A rejection of H_1 is also confirmed by the regression results presented in Table 3. A lagged (by one day) dependent variable is included in these models as a control variable to account for external factors that influence the amount of topic diversity on consecutive days. The standardized regression coefficients show that the effect of time on the percentage of article pairs that share the same topic is in all cases negative and highly significant. The strength of the effect is also comparable between the different countries, except for Denmark, where it is an order of magnitude smaller. Based on the coefficients it is also clear that generally speaking the percentage of article pairs with the same topic is highest between left- and right-wing newspapers and decreases significantly between either left- or right-wing broadsheets and the tabloid newspaper. A notable exception here is Norway, where there is not much difference between the newspaper pairs at all. With an R^2 between .40 (UK) and .09 (Norway) these regression models differ substantially in the amount of variance they explain. In general, a low R^2 is to be expected, as many possible causes for a temporary increase or decrease in topic diversity are not included in these models. However, based on the amount of explained variance, it seems like diversity between newspapers is explained much more by the included predictors in the UK than it is in Norway. Also, the high coefficient of the lagged dependent variable indicates there is more temporal invariance in the amount of article pairs that are about the same topic in the UK than in any of the other countries. In Norway, the included variables do not add to the explanation for the amount of diversity at all, except for time.

Sentiment diversity

(Figure 4 around here)

In contrast to the findings relating to topic diversity, there are no clear trends visible in the bottom row of Figure 4. This figure shows for each country the average normalized level of sentiment diversity between articles that are about the same topic. There do not appear to be any substantial differences between the countries, which is supported by the standardized regression results in Table 4. As the explained variance of the Norwegian regression model is exactly 0, we disregard this model entirely with the comment (like with topic diversity) that the included variables do not predict sentiment diversity at all. Regarding the other countries, the most notable coefficients in these regressions are those indicating the newspaper pairs. All of these are comparable in size and significance between the countries, indicating that the left- and right-wing newspapers differ substantially more from the tabloid newspaper than from each other. In general, the absence of clear trends is also reflected in the insignificant effect of time and more generally in the negligible amount of explained variance in each of the models. The one exception is the UK, where time does have a significant positive effect on sentiment diversity. In combination with Figure 4 these results lead to a rejection of H_2 as there is no clear decrease of sentiment diversity over time in any of the countries.

(Table 4 around here)

Cross-national comparison

Considering the results presented above, and notwithstanding its different media and political system, the UK does not stand out at country level when it comes to topic diversity (bottom row in Figure 3), and the downward trend is remarkably similar to that in other countries. However, a systematic newspaper by newspaper comparison in the various countries as presented in Figure 3 does reveal a different trend in the United Kingdom: topic diversity between the right-leaning (The Daily Telegraph) and tabloid (The Sun) newspaper decreases substantially, while it increases substantially between these newspapers and the left-leaning Guardian. Additionally the results in Figure 4, while not indicating substantial differences between the UK newspapers, do show a modest trend towards increasing sentiment diversity. This indicates that the UK newspapers increasingly differ in the sentiment with which they cover topics, even when they cover the same topics. Both results seem to be indicative for a kind of topic and sentiment polarization or segmentation, which is not found in any of the other countries. In answer to RQ_1 there indeed seems to be a polarizing effect of the political and media system in the UK, even though it is not visible in the country level analyses.

Conclusion

Our study scrutinizes an often-made assumption about journalistic content—that diversity has decreased due to developments such as increasing competition and economic pressures. We put this assumption to a rigid, cross-national empirical test and despite plausible and compelling claims, we find little evidence for a decreasing trend in diversity. If anything, we find evidence for *increasing* topic diversity. The quickly changing media environment might thus not have had the anticipated effect. Potentially, we can attribute the lack of a decreasing diversity trend to the changing role of traditional media, and in particular newspapers. Mellado and colleagues (2017) demonstrate that journalistic role perceptions have become increasingly hybrid. Their study provides compelling evidence about the multilayered hybridization of journalistic cultures at the performative level. Professional roles are varied as well as fluid and dynamic. Compared to several decades ago, the importance of 'bringing the news (first)' has substantially decreased for traditional media. Online and social media have taken over the role of being the first ones to bring news to large audiences. Printed newspapers will only in a minority of instances be the source that brings news and events first. They have moved in the direction of providing interpretation, analysis and opinions instead (Esser & Umbricht, 2014; Soontjens, 2019; e.g. Strömbäck & Aalberg, 2008). This new, more interpretative, role inherently goes hand in hand with a diversification of content, newspapers developing a more distinct profile, that might have canceled out the pressure to less diverse content. It also emphasizes the continuing relevance of newspapers, as they offer something that other news sources do not. In this way, diversity between newspapers continues to affect the news diversity of the entire media landscape. Future research could try to link the results of our study with longitudinal and cross-national data on (changing) role perceptions, for example from the Worlds of Journalism project (Hanitzsch et al., 2019), or to a more detailed analysis of editorial strategies of individual media outlets.

Our findings are indicative of increased instead of decreased diversity, but they do not provide definite answers. First, it might be that decreases in diversity have taken place before the period we scrutinized. After all, changes in the media landscapes and financial pressures originate from well before the end of the previous century (Lewis et al., 2008). Practical constraints, most notably the absence of digital archives for pre-2000 content for a substantial part of our sources, refrain us from establishing whether this is indeed the case.

Second, our measure of diversity is not comprehensive—we focus solely on external diversity and we look only at diversity in terms of topics and sentiment. We do not look at actors, for instance, or at framing. It might well be that external topic diversity is high according to our measure but that internal diversity is low; the readers of a given newspaper might be confronted with low internal diversity although they could find more diverse news by looking also at other outlets. Thus, our measure only partially grasps what actual news consumers are confronted with, as only part of them read several newspapers. This limitation, next to the lack of a comparison with the actual diversity in preferences and opinions in society (through e.g. survey research), thwarts the opportunity to make a comprehensive assessment of our findings and compare them against ideal types such as reflective and open diversity (Van Cuilenburg, 1999).

Third, while our study is cross-national and we engaged in some tentative comparative analyses, we have not been able to truly capitalize on this in terms of providing a systematic comparative account. We would need more countries to statistically test the impact of country level features. Nonetheless, we have found some tantalizing evidence that the increased competition in the Liberal media system of the UK (Hallin & Mancini, 2004) does not lead to relatively less diversity over time than in the Democratic-Corporatist media systems in Denmark, The Netherlands and Norway. Rather, it seems to lead to a more polarized form of diversity, with one newspaper being substantially different from the other two.

Finally, the interpretability of our results is limited because our news diversity measures disregards word order and syntax, and are measured at the article level rather than for example sentence level. The sentiment diversity measure therefore indicates the difference in general tone when the same topic is discussed, but does not account for viewpoint diversity within articles, nor the sentiment associated with specific viewpoints. For topic diversity, the loss of word order and syntax opens up the hypothetical possibility that two articles either use very similar words to describe very different topics or use very dissimilar words to describe very similar topics. However, such cases do not seem likely, as our topic diversity measure (a combination of cosine similarity with tf-idf feature weighting) is shown to be reliable and valid. It strongly correlates with human classifications, works well with languages other than English, and is not overly complex or computationally expensive.

In addition, our study is the first to show trends of increasing news diversity while combining a multiple country comparison with a nearly two-decade time span. One thing that stands out in particular is that the findings are largely similar across countries. In that sense, our paper offers a robust assessment of general patterns that hold across contexts. Patterns, as shown here, that turned out to be quite different from our theoretical expectations.

References

Amsalem, E., Fogel-Dror, Y., Shenhav, S. R., & Sheafer, T. (2020). Fine-Grained Analysis of Diversity Levels in the News. *Communication Methods and Measures*, *14*(4), 266–284. https://doi.org/10.1080/19312458.2020.1825659

Baker, C. E. (2007). *Media concentration and democracy: Why ownership matters*. Cambridge University Press.

