Capturing a News Frame

Comparing Machine-Learning Approaches to Frame Analysis with Different Degrees of Supervision

Eisele, O.; Heidenreich, T.; Litvyak, O.; Boomgaarden, H.G.

DOI
10.1080/19312458.2023.2230560

Publication date
2023

Document Version
Final published version

Published in
Communication Methods and Measures

COMMUNICATION METHODS AND MEASURES
2023, VOL. 17, NO. 3, 205–226
https://doi.org/10.1080/19312458.2023.2230560

Capturing a News Frame – Comparing Machine-Learning Approaches to Frame Analysis with Different Degrees of Supervision

Olga Eisele [a,b], Tobias Heidenreich [a,c], Olga Litvyak [a], and Hajo G. Boomgaarden [a]

[a]Department of Communication, University of Vienna, Vienna, Austria; [b]Amsterdam School of Communication Research (ASCoR), University of Amsterdam, Amsterdam, the Netherlands; [c]The Social Science Centre Berlin (WZB), Berlin, Germany

ABSTRACT

The empirical identification of frames drawing on automated text analysis has been discussed intensely with regard to the validity of measurements. Adding to an evolving discussion on automated frame identification, we systematically contrast different machine-learning approaches with a manually coded gold standard to shed light on the implications of using one or the other: (1) topic modeling, (2) keyword-assisted topic modeling (*keyATM*), and (3) supervised machine learning as three popular and/or promising approaches. Manual coding is based on the Policy Frames codebook, providing an established base that allows future research to dovetail our contribution. Analysing a large dataset of 12 Austrian newspapers' EU coverage over 11 years (2009–2019), we contribute to addressing the methodological challenges that have emerged for social scientists interested in employing automated tools for frame analysis. While results confirm the superiority of supervised machine-learning, the semi-supervised approach (*keyATM*) seems unfit for frame analysis, whereas the topic model covers the middle ground. Results are extensively discussed regarding their implications for the validity of approaches.

Introduction

Framing is among the most prominent conceptual approaches in the social sciences to classify communication contents. However, the academic debate about conceptualizing or operationalizing a frame has always been fractured (Entman, 1993) and has produced a plethora of literature discussing different definitions but also types of frames (e.g., Cacciatore et al., 2016; De Vreese, 2005). In addition to conceptual debates, there is a long-standing divergence in the field regarding the appropriate empirical identification of frames. The difficulty of reliably and validly identifying frames lies in the elusiveness and abstractness of the concept (Matthes & Kohring, 2009, p. 258) and the fact that frames "are embedded in latent dimensions of the communication, and they are generated because of spurious correlations between word (co-)occurrences in communications" (Hellsten et al., 2010, p. 593). Overall, the field appears to have come to accept the variety of frame identification approaches, provided they adhere to quality standards of manual quantitative or qualitative content analysis.

CONTACT Olga Eisele o.eisele@uva.nl Amsterdam School of Communication Research, University of Amsterdam, Nieuwe Achtergracht 166, Amsterdam 1018WV, the Netherlands

This article has been corrected with minor changes. These changes do not impact the academic content of the article.
Supplemental data for this article can be accessed online at https://doi.org/10.1080/19312458.2023.2230560.

The increasing availability of large amounts of digitized text and the growing popularity of computer-assisted content analysis methods in the social sciences have added new dimensions to the challenges of frame analysis (e.g., Günther & Quandt, 2016; Jünger et al., 2022). Researchers now employ a variety of computational tools to capture frames in text corpora, ranging from topic models (e.g., Heidenreich et al., 2019), to dictionaries (e.g., Nassar, 2020), semantic networks (e.g., Calabrese et al., 2019) or supervised machine-learning (e.g., Burscher et al., 2014). Yet, regardless of manual or automated, "[c]ontent analysis is a social scientific methodology that requires researchers who use it to make a strong case for the validity and reliability of their data" (Potter & Levine-Donnerstein, 1999, p. 258). While in manual frame analysis the issue of reliability is probably as important as the issue of validity (e.g. Semetko & Valkenburg, 2000; Van Gorp, 2005, 2007), the shift to computational methods and the fact that computers are reliable per definition, has moved the focus to validity, that is "the extent to which a measuring procedure represents the intended, and only the intended, concept" (Neuendorf, 2002, p. 112). For being able to judge the actual validity of the computer's output, it needs to be compared against and ideally converge on some externally defined "gold standard" or "ground truth," i.e. some form of objective, or at least inter-subjectively valid reference measurement (Grimmer & Stewart, 2013; Song et al., 2020).

Recent literature has started to address the methodological challenges of automated frame analysis tools and the concern that technology has been prioritized at the cost of validity in social science research (see Baden et al., 2021) when it comes to measuring frames. Yet, contributions often discuss "computational methods" as such and highlight general difficulties in automated frame coding exemplifying one popular concept in the social sciences (e.g., Baden et al., 2021; Günther & Quandt, 2016; Grundmann, 2021; also; Jünger et al., 2022). Nicholls and Culpepper (2020) and Kroon et al. (2022) engage in a systematic comparison of approaches used for coding frames automatically. Yet while their studies provide valuable insights, they focus on either unsupervised (Nicholls & Culpepper, 2020) or supervised (Kroon et al., 2022) approaches and do not compare both.

We aim to contribute to this evolving debate by providing a comparison of selected approaches to automated frame analysis that are popular in the field. These approaches have in common that they rely on inductive pattern extraction techniques, with such patterns representing frames. They differ, however, concerning at what stage in the process deductive control is exercised in terms of frame classification. Our contribution hence is a comparison of different automated pattern identification techniques – in other words stretching from less supervised to fully supervised methods – against manual coding, providing a thorough reflection on the implications of using one automated approach over the other. We hence add a comparison of approaches to frame identification that specifically takes into account the role that deductive control at different stages plays for computer-assisted classification, including how inductive procedures perform to find deductive categories in comparison to deductive procedures. For that purpose, we use the Policy Frames codebook developed by Boydstun et al. (2013) which represents one of the first attempts to devise a comparable deductive framework that can be used by any scholar interested in (automated or manual) frames analysis, rather than developing an own framework as is very often done in similar research (Borah, 2011, also Brugman & Burgers, 2018). Being aware of the ongoing conceptual discussions regarding framing, we have chosen this particular notion of framing since it is generic, and lends itself to comparative research, making it well suited for our research purposes.

An all-encompassing comparison of the fast-increasing range of text analysis approaches is a task arguably too comprehensive to address in a single paper (e.g., Boumans & Trilling, 2016; Van Atteveldt & Peng, 2018). Given their growing popularity, we limit ourselves to machine-learning approaches with lower and higher degrees of supervision which have either been popular for automated frame identification in the past (topic models and supervised machine learning) or seem promising in this respect (keyword-assisted topic model). We thereby disregard popular dictionary approaches mainly for pragmatic reasons since the development of dictionaries is often extremely resource-intense (see, e.g., Lind et al. 2020; but see Kroon et al., 2022 for a recent evaluation of the performance of dictionary approaches for frame analysis). We apply the selected approaches to 11 years (2009–2019) of EU print news ($N = 547{,}617$) in Austria. We thus look into political news which, due to the politicization of EU affairs in the country, appear promising for an exemplary comparison of different frames. Our study contributes to the growing literature on the validity of computational social science approaches more generally, and in particular, informs framing researchers about the potential pitfalls and opportunities of a range of machine learning approaches for frame identification.

A fractured paradigm

An ever-growing number of studies utilizing the concept of framing adopt diverse definitions of frames and varying methodologies to identify them. The variation in approaches to frame identification is arguably mirroring the conceptual vagueness of what framing actually is (for overviews of the origins and development of framing see Brugman & Burgers, 2018; Cacciatore et al., 2016; Vliegenthart & Van Zoonen, 2011). Originating in psychology (Bateson, 1955), the framing concept has found extensive application across disciplines. Spreading first to sociology (Goffman, 1974), in particular social movement studies (Snow & Benford, 2005; Snow et al., 1986), and then to communication science and journalism (Gamson & Modigliani, 1989, Gitlin, 1980), with research in this field booming after the publication of Entman's seminal article about "framing as a fractured paradigm" (Entman, 1993), the framing concept has found its application in political science (Baumgartner et al., 2008; Bjarnøe, 2016; Hänggli, 2020). Against this background, research has repeatedly emphasized the lack of a unified approach to framing or even contradictions regarding its main aspects (Cacciatore et al., 2016; Hertog & McLeod, 2001; Matthes, 2009), agreeing that "a more systematic and conceptually precise measurement of framing is warranted" (Vliegenthart, 2012, p. 937).

Existing definitions of frames evolved, narrowing down the initially relatively abstract understanding of frames as frameworks or discourses to restrictive definitions that outline the main features of frames. Goffman (1974) understood frames as "frameworks or schemata of interpretation" in individuals' minds. The idea of "interpretative packages" was adopted by research (Entman, 1993; Gamson & Modigliani, 1987) focusing on "frames in communication". Scholarship in this tradition has been strongly stimulated by Entman's definition of framing (Entman, 1993, p. 52) stating that "to frame is to select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described". This definition highlights the multi-dimensionality of political issues and has largely been used by the scholarship that focuses on the diversity of issue definitions (e.g., Gilardi et al., 2021; Nicholls & Culpepper, 2020). As argued above, while being aware of the conceptual debate we focus on one particular, yet broadly relevant notion of framing that allows for an exploration of computational approaches to framing.

Increasing comparability: a deductive approach to framing

When it comes to empirically identifying frames, the literature proposes two approaches labeled as inductive and deductive (De Vreese, 2005; Matthes, 2009; Semetko & Valkenburg, 2000) that in turn allow identifying generic frames or issue-specific frames. Inductive frames emerge from the data, allowing the researcher to explore the whole range of frames in a more nuanced way than a prior definition would yield (Van Gorp, 2007). These frames are usually only related to the specific issue or context being analyzed. The comparability of frames across different issues, therefore, remains very limited. Contrarily, generic frames may emerge across different issues, time, or cultural contexts (, De Vreese, 2005), such as provided by Semetko and Valkenburg (2000) who discuss five frames commonly appearing across issues – conflict, human interest, economic consequence, morality, and responsibility attribution.

In a similar vein, the more recent Policy Frames Codebook (Boydstun et al., 2013) follows the deductive approach, proposing a general coding scheme that has been used for a comparative analysis of news frames (Card et al., 2015; Field et al., 2018; Kwak et al., 2020). This research adopts Entman's definition and focuses on identifying frames as a "general aspect of linguistic communication about facts and opinions on any issue" (Card et al., 2015, p. 438). Boydstun et al. (2013, p. 4) identify a major shortcoming of the issue-specific coding of frames in the inability to compare across different issues. Relying on examples from empirical studies the authors conclude that similar frames are likely to co-occur together, the appearance of one frame can trigger the appearance of another. Therefore, Boydstun et al. (2013, p. 4) argue in support of their approach that it makes sense to cluster such linked frames into larger framing dimensions cross-cutting different policy issues to allow for comparison. The frames proposed within this codebook "were informed by the framing literature and developed to be general enough to be applied to any policy issue" (Card et al., 2015, p. 439). However, this general coding scheme "can also be specialized in issue-specific ways" (Boydstun et al., 2013, p. 4) and includes an "other" category, allowing to include additional frames, should the need emerge. Though initially developed to analyze frames in English, this approach has been also used for the automatic annotation of frames in Russian (Field et al., 2018). Overall, the Boydstun approach relies on the most popular definition of frames (Entman, 1993) allowing for easy connection to the debate in the field (Baumgartner et al., 2008; Bjarnøe, 2016; Hänggli, 2020). In addition, it allows us to build on a resource that can be used by future research with similar interests. Thus, the comparison of different automated approaches to frame analysis could be continued and would immediately dovetail our efforts if the same codebook was used. We, therefore, regard the approach as a useful basis for any efforts to systematize knowledge about the functioning and validity of different automated approaches to frame analysis.

The (contested) validity of automated frames analysis

Regarding computational approaches to frame analysis, a wide variety of approaches has been applied. In the following, we give a brief overview of the most important ones to allow a better contextualization of the approaches we have selected for our purposes. Existing studies have relied on unsupervised (e.g., Brouwer et al., 2017; Burscher et al., 2016, Egger & Yu, 2022; Lawlor, 2015; Wallace, 2018) and supervised machine-learning techniques (e.g., Burscher et al., 2014), dictionary approaches (e.g., Lawlor & Tolley, 2017; Schultz et al., 2012; Shah et al., 2002), or network analyses (e.g., Calabrese et al., 2019; Hellsten et al., 2010; Jiang et al., 2022; R. A. Lind & Salo, 2002). Studies focusing on the detection of media frames recently turned to text analysis tools developed by computational linguists, such as Bidirectional Encoder Representations from Transformers (BERT) for unsupervised classification of media frames (e.g., Akyürek et al., 2020; Kwak et al., 2020; Liu et al., 2019), also in combination with other approaches (e.g., Guo et al., 2022), Linguistic Inquiry and Word Count (LIWC) for a dictionary-based approach (e.g., Nassar, 2020), Corpus Linguistics (e.g., Dayrell, 2019), and word embeddings (e.g., Kroon et al., 2022). Especially topic modeling has become popular for coding frames (e.g., Gilardi et al., 2021, Heidenreich et al., 2019; Poirier et al., 2020; Ylä-Anttila et al., 2018; Zhang & Trifiro, 2022).

Although automated frame analysis provides advantages such as the opportunity to analyze large text corpora or an increased reliability of frame coding, recent studies have called for a critical assessment of these methods, pointing out their potential limitations (Baden et al., 2021; Grundmann, 2021; Nicholls & Culpepper, 2020). Especially more unsupervised methods, such as topic modeling, have been criticized for not being able to capture the complexity of a frame (e.g., Nicholls & Culpepper, 2020). As at the pre-processing stage topic modeling often removes morphological characteristics of words, stripping them to the lemmata, it is argued that this approach shows linguistic insensitivity (Brookes & McEnery, 2019) and disregards the sequential structure of human language that is valid for all languages (Nerlich et al., 2012) and is crucial for frame analysis (e.g., Jiang et al., 2022). However, while the general effects of preprocessing in automated content analysis have received some attention in the scholarly debate (see Denny & Spirling, 2018), the effects regarding topic modeling in particular remain empirically unsubstantiated.

As such, topic modeling has been increasingly used for frame analysis, whereby scholars equated topics, i.e., the output of the topic model, and frames. However, the topic model is a tool that responds to specific variations in word co-use and as such cannot distinguish between frames and issues; it is therefore likely that analyses of issues or frames, respectively, operate in muddy analytical waters as the output of the topic model will be a mix of issues and frames. In an attempt to increase validity, topic modeling has therefore been complemented with other types of analysis (e.g., Guo et al., 2022; Hubner, 2021) that allow a deeper understanding of the results, such as, for example, Topic Model Networks (Walter & Ophir, 2019) or semantic network analysis (e.g., Calabrese et al., 2019; Jiang et al., 2022).

Recent studies have also discussed a lack of objectivity in the "bags-of-words" (Murakami et al., 2017, p. 244; Nicholls & Culpepper, 2020; Walter & Ophir, 2019), as defining and labeling the "topics" or frames is in the power of the researcher conducting the analysis. Accordingly, a limitation of less supervised approaches lies in their replicability (Roberts et al., 2016; Wilkerson & Casas, 2017). Dictionary approaches seem more promising in this respect, as they rely on the explicit mention of words in the text. However, approaches that aim at creating off-the-shelf dictionaries capturing latent concepts, e.g., sentiment (Rauh, 2018), or frames (e.g. Nassar, 2020) require additional adjustments to provide valid results with data from a domain different from the one the dictionary was developed for (Boukes et al., 2020; Guo et al., 2016; Kroon et al., 2022). Finally, in supervised machine learning approaches the validity of results hinges strongly on the quality of training data as well as potential biases in the training data sample (e.g. Song et al., 2020). As illustrated by the discussion above, computational approaches bear limitations that likely reflect negatively on the validity of outcomes. This may be even more accentuated when it comes to the measurement of a complex construct such as frames.

Three approaches to automated frame analysis

Overall, the use of automated text analysis techniques for frame analysis is growing in popularity, and a critical reflection regarding the implications of using different automated approaches is a necessary step to develop a toolkit for automating frame analysis. Researchers need to be aware of the consequences of applying one or the other approach to identifying frames, which requires systematic comparisons. Drawing on the considerations just discussed, we here contrast three machine-learning approaches that have been popular in frames analysis over the past years (topic models and supervised machine learning) or seem promising new routes to explore (keyword-assisted topic model). These three approaches have in common that they identify frames through an inductive extraction of patterns in the texts at some point in the processes. Unsupervised machine learning (topic models) would extract such patterns without any prior information about what to look for, whereas supervised machine-learning would extract patterns following predefined classifications. For the identification of frames both types are subordinated under deductive control, but at very different stages of the process. In supervised machine learning deductive control is exercised at the very outset, by providing training materials that have been constructed in a deductive manner and pattern extraction would seek to reproduce the deductive classification. In unsupervised machine learning for frame identification deductive control is imposed after the unguided pattern extraction (see Baden et al., 2020), by mapping those patterns onto deductive categories. The latter, therefore, is likely to be more prone to perform worse in an analysis aiming toward identifying concepts specified beforehand (see also Kroon et al., 2022). An approach that sits in the middle between unsupervised or supervised techniques concerning the role of deductive control is the keyword assisted topic model. Utilizing seed words, some deductive control is provided to a priori focus the pattern extraction onto relevant distinctions. The differential role of deductive control exercised by the researcher does reflect on the labor that researchers are required to put into the process, ranging from laborious training data annotations to post hoc word list interpretations. While all approaches differ in their operating modes, they can all be used to answer similar research questions, making a comparison worthwhile in its own right.

(1) *Topic modeling* is a frequently employed inductive machine-learning method that provides insight into the distribution of topics in the form of word clusters in the text without prior expectations (e.g., Dutceac Segesten & Bossetta, 2019; Gilardi et al., 2021). Although the term topic modeling includes a variety of possible cases of application, Latent Dirichlet Allocation (*LDA*) emerged as a frontrunner (Maier et al., 2018): Being a generative probabilistic model of a corpus, the approach aims at representing documents as constructed by drawing upon words associated probabilistically with certain topics. Topics, in turn, can be seen as a distribution of words. Simply put, this technique identifies patterns by clustering words based on their co-occurrence. Without any prior knowledge about the data and topics or possibly frames involved, *LDA* assumes that each document can be seen as an exchangeable distribution of topics where the probabilities for each topic represent a document (Blei et al., 2003: 996ff). As a "bottom-up" approach, with topics emerging from the corpus, researchers thus do not need to rely on a high degree of a priori supervision. However, to achieve reliable and valid results, specific key aspects, for example, concerning the interpretability or the coherence of topics involving human topic inspection, need to be considered (Baden et al., 2020; Maier et al., 2018).

The applicability of topic modeling, or word clustering techniques more generally, to identify frames[1] in the text remains disputed since they cannot, only based on word clusters, deliver the depth of interpretation that a frame would presuppose (e.g., Matthes & Kohring, 2009, p. 261). Simply equating the output of a topic model with a frame blurs the analytical clarity of the concept. The topic model is a tool that responds to specific variations in word co-use and, as such, cannot distinguish between frames and issues but picks up their traces in the text. Scholars have therefore proposed definitions that arguably allow using this method for automated frames analysis, often based on extensive qualitative post-hoc validation or in combination with other methods (Baden et al., 2020; Jacobi et al., 2016; Sides, 2006; Walter & Ophir, 2019; Ylä-Anttila et al., 2018).

(2) The novel *keyword-assisted topic modeling* also relies on the basics of probabilistic topic models, aiming at representing documents as a mixture of underlying topics. Yet, this method, in addition to the inductive exploration of the corpus, also allows researchers to deductively define topics they expect to find in a given collection of documents. Using predefined sets of keywords indicating the topic to be measured, scholars can define a certain number of topics to be found associated with those keywords. In addition, the model provides the possibility to also account for other – unexpected - topics, by allowing the specification of several topics that are not associated with the given keywords, thus basing the approach on a mixture of two distributions (one for keywords, one for all words). The technique places greater importance on the keywords as markers for certain topics but also allows the model to learn about the importance of given keywords for a specific topic. According to Eshima et al. (2020), this yields better results concerning document classification and is less sensitive to the number of topics compared to standard topic models. Additionally, it prevents problems commonly arising when interpreting topics from traditional *LDA* models post-hoc, simplifying interpretation and enhancing the reliability of results.