Beckers, K., Masini, A., Sevenans, J., van der Burg, M., De Smedt, J., Van den Bulck, H., & Walgrave, S. (2019). Are newspapers' news stories becoming more alike? Media content diversity in Belgium, 1983. *Journalism*, *20*(12), 1665–1683. https://doi.org/10.1177/1464884917706860

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). Quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774. https://doi.org/10.21105/joss.00774

Boczkowski, P. J., & Santos, M. de. (2007). When More Media Equals Less News: Patterns of Content Homogenization in Argentina’s Leading Print and Online Newspapers. *Political Communication*, *24*(2), 167–180. https://doi.org/10.1080/10584600701313025

Boukes, M., Van de Velde, B., Araujo, T., & Vliegenthart, R. (2020). What’s the tone? Easy doesn’t do it: Analyzing performance and agreement between off-the-shelf sentiment analysis tools. *Communication Methods and Measures*, *14*(2), 83-104. https://doi.org/10.1080/19312458.2019.1671966

Boumans, J. (2016). *Outsourcing the news? An empirical assessment of the role of sources and news agencies in the contemporary news landscape*. Amsterdam: University of Amsterdam.

Boumans, J., Trilling, D., Vliegenthart, R., & Boomgaarden, H. (2018). The Agency Makes the (Online) News World go Round: The Impact of News Agency Content on Print and Online News. *International Journal of Communication*, *12*(0), 22.

Brüggemann, M., Esser, F., & Humprecht, E. (2012). The Strategic Repertoire of Publishers in the Media Crisis. *Journalism Studies*, *13*(5-6), 742–752. https://doi.org/10.1080/1461670X.2012.664336

Curran, J. (2010). The Future of Journalism. *Journalism Studies*, *11*(4), 464–476. https://doi.org/10.1080/14616701003722444

Dailey, L., Demo, L., & Spillman, M. (2005). The Convergence Continuum: A Model for Studying Collaboration Between Media Newsrooms. *Atlantic Journal of Communication*, *13*(3), 150–168. https://doi.org/10.1207/s15456889ajc1303_2

Damstra, A., & De Swert, K. (2020). The making of economic news: Dutch economic journalists contextualizing their work. *Journalism*, 1464884919897161. https://doi.org/10.1177/1464884919897161

Day, A. G., & Golan, G. (2005). Source and content diversity in Op-Ed Pages: Assessing editorial strategies in The New York Times and the Washington Post. *Journalism Studies*, *6*(1), 61–71. https://doi.org/10.1080/1461670052000328212

De Vreese, C.H., Esser, F., Hopmann, D. N., Esser, F., & Hopmann, D. N. (2016). *Comparing Political Journalism*. Routledge. https://doi.org/10.4324/9781315622286

De Vries, E. (2022). The Sentiment is in the Details. A Language-agnostic Approach to Sentence-level Sentiment Analysis in News Media. *Computational Communication Research*.

Entman, R. M. (2003). Cascading Activation: Contesting the White House’s Frame After 9/11. *Political Communication*, *20*(4), 415–432. https://doi.org/10.1080/10584600390244176

Esser, F., & Umbricht, A. (2014). The Evolution of Objective and Interpretative Journalism in the Western Press: Comparing Six News Systems since the 1960s.

Journalism & Mass Communication Quarterly, *91*(2), 229–249. https://doi.org/10.1177/1077699014527459

Farrell, D. M. (2011). *Electoral Systems: A Comparative Introduction*. Macmillan International Higher Education.

Gandy, O. H. (1980). Information in health: Subsidised news. *Media, Culture & Society*, *2*(2), 103–115. https://doi.org/10.1177/016344378000200201

Hallin, D. C., & Mancini, P. (2004). Comparing Media Systems: Three Models of Media and Politics. In *Cambridge Core*. /core/books/comparing-media-systems/B7A12371782B7A1D62BA1A72C1395E43; Cambridge University Press. https://doi.org/10.1017/CBO9780511790867

Hanitzsch, T., Hanusch, F., Ramaprasad, J., & Beer, A. S. de (Eds.). (2019). *Worlds of Journalism: Journalistic Cultures Around the Globe* (pp. 448 Pages). Columbia University Press.

Hanusch, F. (2015). Transformative Times: Australian Journalists' Perceptions of Changes in Their Work. *Media International Australia*, *155*(1), 38–53. https://doi.org/10.1177/1329878X1515500106

Hendrickx, J., & Ranaivoson, H. (2019). Why and how higher media concentration equals lower news diversity The Mediahuis case. *Journalism*, 1464884919894138. https://doi.org/10.1177/1464884919894138

Hlavac, M. (2018). *Stargazer: Well-Formatted Regression and Summary Statistics Tables.* https://CRAN.R-project.org/package=stargazer.

Hotelling, H. (1929). Stability in Competition. *The Economic Journal*, *39*(153), 41–57. https://doi.org/10.2307/2224214

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*, 11–21.

Joris, G., Grove, F. D., Damme, K. V., & Marez, L. D. (2020). News Diversity Reconsidered: A Systematic Literature Review Unraveling the Diversity in Conceptualizations. *Journalism Studies*, *21*(13), 1893–1912. https://doi.org/10.1080/1461670X.2020.1797527

Jyrkiäinen, J., & Heinonen, A. (2012). Finnish Journalists: The Quest for Quality amidst New Pressures. In *The Global Journalist in the 21st Century*. Routledge.

Lacy, S., & Simon, T. F. (1993). *The economics and regulation of United States newspapers*. Ablex Publishing Corporation.

Langer, A. I., & Gruber, J. B. (2021). Political agenda setting in the hybrid media system: Why legacy media still matter a great deal. *The International Journal of Press/Politics*, *26*(2), 313-340. https://doi.org/10.1177/1940161220925023

Lee, J. K. (2007). The Effect of the Internet on Homogeneity of the Media Agenda: A Test of the Fragmentation Thesis. *Journalism & Mass Communication Quarterly*, *84*(4), 745–760. https://doi.org/10.1177/107769900708400406

Lewis, J., Williams, A., & Franklin, B. (2008). Four Rumours and an Explanation. *Journalism Practice*, *2*(1), 27–45. https://doi.org/10.1080/17512780701768493

Masini, A., & Van Aelst, P. (2017). Actor diversity and viewpoint diversity: Two of a kind? *Communications*, *42*(2), 107–126. https://doi.org/10.1515/commun-2017-0017

Mellado, C., Hellmueller, L., Márquez-Ramírez, M., Humanes, M. L., Sparks, C., Stepinska, A., ... & Wang, H. (2017). The hybridization of journalistic cultures: A comparative study of journalistic role performance. *Journal of Communication*, *67*(6), 944-967. https://doi.org/10.1111/jcom.12339

Napoli, P. M. (1999). Deconstructing the diversity principle. *Journal of Communication*, *49*(4), 7–34. https://doi.org/10.1111/j.1460-2466.1999.tb02815.x

Newman, N., Fletcher, R., Schulz, A., Andi, S., Robertson, C. T., & Nielsen, R. K. (2021). Reuters Institute digital news report 2021. *Reuters Institute for the Study of Journalism*.

Nivre, J., Abrams, M., Agić, Ž., Ahrenberg, L., Antonsen, L., Aplonova, K., Aranzabe, M. J., Arutie, G., Asahara, M., Ateyah, L., Attia, M., Atutxa, A., Augustinus, L., Badmaeva, E., Ballesteros, M., Banerjee, E., Bank, S., Barbu Mititelu, V., Basmov, V., … Zhu, H. (2018). *Universal dependencies 2.3*.