The degree of supervision for this approach is still comparably low as the sets of keywords do not need to be excessively large. However, scholars need to hold certain expectations concerning the frames in the text corpus and come up with explicit keywords for the model. Therefore, and compared to the *LDA* approach, more supervision is needed for the *keyATM*, feeding more information into the algorithm. All keywords selected should be indicative of the topics expected and theoretically motivated. However, as the authors emphasize, keywords need to be present in the data to a certain degree (Eshima et al., 2019), making the selection process top-down (theoretical motivation) as well as

[1]Please note that in the following, we will use the term frame for the output of *LDA* and *keyATM*. While we are of course aware that there is considerable criticism regarding the validity of topic models in general for measuring frames, we aim to increase the clarity and consistency of the paper by using the terms frame. In addition, the output is compared with the frames identified in the manual coding; in that sense, the results of our analysis will show in how far the label "frame" is a valid description of the output of *LDA* and *keyATM*.

bottom-up (sufficient share of mentions in the text). Therefore, the information used for determining keywords should not be detached from the data, requiring intimate knowledge of the corpus.

(3) *Supervised machine learning* (SML) approaches, finally, are usually applied to assign documents to predefined categories and have been used to identify frames in text. The umbrella term represents a variety of different approaches, or classifiers, such as support vector machines or Naive Bayes classifiers.[2] *SML* usually involves the comparison of different models (Burscher et al., 2014) to find an algorithm or a set of classifiers performing best for a task at hand. Imagining the data as quantified information, i.e., as a set of vectors containing features, classifiers are trained on using such vectors that are manually labeled beforehand. Often called the "Gold Standard," this human-coded data is the foundation for the approach. It is used for training as well as validation purposes and is thus largely accountable for reliable and valid results. In the process, the manually coded data is split into a training and validation set. The training part (also called held-in sample) "teaches" classifiers to identify whether a text document contains the concept to be measured (e.g., a frame) while the remaining gold standard data (held-out sample) is subsequently used to check the validity of classifications made by the automated approach by predicting and then comparing labels with results from the manual analysis (Boumans & Trilling, 2016). When proven to be useful, classifiers can be used to predict labels of unseen data.

The approach requires intensive human input. Given the elusiveness and the context-sensitivity of frames as such, it is important to have sufficient data to be able to capture this complex and multi-faceted concept. It is, however, unclear what exactly sufficient would mean beforehand. Thus, an extensive collection of data-based manual coding is needed to ensure satisfying performance. This approach, therefore, is the most elaborate and labor-intensive of the three techniques presented. It is also the one which is fed the most information about the data and the desired measurement entering the algorithm. The validity of results crucially depends on the reliability of the manually coded gold standard material (e.g., Song et al., 2020), making this process costly concerning time and financial resources. Once classifiers are trained and validated, however, they apply to different data, while factors interfering with the validity of classifiers such as the text domain of the training and validation data should always be kept in mind. Considering the automated detection of frames, this approach is shown to be useful for certain definitions of framing concepts (e.g., Burscher et al., 2014).

Data & methods

We apply and compare the methodological approaches discussed above to Austrian print news about the EU, thus political news coverage, for a period of 11 years (2009–2019), thus to German language text. The EU in general is found to suffer from a lack of public visibility, discussed as a communication, and eventually a democratic deficit. However, EU coverage and public contestation over EU issues have generally increased (Boomgaarden & De Vreese, 2016) due to several EU crises occurring during that time, e.g., the financial crisis or the migration crisis (e.g., Hobolt & De Vries, 2016). EU news, also EU frames, have been studied to some degree, allowing us to validate results in the mirror of similar research (e.g., Boomgaarden & De Vreese, 2016; Dutceac Segesten & Bossetta, 2019). As is often the case for political news, we expect a large share of news on the economy; regarding the above-mentioned political differences and the legitimacy deficit of the EU, we also expect that articles will discuss the EU's political problems and suggested solutions. For the Austrian case, in particular, these EU crises have caused heated political debates, leading, for example, to the resignation of the chancellor in the course of the migration crisis (Auel & Pollak, 2016). Overall, the general topic EU and the time frame seem promising in terms of breadth and depth in the material to test different dimensions of frames.

[2]We provide short introductions to the classifiers used in this research in the respective section in the "Data & Methods" chapter and the Appendix.

Our dataset consists of the 12 largest daily print newspapers to gain a broad insight into the Austrian EU discourse (e.g., Greussing & Boomgaarden, 2017 for a similar approach). Sources were selected according to their journalistic routines (tabloid vs. broadsheet), and scope of distribution (regional vs. national scope). In total, we retrieved $n = 547{,}617$ articles via the online database of the Austrian Press Agency (APA) (see Appendix Table A1).

For the retrieval of EU-relevant texts from the database, we formulated an extensive search string (see Appendix Section A1). For that purpose, we drew inspiration from other studies but also inductively coded search terms from a sample of 875 news items published on randomly selected days in our period of analysis. The search string mainly includes EU actors and institutions, e.g., the European Commission or the European Parliament, but also specific buzzwords associated with EU policy for example. The final search string was validated using a fresh sample of 1,500 articles. We ensured that the validation sample included data from every outlet included for a random day for each year. Reaching satisfactory intercoder-reliability (*Krippendorff's* $\alpha = 0.86$), this set was manually annotated by three different coders to validate the performance of the search string (*precision* = 0.92, *recall* = 0.96, *F1* = 0.94).

Before applying the different automated approaches to the data, we went through different steps of data cleaning and preprocessing. This included the removal of a set of predefined stopwords, numbers, or other not-needed characters, as well as lowercasing. We filtered all documents unexpectedly published in different languages than German and removed duplicated news items. Furthermore, we filtered, whenever available, articles according to the section in which they were published to get rid of unwanted documents referring to the TV program or letters to the editor (i.e., non-editorial content). More details on the implementation of each approach are discussed in the following.

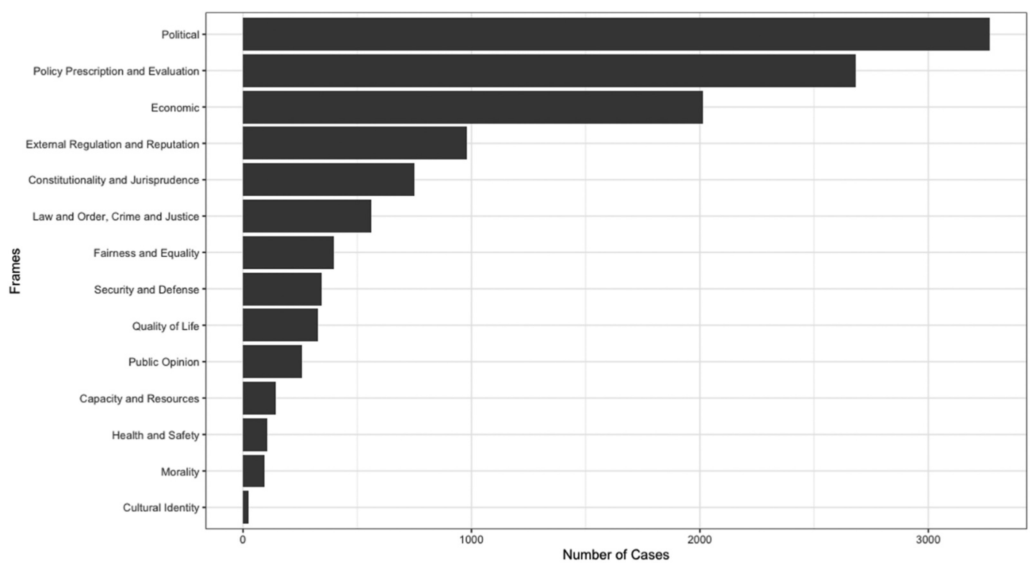

Figure 1. Absolute number of frames manually coded for each category; N = 12481 strings.

The gold standard: manual coding of frames

The Gold Standard, i.e., the training material was coded based on the codebook developed by Boydstun et al. (2013) and includes 14 different frames (see Figure 1; see Table A2 in the Appendix for details). These frames were only coded when they were related to the EU as a frame object in some way, e.g., a frame of the European Commission, an EU policy like external relations, or an EU political event like Brexit. Six trained coders annotated frames from 6×800 articles sampled from the above-described dataset; frames were not coded at the article level but were identified from the article and the relevant text parts were pasted in a frame string variable. This string variable, then, was processed and used as input for our deductive approaches.

After two coding workshops, we tested reliability (I) *before* the start of the coding, (Ia) for the identification of relevant text parts (unitizing reliability) as well as (Ib) for the actual coding. (Ia) Units (frames) needed to be identified by coders from the articles directly. Thus, we needed to make sure that there was reliable agreement as to what text part contained a frame and which does not. Intercoder unitizing reliability was, given the elusiveness of the frame concept (e.g., Matthes & Kohring, 2009) as well as the rather high number of 6 coders, not high ($N = 50$ articles resulting in a set of 110 valid frames; *Percentage Agreement* = 76,1%, *Krippendorff's Alpha* = 0.17, *Brennan & Prediger* = 0.18).[3] The validity of the identification of individual coding units was slightly lower (*percentage agreement* = 71,8%; measured against instructors' coding). Since the automated coding was conducted at the article level (and not at the level of the individual frame), we calculated our final unitizing reliability with respect to how reliably coders identified an article as containing frames overall, which, given the aggregate level, showed higher reliability (*Percentage Agreement* = 86,6%, *Krippendorff's Alpha* = 0.57, *Brennan & Prediger* = 0.57). For our final validity scores, we compared against instructors' identification of relevant articles (i.e., such articles containing frames) which, again, yielded slightly higher scores (*Percentage Agreement* = 76,34%). We trained coders again but given the overall still comparably high score, did not measure reliability/validity for the identification of relevant units or articles anymore. (Ib) In the second step, coders were provided with a set of frames already identified as valid by instructors ($N = 110$) to ensure the comparability of their coding. Based on this, we calculated the final, satisfactory reliability scores of the manual coding which can be found in Appendix Table A3.

Our final intercoder reliability scores revealed some worse-performing frames. We, therefore, decided to monitor reliability also (II) *during* the actual coding and included another set of 50 articles which all coders coded along with their assigned sample. This was to understand if there was a need for a thorough rechecking of the coding in the end. However, all values improved (see Appendix Table A3 for details). Especially frames with high prominence in the coding yielded lower reliability (i.e., *Policy Prescription and Evaluation*, *Political* and *Economic* frames; see Appendix Table A2 for a description of frame categories), suggesting that these frame categories served as "default" categories for some coders in the team.

Topic modelling: latent dirichlet allocation

As mentioned above, topic modeling is sometimes used to identify frames by combining the technique with an intensive post-hoc evaluation to gain better insights into the validity of the topic model output. Similarly, our approach features a manual coding step using the Hybrid Content Analysis (HCA) approach by Baden et al. (2020). Following this method, the output of the *LDA* is annotated using a codebook to ensure that results might be classified toward the targeted frames and, additionally, allow for grouping clusters as well as binary output.

To implement the *LDA* model, we used the R package *topicmodels* (Hornik & Grün, 2011). Guided by the literature (e.g., Jacobi et al., 2016) and based on metrics such as perplexity with varying alpha, we chose *alpha* = $5/k$. The decision on how many k topics the model should calculate was based on multiple considerations. While we aim for measuring at least 14 frames within our data, we anticipated the corpus to entail much more. Therefore, we set up an evaluation process calculating models within

[3]Please note that we use Percentage Agreement here as all variables coded for checking reliability were coded as dummy variables and, given the highly interpretative task, the sample was rather small. As is for example discussed in De Sweert (2012), Krippendorff's alpha is often found to be too conservative in skewed samples with limited variance. However, percentage agreement is often found to be a too liberal measure. The coefficient introduced by Brennan and Prediger is considered a middle ground as it provides a chance-corrected measure that can handle skewed samples as well. We, therefore, decided to provide all three measures to increase the transparency of our analysis (see Egelhofer et al., 2021 for a similar strategy).

a number of k topics ranging from 20 to 300 ($k = \{20, 30, 40, 50, 75, 100, 125, 150, 175, 200, 225, 250, 275, 300\}$). While we can see in the slope of perplexity and loglikelihood (see Figure A1 in the Appendix) that both metrics indicate better-performing models with $k > 75$, we also notice saturation around $k = 100$ where slopes are becoming less steep (perplexity) or even exhibit a reversing pattern (loglikelihood). Based on these considerations, we took a closer look at the models within the range of $k = 100$ (namely $k = \{75, 100, 125\}$). For these models, we manually inspected the top words representing the individual topics but also words that occur frequently within topics but are also exclusive to them (often called frex words). In addition to these qualitative inspections, we also used word intrusion as well as word-set intrusion tests (for more details see the paragraph on validation below) to check for coherence. We finally agreed on $k = 100$ topics as this setting led to better interpretable frames with consistent topics. A relatively high number of topics also make sense from the perspective of HCA, since fine-grained clusters might be grouped together in a later stage, whereas the lower number of k potentially leads to more overarching clusters not representing the individual frames. In the next step, two coders annotated all of the 100 calculated word clusters representing frames using the codebook displayed in Table A2 in Appendix.[4] Frames were labeled using the numerical representation ranging from 1 to 14 with every word cluster being assigned one label, but some labels appeared multiple times across all word clusters. Frames that could be identified as coherent but did not classify as one of the 14 frames under investigation were labeled as "Other Topics," and clusters that did not bear any meaning were annotated as "Uninterpretable." For the few cases for which no consensus was reached concerning the frame label, the coders discussed their decisions and agreed on a label. Following this process, we were able to identify 87 word clusters as meaningful. In turn, 13 were labeled as so-called boilerplate and appeared not to have a consistent meaning, 14 frames did not exhibit any clear relation to the EU or, in other words, were clearly referencing Austrian matters. An overview of the clusters representing EU-related frames with the respective words signifying the frame found by the *LDA* model is given in Table A4 in the Appendix.

True validation procedures yielding reliable and comparable measures for this approach are scarce. To ensure valid results for the *LDA* approach, we assessed the semantic and predictive validity of the model (Grimmer & Stewart, 2013) and implemented additional validation tests to examine semantic meaning (Chang et al., 2009). First, the semantic validity was examined checking whether frames are distinctive and identify a consistent range of articles. For this task, carried out by two independent coders, we labeled frames based on the top terms occurring within these frames as well as frex-words and subsequently read articles annotated with high probabilities for this frame to make sure they actually employ the frame. Second, we looked at the distribution of frames over time and compared how they responded to different real-world events. Articles employing frames concerning the economy, in general, increased after the financial crisis in 2008; similarly, articles addressing the arrival of refugees peaked following 2015, which lends predictive validity to our results. Third, we used the R package oolong (Chan & Sältzer, 2020) to conduct quantitative validity tests to check for coherence. First, word intrusion tasks present the coder with six different words from a calculated frame, one of which is an intruder. If the frame is coherent in itself, the coder should be able to identify the intruding word. Similarly, word-set intrusion presents sets of words belonging to a frame along one set that is not taken from the word distribution of this frame (Chan et al., 2009). Again, if the model outputs are coherent, the coder should be able to identify the intruding word set. For our model, two independent coders performed the tests, resulting in high precision (precision word intrusion test = 0.85; precision word-set intrusion test = 0.9).

Keyword-assisted topic modelling

The keyATM in turn is based on the predefined number of frames expected to occur within the data as well as respective seed words indicating the frames. Guided by the literature, we used 14 different

[4]Intercoder reliability for this task was deemed satisfactory with *Krippendorff's alpha* = .86

frames (Boydstun et al., 2013) as the number of keyword-led frames for the approach. Additionally, the models are adjusted to identify other word clusters inductively to leave room for unexpected findings. Similar to the procedure for the *LDA*, we calculated multiple models with a varying number of non-keyword frames ranging from 5 to 30 (14 keyword frames + $k = \{5, 10, 15, 20, 25, 30\}$). While measures for all models created look serviceable (Eshima et al., 2020), exhibiting a decreasing trend in alpha, as well as perplexity and loglikelihood, we leaned toward manual evaluation steps to determine the most useful model. Again, top- and frex-words for all models were inspected by two independent coders to examine whether they represent a coherent concept. Additionally, independent documents with high probabilities for the individual frames were looked at. Taken together with the word and word-set intrusion tests conducted for the individual models, we finally decided to agree on a model with $k = 29$ (14 keyword based and 15 non-keyword) frames as this number proved to yield the best performance.

For the selection of keywords, we applied a multi-staged process. First, we extracted terms from existing studies as well as based on the qualitative assessment of the individual frames by the authors.[5] Generating long lists with terms indicative of the frames but partially too specific for conventional news coverage. In the second step, therefore, we filtered these lists and looked for words that actually appear within the data to a certain degree (Eshima et al., 2019). Finally, we prepared a list of words occurring frequently within the corpus and scanned for terms that may be useful as markers for one of the frames under investigation. The result of those three steps, eventually, yielded sets of keywords which we used for the individual frames (see Appendix Table A5). While for some frames, we were able to draft extensive lists with up to 23 keywords (*Economic* frame), it was distinctly harder to compile lists for other frames as markers that were deemed appropriate just did not appear within the texts. As a result, we only have a single keyword for the Health and Security frame and only three keywords for the Constitutionality and Jurisprudence frame.

Again, additional validation steps were taken to ensure validity, using the oolong R package. Two independent coders performed word and word-set intrusion tests, deeming the final model useful with a precision of 0.74 and 0.83, respectively.

Supervised machine-learning

Building on the extensive corpus of manually annotated gold standard data as described above, we were able to train binary classifiers for *SML* to identify frames directly, which is generally more effective than through indicators. While existing research shows that the performance of classifiers for this purpose usually increases with the size of available training data (e.g., Burscher et al., 2014) and comparably low numbers of training documents lead to poor performances (Boumans & Trilling, 2016; Grimmer & Stewart, 2013), we decided to limit this part of the analysis to five frames for which we garnered satisfactory amounts of manually coded data.[6] Namely, we include the *Economic*, *External Regulation and Reputation*, *Law and Order*, *Crime and Justice*, *Policy Prescription and Evaluation* as well as the *Political* frame. Furthermore, we selected different classifiers based on their availability, their common use in communication science (e.g., Boumans & Trilling, 2016), and their applicability to our data. We introduced Support Vector Machines, K-Nearest-Neighbors, Gradient Boosting, Decision Tree, and Random Forest Classifier as well as an ensemble approach of the five classifiers (Burscher et al., 2014). Additionally, we implemented a Multi Layer Perceptron (Goldberg, 2017) as a neural network approach, proven to be effective for such classification tasks (F. Lind et al., 2021).[7]

[5]Projects with more resources may additionally have human coders assess texts and extract keywords from documents labeled with the respective frame.

[6]The authors are aware of the possibilities to enrich data to try and fit classifiers also with smaller training sets. However, given the objective of this study, i.e., the application of approaches, we prefer to not dive too deep into the technicalities of individual approaches so as to not get entangled with general challenges of different methods that are not specifically connected to our objective.

[7]For an overview and short introduction of the different classifiers, see Section A2 in the Appendix.

Minding the differences of the frames under investigation, we used an approach allowing different combinations and used varying parameters for each frame. Following this procedure, we compared metrics for 885 settings, contrasting different parameters for the individual classifiers (for details on the parameters varied, please refer to Section A2 in the Appendix)

All classifiers were trained using the manually annotated gold standard data. To this end, we employ 3-fold cross-validation (Japkowicz & Shah, 2016) and split the data into two equal parts to engage in a training and cross-validation procedure. All classifiers were trained using one half of the data, while the remaining part, the holdout sample, was used for validation purposes. Repeating this procedure with alternating turns of parts being used as training or validation data three times, the process yields robust performance metrics. Results show that although the ensemble outperformed single classifiers used for its implementation, the MLP approach yielded better results for all frames under investigation, achieving *F1* scores above 0.7 for all frames with acceptable levels of manually coded data available (a detailed report of the performance is provided in Table A6 in the Appendix).

Results part I: the gold standard

Before turning to the discussion and comparison of different automated approaches, we look into the results of the manual coding that serves as input for the *SML* approach and the comparisons discussed in the following. This allows us to better understand the variance of frames in the data and draw more informed conclusions on how and why the approaches discussed perform the way they do.

Of the 14 frames defined beforehand (Boydstun et al., 2013), not all were equally present in the coverage (see Figure 1). Especially the *Political* and the *Policy Prescription and Evaluation* frames, as well as the *Economic* frame, appeared often and were coded frequently. The strong prevalence of exactly these three frames can be related back to the fact that the EU has been going through a fundamental financial crisis (*Economic* frame) during the period of analysis, resulting in political challenges and intense discussions about how to reform the EU's institutional structure and specific policies (*Political* frame, *Policy Prescription and Evaluation* frame) (e.g., Bijsmans, 2017; Dutceac Segesten & Bossetta, 2019).