Paulussen, S. (2012). Technology and the Transformation of News Work: Are Labor Conditions in (Online) Journalism Changing? In E. Siapera & A. Veglis (Eds.), *The*

Handbook of Global Online Journalism (pp. 192–208). Wiley-Blackwell. https://doi.org/10.1002/9781118313978.ch11

Picard, R. G. (2014). Twilight or New Dawn of Journalism? *Journalism Practice*, *8*(5), 488–498. https://doi.org/10.1080/17512786.2014.905338

Raeymaeckers, K., Paulussen, S., & De Keyser, J. (2012). A Survey of Professional Journalists in Flanders (Belgium). In *The Global Journalist in the 21st Century*. Routledge.

Rodgers, S., Thorson, E., & Antecol, M. (2000). “Reality” in the St. Louis Post-Dispatch. *Newspaper Research Journal*, *21*(3), 51–68. https://doi.org/10.1177/073953290002100305

Roessler, P. (2007). Media Content Diversity: Conceptual Issues and Future Directions for Communication Research. *Annals of the International Communication Association*, *31*(1), 464–520. https://doi.org/10.1080/23808985.2007.11679073

Schudson, M. (2011). *The Sociology of News (2nd edition)*. WW Norton & Company.

Sjøvaag, H. (2016). Media diversity and the global superplayers: Operationalising pluralism for a digital media market. *Journal of Media Business Studies*, *13*(3), 170–186. https://doi.org/10.1080/16522354.2016.1210435

Soontjens, K. (2019). The Rise of Interpretive Journalism. *Journalism Studies*, *20*(7), 952–971. https://doi.org/10.1080/1461670X.2018.1467783

Soroka, S., & McAdams, S. (2015). News, politics, and negativity. *Political Communication*, *32*(1), 1-22. https://doi.org/10.1080/10584609.2014.881942

Straka, M., & Straková, J. (2017). Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 88–99.

Strömbäck, J., & Aalberg, T. (2008). Election News Coverage in Democratic Corporatist Countries: A Comparative Study of Sweden and Norway. *Scandinavian Political Studies*, *31*(1), 91–106. https://doi.org/10.1111/j.1467-9477.2008.00197.x

Van Aelst, P., Strömbäck, J., Aalberg, T., Esser, F., de Vreese, C., Matthes, J., Hopmann, D., Salgado, S., Hubé, N., Stępińska, A., Papathanassopoulos, S., Berganza, R., Legnante, G., Reinemann, C., Sheafer, T., & Stanyer, J. (2017). Political communication in a high-choice media environment: A challenge for democracy? *Annals of the International Communication Association*, *41*(1), 3–27. https://doi.org/10.1080/23808985.2017.1288551

Van Aelst, P., & Walgrave, S. (2016). Information and Arena: The Dual Function of the News Media for Political Elites. *Journal of Communication*, *66*(3), 496–518. https://doi.org/10.1111/jcom.12229

Van Cuilenburg, J. (1999). On Competition, Access and Diversity in Media, Old and New: Some Remarks for Communications Policy in the Information Age. *New Media & Society*, *1*(2), 183–207. https://doi.org/10.1177/14614449922225555

Vliegenthart, R., Damstra, A., Boukes, M., & Jonkman, J. (2021). *Economic News: Antecedents and Effects*. Cambridge: Cambridge University Press.

Vogler, D., Udris, L., & Eisenegger, M. (2020). Measuring Media Content Concentration at a Large Scale Using Automated Text Comparisons. *Journalism Studies*, *0*(0), 1–20. https://doi.org/10.1080/1461670X.2020.1761865

Walgrave, S., Boydstun, A. E., Vliegenthart, R., & Hardy, A. (2017). The nonlinear effect of information on political attention: media storms and US congressional hearings. *Political Communication*, *34*(4), 548-570. https://doi.org/10.1080/10584609.2017.1289288

Tables

Table 1: Newspaper sample

Denmark	Netherlands	Norway	UK
Politiken^1	de Volkskrant^1	Dagbladet^x	The Guardian^1
Jyllands-Posten^2	NRC Handelsblad^2	Aftenposten^2	The Daily Telegraph^2
Ekstra Bladet^3	de Telegraaf^3	VG^3	The Sun^3

Notes: ^{1}Left-leaning; ^{2}Right-leaning; ^{3}Tabloid/Popular; ^{x}Dagbladet is a left-leaning tabloid rather than a broadsheet

Table 2: Similarity matrix example

	Article B1	Article B2	Article B3
Article A1	*.6*	**.4**	.1
Article A2	***.9***	.1	.2
Article A3	.1	.3	***.8***

Notes: Row maximum is indicated in *italic*, column maximum is indicated in **bold**

Table 3: Regression results: Percentage of article pairs about the same topic

	Denmark	Netherlands	Norway	UK
Percentage of article pairs (lagged)	.148***	.110***	.250***	.343***
	(.007)	(.008)	(.007)	(.007)
Time (in years)	−.018***	−.146***	−.114***	−.113***
	(.006)	(.007)	(.007)	(.006)
Left-wing/Tabloid	−.821***	−.469***	.037**	−.855***
	(.016)	(.017)	(.017)	(.017)
Right-wing/Tabloid	−.820***	−.772***	−.021	−.430***
	(.016)	(.018)	(.016)	(.016)
Constant	.547***	.411***	−.005	.445***
	(.011)	(.012)	(.012)	(.012)
Observations	21,686	17,291	20,097	16,688
R^2	.224	.166	.085	.396
Adjusted R^2	.224	.166	.085	.396
Residual Std. Error	.881	.913	.957	.777
F Statistic	1,564.952***	861.050***	468.263***	2,734.278***

Note: *p<0.1; **p<0.05; ***p<0.01

Table 4: Regression results: Sentiment diversity

	Denmark	Netherlands	Norway	UK
Sentiment diversity (lagged)	.006	.008	.001	−.009
	(.008)	(.008)	(.008)	(.008)
Time (in years)	.012	−.013*	−.010	.121***
	(.008)	(.008)	(.008)	(.008)
Left-wing/Tabloid	.139***	.190***	.001	.244***
	(.019)	(.019)	(.019)	(.019)
Right-wing/Tabloid	.171***	.207***	.005	.232***
	(.020)	(.019)	(.018)	(.019)
Constant	−.091***	−.127***	−.004	−.163***
	(.012)	(.013)	(.013)	(.014)
Observations	14,328	16,566	17,050	16,477
R^2	.007	.009	0.000	.027
Adjusted R^2	.007	.009	−0.000	.027
Residual Std. Error	.962	.994	.989	.978
F Statistic	24.851***	39.193***	.498	113.992***

Note: *p<0.1; **p<0.05; ***p<0.01

Figures

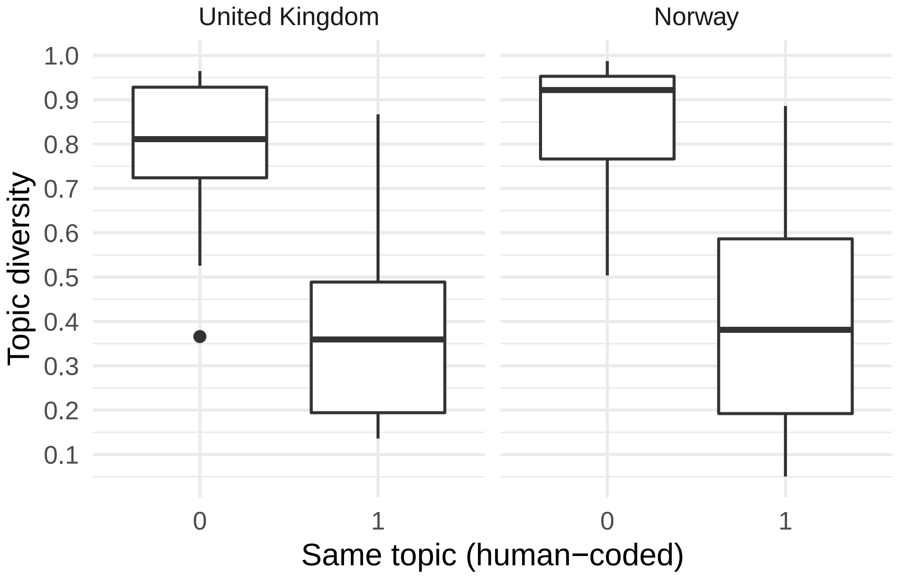


Figure 1: Box plot of topic diversity for articles that are (1) or are not (0) about the same topic