Given the rather high number of 14 frames and the unequal variance in frame prominence (see Figure 1) and as mentioned before, we selected five frames based on their frequent appearance in the manually coded data to train classifiers for the *SML* approach. While *LDA* and *keyATM* do not rely on "sufficient" information coming from the manual coding for fitting purposes, we proceed with all 14 frames for both of these approaches. For the last part of our analysis, the comparison of results across *LDA*, *keyATM*, and *SML*, however, we drew on the five exemplary frames (*Economic* frame, the *External Regulation and Reputation* frame, the *Law and Order* frame, the *Policy Prescription and Evaluation* frame, and the *Political* frame) as those are available for the *SML* but also to limit this part of the analysis to an overseeable amount of frames and thus be able to meaningfully compare and discuss results for specific frames.[8]

Results part II: three approaches to automated frames analysis

Latent Dirichlet Allocation (LDA)

LDA is the only fully inductive approach in our selection, allowing us, at the same time, to explore the contents of Austrian EU print news coverage. To understand how *LDA* may correspond to the frame categories suggested by Boydstun et al. (2013), we followed the HCA approach (Baden et al., 2020) as mentioned above and coded the *LDA* results according to the codebook. Additionally, we clustered such frames that seemed suitable under certain categories. For example, we clustered frames that were

[8]Please note that in order to gather enough gold standard material to train supervised machine learning classifiers, we conducted a second round of manual coding by trying to oversample distinct frames assisted by the results of the *keyATM* approach.

annotated as *Economic* according to the codebook, featuring sub-aspects such as the stock market, banking, currency policy, or economic growth. Similarly, we put word clusters referring to military policy or the war in Syria into the *Security and Defense* frame, corresponding with the frame category in the Boydstun codebook. This allows a clearer comparison with the other approaches (Jacobi et al., 2016). While this strategy can compensate for the somewhat arbitrary nature of granularity that the researcher can choose in the application of *LDA*, it does not guarantee that all frames of interest would be found. In that sense, the researcher's influence on the actual frames that the model identifies is small compared to other approaches, and the outcome may vary according to certain settings (like *alpha* or the number of *k* topics), but not with respect to the characteristics of frames which are usually predefined (e.g., Baden et al., 2020; Grimmer & Stewart, 2013).

As expected, a lot of articles deal with aspects tied to the economy and politics. This corresponds with the results of the manual coding, showing a high prominence of the *Economic* and *Political* frame (see Figure 1). Other frame categories discovered by the *LDA* model concern the *Security and Defense* frame, the *Capacity and Resources* frame as well as *Health and Safety*. Regarding the latter two, our manual coding does not correspond to this result as both frames belong to the 4 least prominent frames overall. The *Policy Prescription and Evaluation* frame found very prominent in the manual coding, is not prominent in the *LDA* output at all. The *LDA* model largely captures the frames one would expect, knowing the types of crises that dominated the political and media agenda of the EU in the last decade (e.g., Dutceac Segesten & Bossetta, 2019; Bijsmans, 2017; see Appendix Table A4 for label terms). However, these results could also suggest that frames not found prominent in the manual coding might be represented to a larger degree in Austrian EU news and vice versa, raising concerns about the validity of *LDA* as a tool for frame analysis. As is vividly discussed in current research (e.g., Nicholls & Culpepper, 2021), *LDA* might be better suited to capture issues or frames as sub-issues (Sides, 2006), as it is relying on explicit markers in the text and therefore not capable of capturing a construct as elusive and latent as a frame. In that sense, *LDA* might be a useful tool for frame analysis, but only against a thorough theoretical reflection of what this means for the definition of frames in general.

KeyATM

The rather novel *keyATM* is a semi-supervised approach (see Appendix Table A6 for a list of predefined keywords and the top words for each frame dimension; see Appendix Table A5 for original versions of key- and top words as well as information on the 15 non-keyword clusters). *KeyATM* allows for deductive as well as inductive analysis; we draw on the 14 manually coded frames (see Figure 1) introduced above for the deductive part.

While the *keyATM* model fits well according to available measures (see Methods section above; Eshima et al., 2020), a first inspection of the output indicates that the approach did not perform very well considering the representativeness of top words for specific frames aimed to measure (see Appendix Table A6). Although many measured concepts appear to be coherent themes, some obviously fail to grasp the concept aimed to capture by the keywords. An example is the *Policy Prescription and Evaluation Frame*, where keywords provided beforehand all connected to the government, measures, solutions, reforms, conditions, or strategies; the output, however, consisted of terms such as Greece, Euro, Brexit, Athens, or Great Britain, with only one of the provided keywords appearing prominently (government). While this became already apparent during the fitting, we mainly emphasized coherence and picked the best model in comparison to others at that stage. Still, for the deductively identified frames, the selected keywords seem to often either not represent a frame very well or not at all whereas, for other frames, we find some of the predefined keywords in the output of top words as well as other words fitting the frame. For example, the *Economic* frame, *External Regulation and Reputation* frame and the *Capacity & Resources* frame are well represented by the top words with four and three keywords appearing (see Appendix Table A6). For most other frames, however, top words do not appear to mirror the frames as defined in our study.

Moreover, according to the manual coding, the *Political* and *Policy Prescription and Evaluation* frames were very prominent, too. However, they show less correspondence with the keywords yielded from the literature and expert assessment: Both frame categories seem to become very specific concerning Austrian party politics and EU crises management with financial aspects and Greece as well as Brexit, respectively.

Supervised machine learning

As we can not show output similar to the top words representing frames for the *LDA* and *keyATM* approaches, we evaluate the results of the *SML* approach according to standard procedures in the field, discussing the comparison of the *SML* classifications with the results of the manual coding.

Calculated *F1* scores in Table A7 in the Appendix show the performance of different classifiers put to test for the classification task at hand. Overall, we see that the MLP approach outperformed other classifiers with up to *F1* = 0.86. Moreover, we find that compared to the other frames, the classifiers for the *Economic* frame perform best with an average *F1* score of 0.86, slightly better than the *Political* frame (0.83). Regarding the performance of the other three frames (*Policy Prescription and Evaluation* frame, average *F1* = 0.82; *External Regulation and Reputation* frame, average *F1* = 0.8; and *Law and Order, Crime and Justice* frame, average *F1 = 0.8*), we similarly obtain satisfactory results.

Results part III: comparing the validity of approaches

It is, arguably, not straightforward to directly compare the approaches implemented in this study due to their very different operating modes. Yet, all of them frequently are or are likely to be used to analyze frames and can thus be compared regarding how well they can capture frames in comparison to the manually coded gold standard data. To enhance the comparability of the approaches, we hence re-coded the output of the *LDA* and *keyATM* approaches to obtain binary classifications of the documents, representing the most prevalent frame (for more information on this step, see Section A3 in the Appendix), a step that is also recommended in the application of the HCA approach chosen for the implementation of *LDA* (Baden et al., 2020). Following this procedure, we are able to calculate recall, precision, and F1 scores for all three approaches, allowing for an easier interpretation.

Starting with the comparison of the different approaches, for the *LDA* approach, we combined all clusters relating to the same frames (for example, all 17 identified as *Economic*; see Table A4 in the Appendix) resulting in one frame, respectively (Jacobi et al., 2016). Subsequently, we compared the automated annotations of all approaches to the manually coded data (Table 1). Findings show that, expectedly, *SML* performs well (see the section on *SML* in the results chapter II). Recall and Precision are well within satisfactory levels throughout the different frames. Recall is, on average, slightly lower than Precision, signaling that the classifier might be a tad conservative and, thus, missing some cases. Nevertheless, the approach delivers strong results given the latent nature of the concept. In addition, *LDA* proves to capture the operationalizations of the *Economic* and the *Political* frame as defined in the process of the manual coding to some degree. While the comparably high mean Precision suggests that the approach is able to find frames for some cases, the low mean Recall shows that it tends to miss out on a lot of instances. The results of the *keyATM* approach, then, do not perform well in comparison to the manually annotated data, indicating that this method picks up different concepts as suggested by the top words in Appendix Table A6. Both, Recall and Precision are distinctly lower for almost all of the frames, with some (e.g., Political) showing similar patterns as for LDA, with slightly higher Precision and very low Recall.

As displayed in Table 1, some of our frames were only coded in very little numbers. Therefore, performance measures for such frames should be interpreted with caution. Frames with very low prominence were not included in the *SML* approach, as explained above.

Table 1. Comparisons of all approaches with gold standard (recall/precision/F1).

Frame	LDA	keyATM	SML	Number of Cases Positive Class
Capacity and Resources	0.21/0.08/0.11	0.06/0.05/0.05	-	118
Constitutionality and Jurisprudence	0.05/0.35/0.09	0.02/0.17/0.04	-	575
Cultural Identity	0.00/0.00/0.00	0.08/0.02/0.03	-	24
Economic	0.35/0.55/0.43	0.10/0.31/0.15	0.80/0.93/0.86	1.322
External Regulation and Reputation	0.08/0.41/0.13	0.03/0.18/0.04	0.69/0.94/0.80	653
Fairness and Equality	0.02/0.50/0.03	0.03/0.09/0.05	-	338
Health and Safety	0.12/0.32/0.18	0.04/0.04/0.04	-	87
Law and Order, Crime and Justice	0.08/0.54/0.14	0.05/0.17/0.07	0.68/0.95/0.80	416
Morality	0.00/0.00/0.00	0.06/0.02/0.03	-	67
Policy Prescription and Evaluation	0.02/0.39/0.03	0.08/0.37/0.13	0.77/0.87/0.82	1.582
Political	0.24/0.64/0.35	0.06/0.44/0.11	0.81/0.86/0.83	1.981
Public Opinion	0.08/0.26/0.13	0.08/0.37/0.13	-	173
Quality of Life	0.02/0.15/0.03	0.02/0.07/0.03	-	272
Security and Defence	0.39/0.19/0.26	0.04/0.07/0.05	-	289

Displayed Scores are Recall/Precision/F1; Scores for *SML* are the F1 scores for the best-performing classifiers (i.e., MLP), for a detailed overview of classifier performances please see Table A7 in the Appendix.

Discussion and conclusion

This study sets out to compare different machine-learning approaches to automatically measure frames regarding the implications for the validity of using one or the other. We established a Gold Standard based on an extensively tested codebook (Boydstun et al., 2013), providing future research with a robust stepping stone to continue building a systematic review of different automated approaches' validity. We then applied three machine-learning models to detect frames in our corpus, which, however, clearly differed in terms of at what place in the process deductive control over the outcomes is exercised. While for the *LDA* model deductive control only comes into play after patterns have been detected in terms of matching the outcomes to specified frames, *keyATM* requires some deductive control in terms of theoretically justified keywords that are present in the corpus before pattern extraction and again in output interpretation. Similarly, but much more intensive, deductive control is required at the onset of the *SML* model in terms of a fully coded training data set.