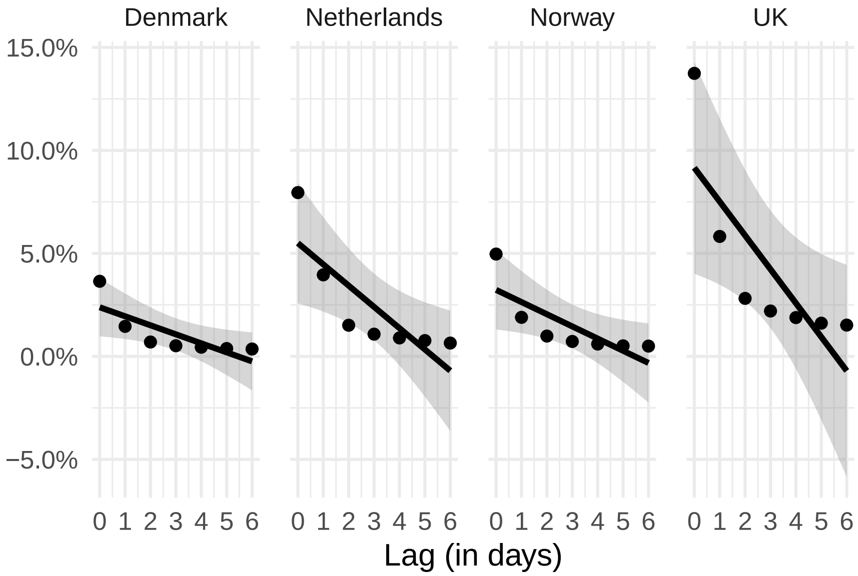


Figure 2: Linear regression showing the effect of publication lag (i.e. article pairs that are published on different days, with a lag from 0-6 days) on the percentage of articles with a topic diversity below .6

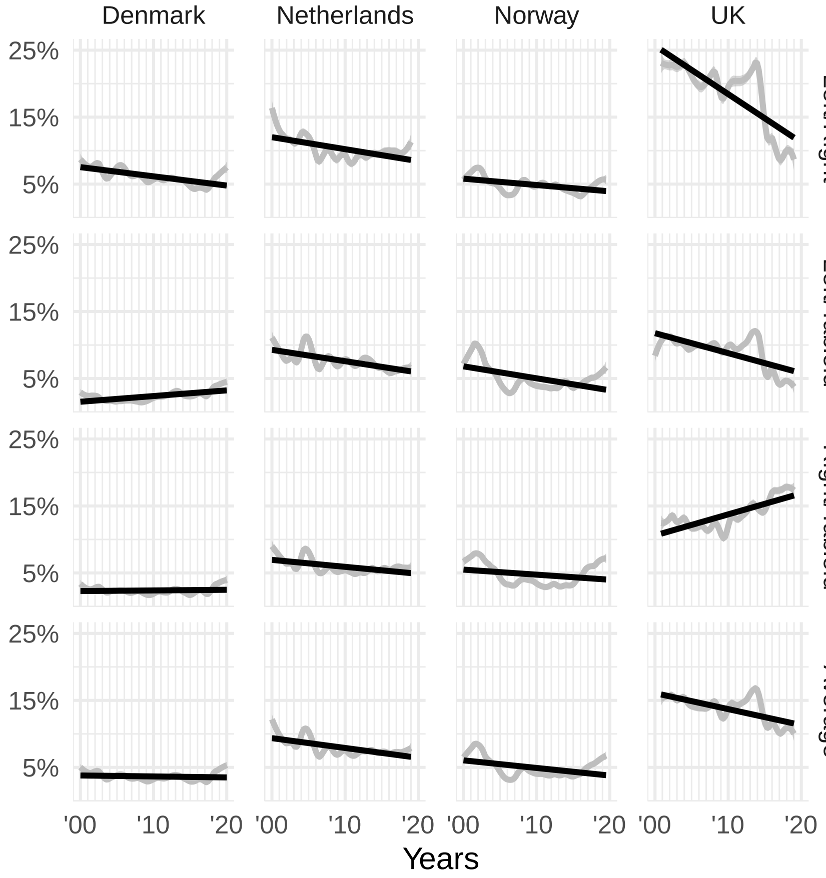


Figure 3: Weighted (by word count) percentage of article pairs with a diversity of .6 or lower, by country and newspaper pair

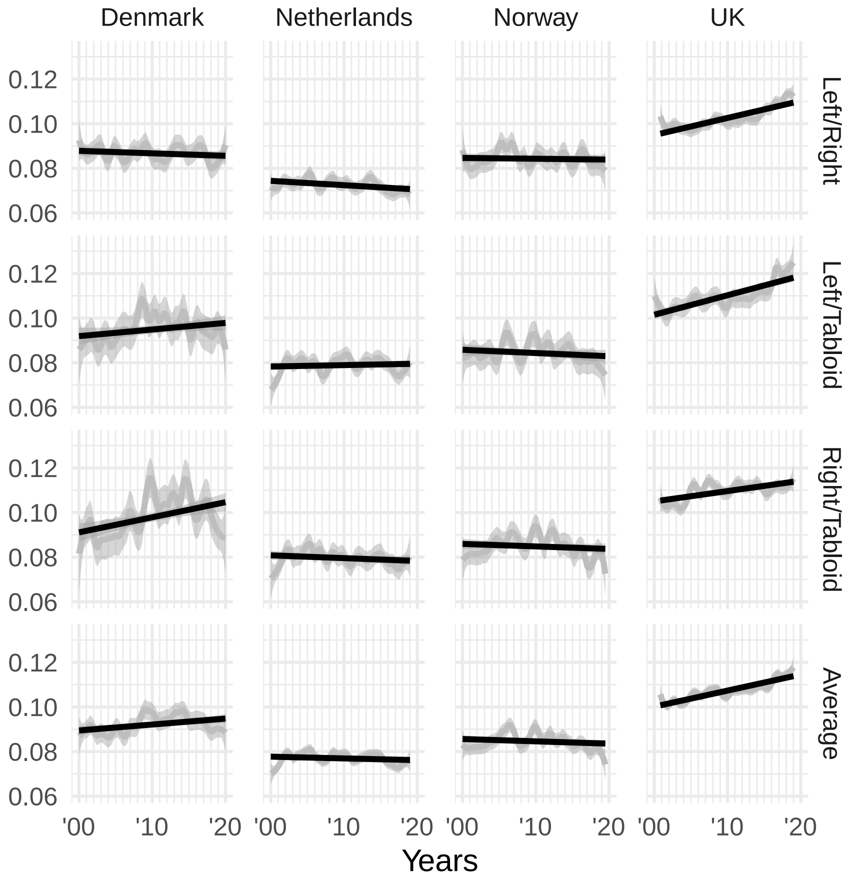


Figure 4: Sentiment diversity of article pairs with a diversity of .6 or lower, by country and newspaper pair

Supplementary Materials:
Telling a Different Story

Erik de Vries, Stefaan Walgrave, Rens Vliegenthart

Overview of irrelevant article categories

Weather reports, traffic announcements, “what happened this date XX years ago”, birthdays, brief notices about appointments, training/fitness (not if focusing on health, but articles with an emphasis on training methods or “how to get a perfect whatever. . . ”), testing of products (mobile phones, cars, foods, etc.), articles that summarise current news or refer to several stories, obituaries, restaurant reviews, food recipes. Proverbs and short quotes without context.

Reports from all sports and sporting events (including chess and bridge). Interviews with athletes / performers, reports about a particular club or team (without wider discussions about politics, government funding or regulations).

Reports from concerts, theater and entertainment events, reviews of books, music, etc. Prizes, awards, hirings in the cultural/entertainment sector. Trends, lifestyle, entertainment industry, television shows, celebrities, etc. “Sensational stories” with no real news value (ie. penis in ketchup bottle, kissing record and that sort of stuff). Also, the royal house as entertainment news: royal divorces, weddings, baptisms, privacy/past/activities of royals.

Descriptive statistics

Table 1: Descriptive statistics by country and comparison type

	Mean	SD	Skewness	Kurtosis	N
Denmark: Left-wing/Right-wing					
Topic diversity	0.06	0.04	0.90	0.97	7228
Sentiment diversity	0.09	0.05	1.46	5.48	6884
Time (in years)	10.03	5.78	-0.01	-1.20	7228
Denmark: Left-wing/Tabloid					
Topic diversity	0.02	0.03	1.89	5.31	7226
Sentiment diversity	0.10	0.07	1.55	6.06	5450
Time (in years)	10.04	5.78	-0.01	-1.20	7226
Denmark: Right-wing/Tabloid					
Topic diversity	0.02	0.03	1.96	5.42	7233
Sentiment diversity	0.10	0.07	1.43	3.22	4965
Time (in years)	10.03	5.78	-0.01	-1.20	7233
Netherlands: Left-wing/Right-wing					
Topic diversity	0.10	0.05	0.55	0.34	5834
Sentiment diversity	0.07	0.03	1.19	4.85	5796
Time (in years)	9.50	5.48	0.00	-1.20	5834
Netherlands: Left-wing/Tabloid					
Topic diversity	0.08	0.04	0.77	1.35	5729
Sentiment diversity	0.08	0.03	0.95	2.78	5610
Time (in years)	9.62	5.43	-0.02	-1.18	5729
Netherlands: Right-wing/Tabloid					
Topic diversity	0.06	0.04	0.86	0.97	5729
Sentiment diversity	0.08	0.04	1.36	5.00	5507
Time (in years)	9.62	5.43	-0.02	-1.18	5729
Norway: Left-wing/Right-wing					
Topic diversity	0.05	0.04	1.13	2.04	6583
Sentiment diversity	0.08	0.04	1.30	4.08	6264
Time (in years)	9.50	5.53	0.05	-1.15	6583
Norway: Left-wing/Tabloid					
Topic diversity	0.05	0.05	1.51	3.26	6630
Sentiment diversity	0.08	0.05	1.41	3.70	5768
Time (in years)	9.50	5.52	0.05	-1.15	6630
Norway: Right-wing/Tabloid					
Topic diversity	0.05	0.04	1.35	2.50	6885
Sentiment diversity	0.08	0.05	1.51	5.09	6405
Time (in years)	9.82	5.63	-0.02	-1.20	6885
United Kingdom: Left-wing/Right-wing					

Topic diversity	0.19	0.08	0.14	-0.50	5323
Sentiment diversity	0.10	0.02	0.98	4.65	5305
Time (in years)	9.81	5.29	0.06	-1.20	5323
United Kingdom: Left-wing/Tabloid					
Topic diversity	0.09	0.05	0.60	0.72	5975
Sentiment diversity	0.11	0.04	1.09	4.08	5889
Time (in years)	9.81	5.56	-0.06	-1.23	5975
United Kingdom: Right-wing/Tabloid					
Topic diversity	0.14	0.05	0.31	-0.02	5391
Sentiment diversity	0.11	0.03	0.47	1.52	5382
Time (in years)	9.80	5.30	0.06	-1.22	5391

Example article pairs at different levels of diversity

Table 2: Examples of various diversity scores from British newspapers

		Div.
TECH GROUP MOULDINGS SHED 20 WORKERS The Tech Group Mouldings company in Mulhuddart, Co Dublin, yesterday shed 20 workers. The firm makes mouldings for Xerox and Hewlett Packard. PUB-TYPE: Newspaper.	Call centre managers see surge in salaries. MANAGERS of larger call centres have seen their salaries rise more than 20pc during the past year, according to research by management consultants Hay Group. The salaries of managers of call centres with more than 500 staff have increased from an average of pounds 59,245 to pounds 71,586 over the past 12 months. In comparison, Bank of England figures show headline average earnings growth rose 4.5pc in the year to May. Call centre team leader's salaries have risen by nearly 11pc from an average of pounds 21,408 to pounds 23,698. Anthony McNulty, at Hay Group, said: "Call centres are trying to shed their sweat shop image and remodel themselves as contact centres, providing high level technical advice to customers. This in turn means that staff must be of a correspondingly high calibre." [PS]City: [ES].	0.9621

		Div.
Edition 1; Scotland '£500m in coins' on sunk boat. THE first HMS Victory is to be raised from the sea bed 268 years after it sank - along with a possible £500million in gold coins. The vessel, predecessor of Nelson's flagship, went down in a storm off the Channel Islands in 1744, killing about 1,000 soldiers. Some say it was carrying the coins from Lisbon to Britain alongside a bronze cannon collection. The wreck is set to be handed to the Maritime Heritage Foundation and its guns and artefacts displayed in UK museums. But the bulk of any treasure is tipped to go to Odyssey Marine Exploration, a US firm that found the vessel four years ago. GRAPHIC: Treasure .. first Victory.	Edition 2; National Edition Treasure hunters to raise HMS Victory; In Brief. The remains of the original HMS Victory are to be raised from the sea bed 250 years after it sank, it has been reported. The predecessor of Nelson's flagship, it went down in a storm off the Channel Islands in 1744, apparently carrying gold coins worth £500million.	0.4345

		Div.
Edition 1; Scotland No charges for pink cat's owner; News Bulletin. A woman who dyed her cat bright pink to match her hair will have her pet returned to her after the RSPCA decided she had not committed a crime. An officer from the animal charity will visit the cat's owner to offer advice about the potential hazards and consequences of dyeing cats. Natasha Gregory, 22, said she dyed the naturally white cat using food colouring to match her own pink hair. The cat was taken into the RSPCA after it was found in a garden. Miss Gregory, of Swindon, said she got the idea to dye her pet, called Oi! Kitty, from a US television show.	Edition 1; National Edition Return of pink puss. PRANKSTER Natasha Gregory is to be reunited with the cat she dyed pink to match her hair. The RSPCA will hand back the cat, named Oi! Kitty, after a vet ruled it was well cared for. Natasha, 22, who used food dye, said: "I'm glad she's coming home. I won't dye her again." Animal officers who were handed the cat after it was found in a Swindon garden initially feared it had been ill-treated.	0.3363

		Div.
Edition 1; Scotland Ovarian cancer risk cut by low doses of aspirin; NEWS BULLETIN. Taking a third of an aspirin a day could reduce the risk of women developing ovarian cancer, according to research. Women who reported regular use of low-dose aspirin (100mg or less) had a 23 per cent lower chance of developing the cancer when compared with women who did not take it, a US study published in JAMA Oncology, found. However, long-term heavy use of non-steroidal anti-inflammatory drugs - such as ibuprofen - may be associated with an increased risk. Aspirin is typically taken at a low-dose to prevent heart disease. Ovarian cancer is the sixth most common cancer in females in the UK, with about 7,400 new cases every year.	Edition 1; Ireland Aspirin cuts risk of Big C. WOMEN who take a low dose of aspirin regularly have a lower risk of getting ovarian cancer when compared with women who don't take the painkiller, a new study shows. The US research also found long-term heavy use of non-aspirin nonsteroidal anti-inflammatory drugs - such as ibuprofen and high-dose aspirin - may be linked with an increased risk of ovarian cancer. The research was published in JAMA Oncology. Around 272 women in Ireland die every year from ovarian cancer. Mollie Barnard, who led the study at Harvard University in Boston, said: "Our findings emphasise that research on aspirin use and cancer risk must consider aspirin dose." GRAPHIC: Low dose .. aspirin.	0.1824