Regarding results for each approach individually, *LDA* as a less supervised approach is better suited to identify frames such as the *Economic* frame, represented by very salient, explicit, and very detailed clusters in the text. However, it needs to be highlighted that this could be specific to our corpus: Since the topic model cannot distinguish between frames and issues per se, the output of our analyses might also mirror the high volume of frames in contrast to issues in the media coverage of Austrian European Union politics. In that sense, many issues in EU politics could have been discussed with an *Economic* frame rather than discussing the economy as such. Moreover, researchers have no influence on what to search for with *LDA*. Results show that while varying the granularity (i.e., number of *k* topics) might increase chances of picking up less prominent frames, the *LDA* did not find certain frames of interest, a shortcoming when interested in finding predefined frames. The approach is best suited to explore corpora without prior expectations or knowledge (and hence no need for deductive control) and visualize frames as connected concepts within textual data (e.g., Ylä-Anttila et al., 2018, also; Sides, 2006). This is arguably more often the case when investigating issue-specific frames for rather novel topics. Hence, using *LDA* for frame identification should really only be applied in such cases.

The *keyATM* can accommodate the same features as *LDA* but also allows for a deductive analysis based on predefined keywords that researchers need to select beforehand and that, thus, should fit explicitly to target frames. Keywords do, however, not necessarily need to come up prominently in resulting word clusters: *keyATM* can tell us what words the predefined keywords co-occur with, but these keywords do not necessarily dominate the frame in question. Accordingly, results show that frames uncovered by this approach might deviate considerably from other machine-learning methods. The definition of keywords needs careful consideration. While the top words for the *Economy* frame, for example, appeared to reflect the concept well, annotations did not correspond with the manual coding. We gained keywords from the literature and expert assessment in accordance with frequent occurrences within the data but keywords could also be identified in other ways (e.g., keywords from manually coded material, keyword-in-context analysis, etc.). While we acknowledge that the *keyATM* approach did uncover patterns in the data that make sense and were interpretable, we note that these patterns do not fit well with our gold standard data. In that sense, it seems that our keywords which were extensively tested and discussed among the authors did not make for good frame markers. So whereas it seems that *keyATM* does not work well under the theoretical and empirical conditions we specified for this study, future research may attempt to understand if *keyATM* may be suitable under different theoretical or empirical conditions to identify frames.

Regarding *SML*, finally, classification approaches using supervised machine learning deliver the most valid results for automated frame coding. This was to be expected given that *SML* did also require the highest degree of input data. In most instances, manual coding relies on a random sampling of units as no information is available on how frames might be distributed in the data. However, the overall presence of a frame matters when it comes to manual coding, and subsequently, the application of *SML* as a sufficient amount of gold standard data is needed (e.g., Burscher et al., 2014). Our findings indicate that frames can be measured reliably and validly using *SML*, but the process can become very resource-intensive, especially when frames are only marginally present in the manually coded gold standard data and additional rounds of manual coding with strategies for oversampling are necessary. *SML* might not yield valid results with very sparse training data, yet, compared to *LDA*, oversampling of rare instances or using artificially augmented data could be capable of dealing with marginally present frames. *LDA* will only find what is somewhat salient in the text and rare instances of frames might only be captured by a very fine granularity.

Overall, thus, *SML*, as expected, clearly outperforms the other approaches. And while the performance for the *Economic* frame as measured through *LDA* is surprisingly good, our results support the criticism expressed by other researchers (e.g., Nicholls and Culpepper, 2021), in that topic modeling should be only one step in the automated analysis of frames to ensure the validity of the measurement. In that sense, automated measures in general might be best used in a complementary way to increase validity (e.g., Guo et al., 2022; Walter & Ophir, 2019) while clearly acknowledging the role of prior knowledge and from such knowledge in the processes.

While in our study *LDA* and *SML* rely on combinations of words (that is feature vectors for *SML* and co-occurrences in the case of *LDA*), both approaches are not bound to specific markers but "learn" them through the data. It is interesting to note that both work better than the approach that is in-between in terms of the degree of supervision and deductive control. This is of course less surprising concerning the *SML* approach because it relies on most a priori information in line with what we were looking for. For *LDA*, it seems that our approach to framing may identify frames that are hard to bind to specific markers, that is keywords, in the text. Given that *keyATM* does, however, require the specification of such markers from both a theoretical, yet also empirical perspective, we can infer that a simple specification of a few of such keywords does not do the trick for the comprehensive identification of frames as studied here. The frames do connect to specific combinations of words/markers as we see in the more successful application of the *LDA* model, but it would be incredibly hard to know these beforehand as required for *keyATM*. Despite seeming a useful approach, we, therefore, recommend being very cautious about using *keyATM* for frame identification for a conceptualization of frames that speaks to the mainstream of framing research.

The tension between automation and interpretation (Jünger et al., 2022) could be further alleviated by researchers more consciously and explicitly acknowledging the conceptual limitations of the selected approach. Thus, studies using topic modeling should be more reflective about what they cannot measure and tone down their ambitions accordingly. A "frame" may, for example, correspond rather to a sub-issue (Sides, 2006), and thus provide a valuable and insightful contribution to the discussion, satisfying a less demanding definition of a frame.

Of course our study has its limitations. First, it must be noted that the reliability of our Gold Standard data is not ideal. While this is often the case in frames analysis (see, e.g., Matthes & Kohring, 2009, p. 258), given the elusiveness and context sensitivity of the concept, the conclusions drawn from our research, especially for fully supervised approaches, need to be taken with a grain of salt. Furthermore, our study is limited in that one specific definition and the codebook was used. As discussed in our study, the diversity of approaches to framing and the resulting fuzziness of empirical approaches to measuring frames represents an important hurdle to universal methods of frame identification. We tried to alleviate this shortcoming by relying on the most prevalent definition of framing in the field (Entman, 1993) and a codebook that can be used by other researchers with similar interests (Boydstun et al., 2013) and lends itself for comparisons across contexts. However, future research should also test our results against other perspectives on news frames, both generic (as for example, used by Semetko and Valkenburg (2000)) and issue-specific frames (as for example, applied by Nelson and Oxley (1999) or Cappella and Jamieson (1997)) to allow a more comprehensive discussion on different approaches to frames and their identification in an automated way.

Moreover, we detail a specific way of applying approaches in this study, while there are many parameters that could be set differently with potentially different results. This concerns, for example, preprocessing steps, e.g., lemmatizing or removing stopwords, but also hyperparameters of the model itself, deleting named entities from the text, and more. For *SML*, the researcher will usually look for the algorithm performing best against the Gold Standard. In this respect, it is less clear for *LDA* or *keyATM* when the model output has reached sufficient validity. For *LDA*, specifically, given that it is a stochastic approach, rerunning the same code on a different computer may also slightly alter the results of the topic model again. In addition, we moved beyond the dominant focus on the English language by analyzing a large corpus of 11 years of German language Austrian EU news; but it would be crucial to building a thorough basis of information for researchers interested in automated frame analysis in languages beyond the Western focus.

Last but not least, the field of computational text analysis is a fast-developing one and in that sense, the selection of approaches in this study is limited in that more recent approaches, such as BERTopic (Grootendorst, 2022) for topic modeling or other transformer-based approaches, for example, were not considered. Future research should build on our approach to expand our systematic knowledge of the comparative performance of the wide array of different approaches to automated frame analysis.

In conclusion, this paper set out to systematically compare different approaches to automated frame analysis with a special focus on the validity. We discussed machine-learning approaches due to their increasing popularity, limiting ourselves to three inductive and deductive approaches. Comparing the results with the manually annotated Gold Standard shows that valid and reliable results do need resources. But also we need to better acknowledge where and to what degree we exercise deductive control over our procedures and thus outcomes. Further, in order for social science to implement automated approaches fruitfully, we will need to develop more accessible ways of using new technologies for alleviating a "digital divide" (e.g., Guo et al., 2022; Watanabe, 2021). In order to foster the validity of research in the social sciences, we must also work toward further establishing the insight that letting the computer do the work does not make our lives easier but requires careful validation of its output (e.g., Song et al., 2020). In that sense, as also demonstrated in our study, less input may come at the cost of validity, especially regarding a complex and latent concept like frames.

The variety of automated text analysis applied in the social sciences is expanding fast, which is why future research needs to provide the necessary guidance for researchers to enable them to implement approaches in a valid and reliable manner. The analyses detailed here, while limited in their range, provide a stepping stone; future studies could add to the discussion by, for example, comparing other approaches based on the same frame codebook (Boydstun et al., 2013). In that way, the social sciences could establish a reference framework for a more informed and valid implementation of automated frames analysis.

Acknowledgments

Please note that the two first authors shared the lead on this paper. The work was supported by the Austrian Science Fund (FWF) under Grant T-989 G27; and by the Austrian National Bank (Anniversary Fund) under Grant 18120. In addition, we used the Application Programming Interface of the Austria Press Agency (APA-API), represented by the UniVie Data Project, for downloading Austrian news content for our analyses. Last but not least, the authors would like to thank Huizhi Bao for her support.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Olga Eisele (Ph.D., University of Vienna) is an Assistant Professor at the Amsterdam School of Communication Research (ASCoR), University of Amsterdam. Her research interests lie in crisis communication in the context of European integration, with a special interest in political legitimation processes and the relationship between media and politics. She also works on advancements in text analysis methods.

Tobias Heidenreich (Ph.D., University of Vienna) was a research associate at the Department of Communication, University of Vienna and is now a postdoctoral researcher at the Berlin Social Science Center (WZB). His main research interests include quantitative text analysis, political communication, and social media.

Olga Litvyak (Ph.D., University of Lausanne) is a postdoctoral researcher at the Department of Communication, University of Vienna. Her research focuses on party competition and political communication in Europe, especially issue and framing dynamics.

Hajo G. Boomgaarden (Ph.D., University of Amsterdam) is professor for empirical social science methods with a focus on text analysis at the Department of Communication at University of Vienna. His research interests include the coverage and effects of political information on citizens' cognitions, attitudes and behaviors in various domains of media and politics, and developments in automated content analysis techniques.