Newspaper Favorites? A Comparative Assessment of Political Parallelism Across Two Decades

Erik de Vries[a]* and Gunnar Thesen[a]

[a] *Institute for Media and Social Sciences, University of Stavanger, Stavanger, Norway*

Email: erik.devries@uis.no

The MaML datasets:
Political actors, topics and tone in 15 newspapers from Norway, Denmark, Belgium, the Netherlands and United Kingdom

Gunnar Thesen
University of Stavanger, Norway
gunnar.thesen@uis.no

Erik De Vries
University of Stavanger, Norway
erik.devries@uis.no

August 2022

For using any of the described data, in parts or in full, or for quoting this note, please cite the resource as:

Thesen, Gunnar and Erik De Vries (2022) 'The MaML datasets: Political actors, topics and tone in 15 newspapers from Norway, Denmark, Belgium, the Netherlands and United Kingdom', Harvard Dataverse, https://doi.org/....

Content

1. Introduction

This note documents the data collection for a comparative project on news and party support entitled "Media as the missing link" (MaML). The data comprises a corpus of newspaper articles from Norway, Denmark, Netherlands, Belgium and UK, covering the period from 2000 until 2018/19. Sources were selected based on de Vreese et al (2017): Comparing Political Journalism. The corpus covers the full content of three newspapers from each country: 1 mass-market, 1 left-leaning and 1 right-leaning upmarket.

Section 2 elaborates on the choice of news sources and the processes involved in the preparation of the corpus. Section 3 to 5 subsequently explain the coding of the three sets of variables that has been added to the corpus through a combination of automated and manual coding processes:

- 3, the presence of political actors in the news
- 4, the topic content in the news
- 5, the sentiment in the news

Data files and additional resources are available in our dataverse, see link. Throughout the note, we refer to these resources in the relevant sections. Additionally, section 6 provides an overview of the files in the dataverse. Five country datasets are available for download, together with a brief documentation note explaining the datasets and the variables.

2. Selection of news sources and corpus construction

The selection of newspapers (see Table 2.1) was based on a recent comparative analysis of political journalism (De Vreese, Esser and Hopmann 2017) which included the leading left-leaning broadsheet, the leading right-leaning broadsheet and one mass-market newspaper in each country. However, we have made one change to the sources applied in De Vreese, Esser and Hopmann (2017). The left-leaning broadsheet Dagsavisen, which was part of the Norwegian sample in their study, was not an option due to low data quality and accessibility. One the one hand, this arguably less of a problem in Norway compared to the other countries, since there are no strong candidates for a left-leaning broadsheet with a sizeable and national circulation in Norway. Dagsavisen for instance, has a limited circulation which amounts to less than 10% of the leading right-leaning broadsheet Aftenposten. Furthermore, nearly all of its readers are located in the capital region (Høst 2019). In the end we chose to include the tabloid Dagbladet as a replacement, meaning that we capture a substantially larger share of the national news market. Dagbladet is clearly not a left-leaning newspaper in the same way as Dagsavisen, although it has occasionally issued editorial warnings against incumbent conservative coalitions in the period we cover (e.g., Pettersen 2009).

Table 2.1. News sources by country, political leaning and format.

Country	Left of center	Right of center	Mass-market	Period
Denmark	Politiken	Jyllandsposten	Ekstra Bladet	2000 to 2019
Belgium	De Morgen	De Standaard	Het Laatste Nieuws	2000 to 2019
Netherlands	De Volkskrant	NRC Handelsblad	De Telegraaf	2000 to 2018
UK	The Guardian	Daily Telegraph	The Sun	2000 to 2018
Norway	*Dagbladet**	Aftenposten	VG	2000 to June 2019

For each of these 15 sources, the corpus contains all news published in the period 2000 until 2018/19, summing to a total of over 7 million news articles. Note that this is the final count of articles *after* pre-processing the corpus. The pre-processing contains several important steps. First, items with a word count below 30 have been removed from the corpus, because they nearly always refer to longer articles on other pages in the newspaper. The number of items removed through this procedure was approximately 641.000.

Second, the original newspaper data contained a substantial number of duplicate articles (within the same day). Duplicates were not randomly distributed across time and sources. Therefore, to increase the validity of comparisons, duplicate entries have been removed based on the cosine similarity of article pairs published on the same day by the same newspaper. When article pairs have a similarity of 0.85 or above, one of the articles is (at random) removed from the dataset. These similarity scores have been based on the first 300 words of each article. This cutoff was chosen to reduce computational complexity, and to make comparisons between short and long articles more equal. For instance, we encountered cases where one article contained more words than the other, but where the articles were still clearly duplicates of each other. Had we used the full article text of both documents, their cosine similarity might not have been high enough to detect them as duplicates. The number of articles removed through this procedure was 964.311.

Next, the remaining articles went through Natural Language Processing (NLP) using the R package UDPipe (Straka & Straková, 2017) and version 2.3 of the Danish DDT, Dutch Alpino, Norwegian Bokmål and English EWT Universal Dependencies Models (Nivre et al., 2018). The goal of this procedure is to remove the inherent complexity of natural language by reducing all words in an article to their dictionary lemma. In addition, the process of NLP produces Universal Part-Of-Speech (UPOS) tags for each lemma, indicating the function of the word in a sentence. In this way, words that are written the same way but carry different meanings can be disambiguated.[1]

Finally, before analyzing and coding the corpus, we also removed news stories that deal mainly with sports and entertainment. To classify these ("irrelevant") articles, around 12,000 news articles have been hand-coded in English, and between 6,000 and 7,000 in Danish, Dutch and Norwegian. The reason for the difference between English and the other languages is because similar classification performance for all countries needs to be obtained, and this required more data in English than the other languages. Research assistants have classified these articles based on the categories "Culture/art events and entertainment", "Sporting events and athletes" and "Miscellaneous". If articles fall into any of these three categories, they are considered irrelevant, if not, they are relevant. The miscellaneous category contains all articles that cannot be classified in any of the other categories in the adapted Comparative Policy Agendas codebook (see /Topics/Codebook for manual topic coding of MaML news.pdf). Prior to the coding, the RAs went through several rounds of training. The intercoder reliability scores for this coding reached a level of 0.81 to 0.86, see table 2.2 below.

*Table 2.2. Intercoder reliability human coding of irrelevant articles**

Country	Final round	Min	Max	No. of coders	No. of training rounds
Belgium	0.84	0.77	0.87	7	5
Denmark	0.86	0.81	0.88	4	3
Netherlands	0.81	0.71	0.89	7	3
UK	0.86	0.84	0.89	6	3

* Intercoder reliability measured with Krippendorff's alpha. N=200 for each round of coding. Alpha in table refers to all RAs and one project member. Min refers to lowest score among all coders in final round, max refers to highest score. For Norway, irrelevant articles were not filtered out in advance but instead during the topic coding process.

The hand-coded articles are then used as input for a Naive Bayes classifier. The input features for this model are the tf-idf weighted lemmas and UPOS tags generated in the NLP procedure described in the paper. The "format" of each word/feature in an article becomes lemma_UPOS. For getting the best-performing model for each country, a 3 by 5 nested cross-validation procedure is used, with the 3 outer folds being used for performance estimation of the final model, and the 5 inner folds of each outer fold being used for parameter optimization. In this case, parameter optimization consists of only a single parameter, for feature selection. Features are selected based on the chi2 measure to determine

[1] Eg. book as a noun and book as a verb.

which features are most and least strongly associated with the “irrelevant” topic. Using the absolute chi2 values, the top x-th percentile of features are kept to construct a model. Through the nested cross-validation procedure described above, the optimum cutoff values for feature selection are determined as follows: 0.99 (BE), 0.995 (DK), 0.996 (NL), 0.994 (NO), 0.994 (UK). Using these parameters, the final models achieve a precision of between 0.87 (DK) and 0.94 (UK). Precision is used as optimization measure to avoid as much as possible that relevant articles are classified as irrelevant, allowing for some relevant articles to remain in the relevant articles category. More details on the process of removing irrelevant article can be found in De Vries (2022).

3. Coding the news appearances of political actors

The news appearances of political actors are captured by running queries in the corpus. Political actors are limited to political parties and individual politicians serving as either MPs, ministers, or party leaders.

Mentions of political parties are collected using case-sensitive queries on either the full party name, or the most commonly used party abbreviations. When necessary, special characters like opening and closing brackets for the abbreviations (con) and (lab) in the UK, are also taken into account. In Norway and Denmark, several of the major political parties have single letter abbreviations. In these specific cases, regular expression filters are used to filter out common mistakes, like V (the abbreviation for the left-wing party Venstre) as a roman number 5 in the names of monarchs.

Queries for individual politicians (ministers, party leaders and MPs), are constructed by looking for the combination of the (first) given name and surname within 5 words of each other. A larger distance between the two would result in too many false positives, and a smaller distance in too many false negatives. The queries are also limited to articles published during the time the politician was in office. For ministers the queries include their formal title as an alternative for their given name (e.g. both Secretary Johnson and Boris Johnson are valid hits).

A list of the parties and politicians that were queried is available in our dataverse (/Actors/maml_actors.xlsx). This list also contains the regular expression filters used for dealing with common mistakes.

4. Coding the topic content in the news

The content of each article in the corpus has been coded according to an abbreviated and adjusted version of the classification scheme applied by the Comparative Policy Agendas project (www.comparativeagendas.net). The codebook and coding instructions are available in our dataverse, see /Topics/Codebook for manual topic coding of MaML news.pdf. The procedure involves handcoding a random sample for each language containing between 35 to 45 000 news articles. These samples are subsequently used to train an automated classifier. Before handcoding the sample, RAs were trained in several rounds. The level of intercoder reliability in the final test was between 0.71 and 0.81, see table 4.1 below.

*Table 4.1. Intercoder reliability human coding of article topic**

Country	Final round	Min	Max	No. of coders	No. of training rounds
Norway	0.81	0.79	0.86	3	4
Belgium	0.71	0.68	0.74	4	6
Denmark	0.79	0.76	0.81	4	7
Netherlands	0.72	0.66	0.77	7	9
UK	0.81	0.81	0.82	3	6

* Intercoder reliability measured with Krippendorff's alpha on the major topic level. N=200 for each round of coding. Alpha in table refers to all RAs and one project member. Min refers to lowest score among all coders in final round, max refers to highest score.

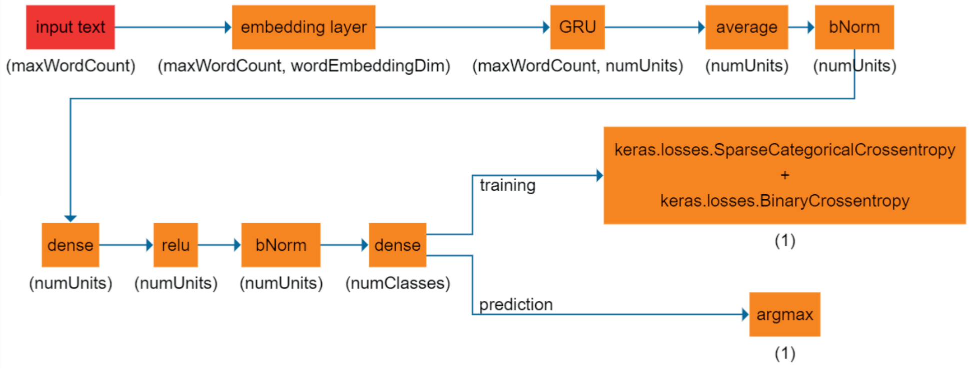
Prior to training the automated classifier (described below), a few changes were made to the handcoded data in order to reach a satisfying level of performance. It is important to be aware of these adjustments when using the MaML-data in combination with other CAP-coded data:

- All subtopic labels were recoded to their major topic label, eg. 101 to 1, 1302 to 13 etc.
- Major topic 18 (Foreign trade) and 21 (Public land, spatial planning and resource management) occurred with a very low frequency in the corpus, causing problems with their classification. As a consequence, these two categories were merged with the respective major topics which they are most similar to. Hence 18 was relabelled 19, while 21 was relabelled 7. This means that the automated classification of topic 19 indicates articles on foreign trade as well as foreign affairs, development aid and international economy. While the automated classification of topic 7 indicates public land etc in addition to environment.

Finally, note that topics 91, 92 and 93 of the codebook are dealt with through the classification of ‘irrelevant’ articles (see section 2). These categories are therefore not applied in the automated topic coding.

To classify the topic of an article, and articles that are exclusively about foreign news topics (domestic vs non-domestic), we use the deep learning framework Keras (Srinivasa-Desikan 2018). The procedure follows the data flow illustrated in Figure 4.1 below.

Figure 4.1. The data flow of the topic classification procedure.



The input text vocabulary is limited to the 10.000 words most commonly found in the train data, and the input text is limited to the first 800 words (maxWordCount). From this input, a word embedding model is generated, using 512 dimensions (tuned). The word embedded representation of the articles is then used as input for a recurrent neural network as described in Cho et al. (2014). After passing through some intermediate neural network layers, the dense layer produces a vector of predicted probabilities for each document, for both the topic and foreign news variables. During training, these are parsed through a sparse categorical crossentropy loss function for the topic labeling, and through a binary crossentropy loss function for the foreign news labeling. Both are added together, and the foreign news loss is multiplied by a factor of 7 (tuned). For labeling, the label with the highest probability is selected.

Optimization of this model has been conducted using 90% of the manually labeled data as training data, and 10% as testing/validation data. Once the optimum model parameters have been determined, these are used in a 10-fold cross-validation procedure to determine the final performance of the model. While it is true that this 10-fold CV includes the 10% test data used to determine the optimal parameters, the stability in performance between the individual folds leads us to the conclusion that the performance estimates provided by the 10-fold CV procedure are reliable.

Using this model, the resulting topic classifications achieve a weighted F1 (the harmonic mean of precision and recall) of between 0.64 (NL/BE) and 0.69 (UK). Given that the model is predicting 22 topic classes, and the considerable complexity of news articles that usually touch upon several topics, this is an acceptable performance. Table 4.2 displays the results per topic category and country. Wojcieszak et al (2021) report exceptionally high performance (accuracy up 0.78) when applying the CAP codebook to tweets, which are shorter than news articles and easier to classify. We are not aware of machine learning classifiers performing substantially better than ours when categorizing traditional news articles based on a multiclass scheme with a high number of topics that are also highly unbalanced (see Burscher et al. 2015). Sebők, M., & Kacsuk recently succeeded in assigning CAP-labels with a precision of over 80 % for most topics, but this approach leaves over 40 % of their corpus unlabeled.

Table 4.2. Classifier performance (F1), across languages and topics.