ORCID

Olga Eisele http://orcid.org/0000-0002-6604-3498
Tobias Heidenreich http://orcid.org/0000-0001-9070-0550
Olga Litvyak http://orcid.org/0000-0002-1277-827X
Hajo G. Boomgaarden http://orcid.org/0000-0002-5260-1284

References

Akyürek, A. F., Guo, L., Elanwar, R., Ishwar, P., Betke, M., & Wijaya, D. T. (2020). Multi-label and multilingual news framing analysis. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8614–8624. https://doi.org/10.18653/v1/2020.acl-main.763

Auel, K., & Pollak, J. (2016). Österreich. In *Jahrbuch der Europäischen Integration 2016* (pp. 547–552). Institut für Europäische Politik. https://doi.org/10.5771/9783845275642-546

Baden, C., Kligler-Vilenchik, N., & Yarchi, M. (2020). Hybrid content analysis: Toward a strategy for the theory-driven, computer-assisted classification of large text corpora. *Communication Methods and Measures*, *14*(3), 165–183. https://doi.org/10.1080/19312458.2020.1803247

Baden, C., Pipal, C., Schoonvelde, M., & van der Velden, M. A. C. G. (2021). Three gaps in computational text analysis methods for social sciences: A research agenda. *Communication Methods and Measures*, *16*(1), 1–18. https://doi.org/10.1080/19312458.2021.2015574

Bateson, G. (1955). A theory of play and fantasy; a report on theoretical aspects of the project of study of the role of the paradoxes of abstraction in communication. *Psychiatric Research Reports*, *2*, 39–51.

Baumgartner, F. R., De Boef, S. L., & Boydstun, A. E. Q. (2008). *The decline of the death penalty and the discovery of innocence*. Cambridge University Press.

Bijsmans, P. (2017). EU media coverage in times of crisis: Euroscepticism becoming mainstream? In *Euroscepticism, democracy and the media* (pp. 73–94). Palgrave Macmillan UK. https://doi.org/10.1057/978-1-137-59643-7_4

Bjarnøe Jensen, C. (2016). Evolution in Frames: Framing and Reframing of Policy Questions [University of Aarhus]. https://pure.au.dk/portal/en/publications/evolution-in-frames-framing-and-reframing-of-policy-questions(59bba768-40ac-4063-9d06-082249a06e50).html

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Boomgaarden, H. G., & De Vreese, C. H. (2016). Do European elections create a European public sphere? In W. Van der Brug & C. H. De Vreese (Eds.), *(Un)intended consequences of european parliamentary elections* (pp. 1., pp. 19–35). Oxford University Press.

Boumans, J. W., & Trilling, D. (2016). Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital Journalism*, *4*(1), 8–23. https://doi.org/10.1080/21670811.2015.1096598

Boydstun, A. E., Gross, J. H., Resnik, P., & Smith, N. A. (2013). Identifying media frames and frame dynamics within and across policy issues. *New Directions in Analyzing Text as Data Workshop*, 1–13. http://www.cs.cmu.edu/~nasmith/temp/frames-2013.pdf

Brookes, G., & McEnery, T. (2019). The utility of topic modelling for discourse studies: A critical evaluation. *Discourse Studies*, *21*(1), 3–21. https://doi.org/10.1177/1461445618814032

Brouwer, J., van der Woude, M., & van der Leun, J. (2017). Framing migration and the process of crimmigration: A systematic analysis of the media representation of unauthorized immigrants in the Netherlands. *European Journal of Criminology*, *14*(1), 100–119. https://doi.org/10.1177/1477370816640136

Brugman, B. C., & Burgers, C. (2018). Political framing across disciplines: Evidence from 21st-century experiments. *Research & Politics*, *5*(2), 205316801878337. https://doi.org/10.1177/2053168018783370

Burscher, B., Odijk, D., Vliegenthart, R., de Rijke, M., & de Vreese, C. H. (2014). Teaching the computer to code frames in news: Comparing two supervised machine learning approaches to frame analysis. *Communication Methods and Measures*, *8*(3), 190–206. https://doi.org/10.1080/19312458.2014.937527

Burscher, B., Vliegenthart, R., & Vreese, C. H. D. (2016). Frames beyond words: Applying cluster and sentiment analysis to news coverage of the nuclear power issue. *Social Science Computer Review*, *34*(5), 530–545. https://doi.org/10.1177/0894439315596385

Borah, P. (2011). Conceptual Issues in Framing Theory: A Systematic Examination of a Decade's Literature. *Journal of Communication*, *61*(2), 246–263.

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

Cacciatore, M. A., Scheufele, D. A., & Iyengar, S. (2016). The end of framing as we know it . . . and the future of media effects. *Mass Communication and Society*, *19*(1), 7–23. https://doi.org/10.1080/15205436.2015.1068811

Calabrese, C., Anderton, B. N., & Barnett, G. A. (2019). Online representations of "Genome Editing" uncover opportunities for encouraging engagement: A semantic network analysis. *Science Communication*, *41*(2), 222–242. https://doi.org/10.1177/1075547018824709

Cappella, J. N., & Jamieson, K. H. (1997). *Spiral of cynicism: The press and the public good*. Oxford University Press.

Card, D., Boydstun, A. E., Gross, J. H., Resnik, P., & Smith, N. A. (2015). The media frames corpus: Annotations of frames across issues. ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference, 2, 438–444 https://doi.org/10.3115/v1/p15-2072

Chan, C. H., & Sältzer, M. (2020). Oolong: An R package for validating automated content analysis tools. *Journal of Open Source Software*, *5*(55), 2461. https://doi.org/10.21105/joss.02461

Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., & Blei, D. M. (2009, December). Reading tea leaves: How humans interpret topic models. *Neural Information Processing Systems*, *22*, 288–296. https://proceedings.neurips.cc/paper_files/paper/2009/file/f92586a25bb3145facd64ab20fd554ff-Paper.pdf

Dayrell, C. (2019). Discourses around climate change in Brazilian newspapers: 2003–2013. *Discourse & Communication*, *13*(2), 149–171. https://doi.org/10.1177/1750481318817620

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

De Swert, K. (2012). *Calculating inter-coder reliability in media content analysis using Krippendorff's Alpha*. Center for Politics and Communication, University of Amsterdam. http://www.polcomm.org/wp-content/uploads/ICR01022012.pdf

De Vreese, C. H. (2005). News framing: Theory and typology. *Information Design Journal + Document Design*, *13*(1), 51–62. https://doi.org/10.1075/idjdd.13.1.06vre

Dutceac Segesten, A., & Bossetta, M. (2019). Can Euroscepticism contribute to a European public sphere? The Europeanization of media discourses on euroscepticism across six countries. *JCMS: Journal of Common Market Studies*, *57*(5), 1051–1070. https://doi.org/10.1111/jcms.12871

Egger, R., & Yu, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*, *7*, 886498.

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, *43*(4), 51–58. https://doi.org/10.1111/j.1460-2466.1993.tb01304.x

Eshima, S., Imai, K., & Sasaki, T. (2020). *Keyword Assisted Topic Models*. http://arxiv.org/abs/2004.05964

Eshima, S., Sasaki, T., & Imai, K. (2019). *keyAtm: Keyword Assisted Topic Models*. R Package Version 0.4.0. https://keyatm.github.io/keyATM/

Field, A., Kliger, D., Wintner, S., Pan, J., Jurafsky, D., & Tsvetkov, Y. (2018). Framing and agenda-setting in Russian news: A computational analysis of intricate political strategies. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3570–3580. https://doi.org/10.18653/v1/D18-1393

Gamson, W. A., & Modigliani, A. (1989). Media Discourse and Public Opinion on Nuclear Power: A Constructionist Approach. *The American Journal of Sociology*, *95*(1), 1–37. https://www.jstor.org/stable/2780405

Gilardi, F., Shipan, C. R., & Wüest, B. (2021). Policy diffusion: The issue-definition stage. *American Journal of Political Science*, *65*(1), 21–35. https://doi.org/10.1111/ajps.12521

Gitlin, T. (1980). *The whole world is watching: Mass media in the making & unmaking of the New Left*. University of California Press.

Goldberg, Y. (2017). Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, *10*(1), 1–309. https://doi.org/10.2200/S00762ED1V01Y201703HLT037

Goffman, E. (1974). *Frame analysis: An essay on the organization of experience*. Harvard University Press.

Greussing, E., & Boomgaarden, H. G. (2017). Shifting the refugee narrative? An automated frame analysis of Europe's 2015 refugee crisis. *Journal of Ethnic and Migration Studies*, *43*(11), 1749–1774. https://doi.org/10.1080/1369183X.2017.1282813

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

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794). arXiv. https://doi.org/10.48550/arXiv.2203.05794

Grundmann, R. (2021). Using large text news archives for the analysis of climate change discourse: Some methodological observations. *Journal of Risk Research*, 1–12. https://doi.org/10.1080/13669877.2021.1894471

Günther, E., & Quandt, T. (2016). Word counts and topic models. *Digital Journalism*, *4*(1), 75–88. https://doi.org/10.1080/21670811.2015.1093270

Guo, L., Su, C., Paik, S., Bhatia, V., Akavoor, V. P., Gao, G., Betke, M., & Wijaya, D. (2022). Proposing an open-sourced tool for computational framing analysis of multilingual data. *Digital Journalism*, *11*(2), 276–297. https://doi.org/10.1080/21670811.2022.2031241

Guo, L., Vargo, C. J., Pan, Z., Ding, W., & Ishwar, P. (2016). Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling. *Journalism and Mass Communication Quarterly*, *93*(2), 322–359. https://doi.org/10.1177/1077699016639231

Hänggli, R. (2020). *The Origin of Dialogue in the News Media*. Palgrave Macmillan. https://doi.org/10.1007/978-3-030-26582-3

Heidenreich T, Lind F, Eberl J and Boomgaarden H G. (2019). Media Framing Dynamics of the 'European Refugee Crisis': A Comparative Topic Modelling Approach. Journal of Refugee Studies, 32(Special_Issue_1), i172–i182. 10.1093/jrs/fez025

Hellsten, I., Dawson, J., & Leydesdorff, L. (2010). Implicit media frames: Automated analysis of public debate on artificial sweeteners. *Public Understanding of Science*, *19*(5), 590–608. https://doi.org/10.1177/0963662509343136

Hertog, J. K., & McLeod, D. M. (2001). A multiperspectival approach to framing analysis: A field guide. In S. D. Reese, O. H. Gandy, & A. E. Grant (Eds.), *Framing public life: Perspectives of media and our understanding of the social world* (pp. 157–158). Routledge.