Code	Label	BEL/NLD	DNK	NOR	GBR
1	Macroeconomics	0.52	0.49	0.51	0.57
2	Civil rights and liberties	0.32	0.40	0.43	0.46
3	Health	0.75	0.76	0.76	0.80
4	Agriculture, fisheries and food	0.55	0.64	0.64	0.51
5	Labour	0.58	0.54	0.59	0.50
6	Education	0.76	0.76	0.76	0.81
7	Environment	0.53	0.61	0.54	0.62
8	Energy	0.59	0.65	0.66	0.70
9	*Refugees and immigration*	*0.65*	*0.63*	*0.66*	*0.66*
10	Transport	0.70	0.73	0.72	0.67
12	Crime and justice	0.77	0.75	0.82	0.83
13	Social welfare and social affairs	0.47	0.55	0.55	0.52
14	Housing and urban/rural development	0.50	0.62	0.66	0.71
15	Commerce, banking and consumer issues	0.53	0.70	0.62	0.68
16	Defense and security	0.74	0.71	0.74	0.76
17	Research, technology, IT and mass media	0.56	0.57	0.63	0.67
19	Foreign affairs, trade, international economy	0.62	0.55	0.55	0.55
20	Public sector and politics in general	0.74	0.64	0.69	0.70
23	Culture, art	0.44	0.44	0.69	0.52
24	Sports	0.60	0.71	0.74	0.67
25	Natural disasters, fires, preparedness	0.61	0.53	0.63	0.63
26	Religion and churches	0.58	0.61	0.71	0.56
	Average performance (weighted F1)	*0.64*	*0.64*	*0.67*	*0.69*

5. Coding the sentiment in the news

To measure the tone or sentiment in the news we rely on a method involving word embedding models similar to recent applications in political communication and political science (Rheault and Cochrane 2020; Rudkowsky et al. 2018). The method was proposed by Rheault et al. (2016) and further developed by Erik de Vries for this project. A detailed account of the approach can be found in De Vries (2022). Note that code and replication material can be found at GitHub - vriezer/sentiment: Replication materials

The process starts with a small and context-independent "seed dictionary" containing one hundred unambiguous negative and positive words. The seed dictionaries for the different languages are available in the dataverse. This is used to build an extensive sentiment dictionary adapted to the news corpus. We do this with the help of machine-learning models that extract meaning from a text by estimating a multi-dimensional vector space in which each word is positioned. Essentially, this approach finds the words that appear in proximity to our words of interest from the seed dictionary, based on the assumption that neighboring words will share an association to the latent semantic meaning of a piece of text (Mikolov et al. 2013).

The result is a longer dictionary (also available in the dataverse) with sentiment values for each word: higher positive values indicate closer proximity to the positive seed words, and higher negative values indicate closer proximity to the negative seed words. Based on the expanded dictionary, we then calculate how positive or negative each sentence and article in our corpus is.

The procedure is validated by having trained research assistants code a random sample of sentences. The instructions for this coding can be found in our dataverse, see /Tone/Codebook for tone coding.pdf. The intercoder reliability ranged from 0.71 (UK) through 0.75 (Denmark) and 0.79 (Norway) to 0.84 (Dutch). Comparing the human-coded sentiment to the sentiment based on the word embedding dictionary, our F1-scores range from 0.61 to 0.64.

De Vries (2022) provides more details on the validation, in addition to examples of applications that indicate good predictive validity. For instance, the well-established negativity bias of political news can be reproduced with our data. Furthermore, in each country, the tabloid is more negative than the broadsheets. Summarizing, the performance is on a par with the best performing non-manual sentiment model (based on non-human coding) tested in a recent study by Van Atteveldt, van der Velden, and Boukes (2021, see overview of results page 128).

6. Overview of files in dataverse

Table 6.1. File location, names and content.

Location	Name	Content
Root folder	MaML_documentation_and_release_note.pdf	The present document
/Actors/	maml_actors.xlsx	List of all parties and individual politicians queried in the corpus
	ActorSearch.R	Code for running actor queries.
/Topics/	Codebook for manual topic coding of MaML news.pdf	The adjusted CAP classification scheme, including coding
	Intercoder_Reliability_Irrelevant_Articles.xlsx	Intercoder reliability scores for human coding of irrelevant articles
	Intercoder_Reliability_Major_topics.xlsx	Intercoder reliability scores for human coding of major topics
	About the topic classifier.pdf	Short description of the deep learning framework applied by Markus Fjellheim for the topic classification
	performance_topic_classifier.csv	Performance scores for the topic classifier
/Tone/	Codebook for tone coding.pdf	Coding instructions for human coding of tone in the news
	Intercoder_Reliability_Tone.xlsx	Intercoder reliability scores for human coding of tone
	validation_results_sentiment_*country*.Rds	Four rds-files containing the results from validating the automated sentiment coding against human coded sentiment.
/Tone/ Dictionaries/	seed_dict_*country*.csv	Four seed dictionaries (benl=Belgium and the Netherlands, no=Norway, dk=Denmark, uk=United Kingdom) used as a starting point to create full sentiment dictionaries.
	full-lexicon-*country*.txt	Four full sentiment dictionaries created through the word embedding models
/Data/	articleXactor_level_data_*country *.dta	Five country datasets in Stata-format, containing information on the article X actor level
	Documentation for article level MaML datasets.pdf	Documentation of the article X actor level datasets and variables

7. Funding and acknowledgements

The MaML-project was funded by the University Fund for Rogaland, Stavanger, Norway.

We are indebted to the rest of our project group who has coordinated data collection efforts outside Norway: Christoffer Green-Pedersen and Peter Bjerre Mortensen in Denmark; Annelien Van Remoortere, Rens Vliegenthart and Stefaan Walgrave in Belgium and the Netherlands; Will Jennings in the UK.

A special thanks to data analyst Markus Fjellheim who is responsible for the topic classifier described in section 4.

Finally, we are grateful for the excellent work of our 25+ research assistants who have coded news across the five countries.

8. References

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 Science, 659(1), 122-131.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

De Vreese, C., Esser, F., & Hopmann, D. N. (eds). (2017). Comparing political journalism. London: Routledge.

De Vries, E. (2022). The Sentiment is in the Details: A Language-Agnostic Approach to Sentence-Level Sentiment Analysis in News Media. *Computational Communication Research*. doi:10.31235/osf.io/8y3jq.

Høst, S. (2019). Papiraviser og betalte nettaviser 2018. Statistikk og kommentarer. Report no. 90/2019, Volda University College.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28, 11–21.

Pettersen, Ø. B. (2009). Pressen og partiene - Partitilhørighet i 2005. En analyse av Aftenposten, Dagsavisen, Dagbladets og Dagens Næringslivs leder- og kommentarartikler under valgkampen i 2005. Master Thesis, Department of Media and Communication, University of Oslo.

Sebők, M., & Kacsuk, Z. (2021). The multiclass classification of newspaper articles with machine learning: The hybrid binary snowball approach*. Political Analysis*, 29(2), 236-249.

Straka, M., & Straková, J. (2017). Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with UDPipe. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*.

Srinivasa-Desikan, B. (2018). Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras. Packt Publishing Ltd.

Rheault, L., & Cochrane, C. (2020). Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora. *Political Analysis* 28 (1): 112–33.

Rheault, L., Beelen, K., Cochrane, C., & Hirst, G. (2016). Measuring Emotion in Parliamentary Debates with Automated Textual Analysis. *PloS ONE* 11 (12): e0168843.

Rudkowsky, E., Haselmayer, H., Wastian, M., Jenny, M., Emrich, S., & Sedlmair, M. (2018). More than Bags of Words: Sentiment Analysis with Word Embeddings. *Communication Methods and Measures* 12 (2-3): 140–57.

Van Atteveldt, W., van der Velden, M., & Boukes, M. (2021). The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms. *Communication Methods and Measures* 15 (2): 121–40.

Wojcieszak, M., Casas, A., Yu, X., Nagler, J., & Tucker, J. A. (2021). Echo chambers revisited: The (overwhelming) sharing of in-group politicians, pundits and media on Twitter.

Zeman, D., Nivre, J., & Abrams, M. (2019). "Universal Dependencies 2.5. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (UFAL)." Faculty of Mathematics and Physics. Charles University.