Hobolt, S. B., & De Vries, C. E. (2016). Public support for European integration. *Annual Review of Political Science*, *19*(1), 413–432. https://doi.org/10.1146/annurev-polisci-042214-044157

Hornik, K., & Grün, B. (2011). Topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, *40*(13), 1–30. https://doi.org/10.18637/jss.v040.i13

Hubner, A. (2011). How did we get here? A framing and source analysis of early COVID-19 media coverage. *Communication Research Reports*, *38*(2), 112–120. https://doi.org/10.1080/08824096.2021.1894112

Jacobi, C., Van Atteveldt, W., & Welbers, K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, *4*(1), 89–106. https://doi.org/10.1080/21670811.2015.1093271

Japkowicz, N., & Shah, M. (2011). *Evaluating learning algorithms: A classification perspective*. Cambridge University Press.

Jiang, K., Barnett, G. A., & Taylor, L. D. (2016). Dynamics of culture frames in international news coverage: A semantic network analysis. *International Journal of Communication*, Vol. *10*, 27. https://ijoc.org/index.php/ijoc/article/view/4484

Jünger, J., Geise, S., & Hänelt, M. (2022). Unboxing computational social media research from a data hermeneutical perspective: How do scholars address the tension between automation and interpretation? *International Journal of Communication*, Vol. *16*, 1482–1505. https://ijoc.org/index.php/ijoc/article/view/1747

Kroon, A., Van der Meer, T., & Vliegenthart, R. (2022). Beyond counting words: Assessing performance of dictionaries, supervised machine learning, and embeddings in topic and frame classification. *Computational Communication Research*, *4*(2), 528–570. https://doi.org/10.5117/CCR2022.2.006.KROO

Kwak, H., An, J., & Ahn, Y. Y. (2020, July). A systematic media frame analysis of 1.5 million new york times articles from 2000 to 2017. In 12th ACM Conference on Web Science. (pp. 305–314). New York, NY, United States: Association for Computing Machinery. https://doi.org/10.1145/3394231.3397921

Lawlor, A. (2015). Framing immigration in the Canadian and British news media. *Canadian Journal of Political Science/Revue Canadienne de Science Politique*, *48*(2), 329–355. https://doi.org/10.1017/S0008423915000499

Lawlor, A., & Tolley, E. (2017). Deciding who's legitimate: News media framing of immigrants and refugees. *International Journal of Communication*, *11*(0), 25. https://ijoc.org/index.php/ijoc/article/view/6273/1946

Lind, F., Eberl, J.-M., Heidenreich, T., & Boomgaarden, H. G. (2020). When the Journey is as Important as the Goal: A Roadmap to Multilingual Dictionary Construction. *International Journal of Communication*, *13*, 4000–4020. https://ijoc.org/index.php/ijoc/article/view/10578/2768

Lind, F., Heidenreich, T., Kralj, C., & Boomgaarden, H. (2021). Greasing the wheels for comparative communication research: Supervised text classification for multilingual corpora. *Computational Communication Research*, *3*(3), 1–30. https://doi.org/10.5117/CCR2021.3.001.LIND

Lind, R. A., & Salo, C. (2002). The framing of feminists and feminism in news and public affairs programs in U.S. electronic media. *Journal of Communication*, *52*(1), 211–228. https://doi.org/10.1111/j.1460-2466.2002.tb02540.x

Liu, S., Guo, L., Mays, K., Betke, M., & Wijaya, D. T. (2019). Detecting frames in news headlines and its application to analyzing news framing trends surrounding U.S. gun violence. *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 504–514. https://doi.org/10.18653/v1/K19-1047

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., & Adam, S. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, *12*(2–3), 93–118. https://doi.org/10.1080/19312458.2018.1430754

Matthes, J. (2009). What's in a frame? A content analysis of media framing studies in the world's leading communication journals, 1990-2005. *Journalism & Mass Communication Quarterly*, *86*(2), 349–367. https://doi.org/10.1177/107769900908600206

Matthes, J., & Kohring, M. (2009). The content analysis of media frames: Toward improving reliability and validity. *Journal of Communication*, *58*(2), 258–279. https://doi.org/10.1111/j.1460-2466.2008.00384.x

Murakami, A., Thompson, P., Hunston, S., & Vajn, D. (2017). 'What is this corpus about?': Using topic modelling to explore a specialised corpus. *Corpora*, *12*(2), 243–277. https://doi.org/10.3366/cor.2017.0118

Nassar, R. (2020). Framing refugees: The impact of religious frames on U.S. partisans and consumers of cable news media. *Political Communication*, *37*(5), 593–611. https://doi.org/10.1080/10584609.2020.1723753

Nelson, T. E., & Oxley, Z. M. (1999). Issue framing effects on belief importance and opinion. *The Journal of Politics*, *61*(4), 1040–1067. https://doi.org/10.2307/2647553

Nerlich, B., Forsyth, R., & Clarke, D. (2012). Climate in the news: How differences in media discourse between the US and UK reflect national priorities. *Environmental Communication*, *6*(1), 44–63. https://doi.org/10.1080/17524032.2011.644633

Neuendorf, K. (2002). *The content analysis guidebook*. SAGE Publications.

Nicholls, T., & Culpepper, P. D. (2021). Computational Identification of Media Frames: Strengths, Weaknesses, and Opportunities. *Political Communication*, *38*(1–2), 159–181. https://doi.org/10.1080/10584609.2020.1812777

Nicholls, T., & Culpepper, P. D. (2020). Computational identification of media frames: strengths, weaknesses, and opportunities. *Political Communication*. https://doi.org/10.1080/10584609.2020.1812777

Poirier, W., Ouellet, C., Rancourt, M.-A., Béchard, J., & Dufresne, Y. (2020). (Un)covering the COVID-19 pandemic: Framing analysis of the crisis in Canada. *Canadian Journal of Political Science/Revue Canadienne de Science Politique*, *53*(2), 365–371. https://doi.org/10.1017/S0008423920000372

Potter, W. J., & Levine-Donnerstein, D. (1999). Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research*, *27*(3), 258–284. https://doi.org/10.1080/00909889909365539

Roberts, M. E., Stewart, B. M., & Tingley, D. (2016). Navigating the local modes of big data: The case of topic models. In R. M. Alvarez (Ed.), *Computational social science: Discovery and prediction* (pp. 51–97). Cambridge University Press. https://doi.org/10.1017/CBO9781316257340.004

Schultz, F., Kleinnijenhuis, J., Oegema, D., Utz, S., & van Atteveldt, W. (2012). Strategic framing in the BP crisis: A semantic network analysis of associative frames. *Public Relations Review*, *38*(1), 97–107. https://doi.org/10.1016/j.pubrev.2011.08.003

Semetko, H. A., & Valkenburg, P. M. (2000). Framing European politics: A content analysis of press and television news. *Journal of Communication*, *50*(2), 93–109. https://doi.org/10.1111/j.1460-2466.2000.tb02843.x

Shah, D. V., Watts, M. D., Domke, D., & Fan, D. P. (2002). News framing and cueing of issue regimes: Explaining clinton's public approval in spite of scandal*. *Public Opinion Quarterly*, *66*(3), 339–370. https://doi.org/10.1086/341396

Sides, J. (2006). The origins of campaign agendas. *British Journal of Political Science*, *36*(3), 407–436. https://doi.org/10.1017/S0007123406000226

Snow, D. A., & Benford, R. D. (2005). Clarifying the Relationship between Framing and Ideology. In H. Johnston & J. A. Noakes (Eds.), *Frames of Protest: Social Movements and the Framing Perspective* (pp. 205–216). Rowman & Littlefield Publishers.

Snow, D. A., Rochford, E. B., Worden, S. K., & Benford, R. D. (1986). Frame Alignment Processes, Micromobilization, and Movement Participation. *American sociological review*, *51*(4), 464–464. https://doi.org/10.2307/2095581

Song, H., Tolochko, P., Eberl, J. M., Eisele, O., Greussing, E., Heidenreich, T., Lind, F., Galyga, S., & Boomgaarden, H. G. (2020). In validations we trust? The impact of imperfect human annotations as a gold standard on the quality of validation of automated content analysis. *Political Communication*. https://doi.org/10.1080/10584609.2020.1723752

Rauh, C. (2018). Validating a sentiment dictionary for German political language—A workbench note. *Journal of Information Technology & Politics*, *15*(4), 319–343. https://doi.org/10.1080/19331681.2018.1485608

Van Atteveldt, W., & Peng, T. Q. (2018). When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. *Communication Methods and Measures*, *12*(2–3), 81–92. https://doi.org/10.1080/19312458.2018.1458084

Van Gorp, B. (2005). Where is the frame? *European Journal of Communication*, *20*(4), 484–507. https://doi.org/10.1177/0267323105058253

Van Gorp, B. (2007). The constructionist approach to framing: Bringing culture back in. *Journal of Communication*, *57*(1), 60–78. https://doi.org/10.1111/j.1460-2466.2006.00329_4.x

Vliegenthart, R. (2012). Framing in mass communication research - an overview and assessment. *Sociology Compass*, *6*(12), 937–948. https://doi.org/10.1111/soc4.12003

Vliegenthart, R., & Van Zoonen, L. (2011). Power to the frame: Bringing sociology back to frame analysis. *European Journal of Communication*, *26*(2), 101–115. https://doi.org/10.1177/0267323111404838

Wallace, R. (2018). Contextualizing the crisis: The framing of Syrian refugees in Canadian print media. *Canadian Journal of Political Science/Revue Canadienne de Science Politique*, *51*(2), 207–231. https://doi.org/10.1017/S0008423917001482

Walter, D., & Ophir, Y. (2019). News frame analysis: An inductive mixed-method computational approach. *Communication Methods and Measures*, *13*(4), 248–266. https://doi.org/10.1080/19312458.2019.1639145

Watanabe, K. (2021). Latent semantic scaling: A semisupervised text analysis technique for new domains and languages. *Communication Methods and Measures*, *15*(2), 81–102. https://doi.org/10.1080/19312458.2020.1832976

Wilkerson, J., & Casas, A. (2017). Large-scale computerized text analysis in political science: Opportunities and challenges. *Annual Review of Political Science*, *20*(1), 529–544. https://doi.org/10.1146/annurev-polisci-052615-025542

Ylä-Anttila, T., Eranti, V., & Kukkonen, A. (2018). Topic modeling as a method for frame analysis: Data mining the climate change debate in India and the USA OSF Preprints . https://doi.org/10.31235/osf.io/dgc38

Zhang, Y., & Trifiro, B. (2022). Who portrayed it as "The Chinese Virus"? An analysis of the multiplatform partisan framing in U.S. news coverage about China in the COVID-19 pandemic. *International Journal of Communication*, *16*(0), 24.