SC|M

Studies in Communication and Media

FULL PAPER

Developing a synthetic news corpus to validate generic frame detection methods

Entwicklung eines synthetischen Nachrichten-Corpus zur Validierung von Methoden zur Detektion generischer Frames

Chung-hong Chan, Rainer Freudenthaler & Philipp Müller

Chung-hong Chan (Dr.), GESIS – Leibniz-Institut für Sozialwissenschaften, Unter Sachsenhausen 6-8, 50667 Cologne, Germany. Contact: chung-hong.chan(at)gesis.org. ORCID: https://orcid.org/0000-0002-6232-7530

Rainer Freudenthaler (Dr.), Mannheimer Zentrum für Europäische Sozialforschung, A5, 6 (Bauteil A), 68159 Mannheim, Germany. Contact: rainer.freudenthaler(at)uni-mannheim.de

Philipp Müller (Dr.), Institute for Media and Communication Studies, University of Mannheim B 6, 30–32, 68159 Mannheim, Germany. Contact: p.mueller(at)uni-mannheim.de. ORCID: https://orcid.org/0000-0002-5351-0608

FULL PAPER

Developing a synthetic news corpus to validate generic frame detection methods

Entwicklung eines synthetischen Nachrichten-Corpus zur Validierung von Methoden zur Detektion generischer Frames

Chung-hong Chan, Rainer Freudenthaler & Philipp Müller

Abstract: Frames are a central concept in communication research. Based on our literature review, we propose that frame identification is an act of identifying selected reality and communicative intention. We then highlight the conceptual and methodological issues of frame identification using computational methods. To avoid the correlation between topics and frames, we provide a synthetic dataset for evaluating frames found in multi-topical news content, using the detection of generic frames as a test case. With this dataset, for the first time, we benchmark manual coding and various automatic and semi-supervised methods. Based on the preliminary benchmark results, this study provides evidence that generic frame identification using both manual coding and automatic methods might not be accurate.

Keywords: Frame, unsupervised method, topic model, semi-supervised method, validity.

Zusammenfassung: Frames stellen ein zentrales Konzept der Kommunikationsforschung dar. Auf der Basis eines Literaturüberblicks schlagen wir vor, dass das Erfassen von Frames als Akt der Identifikation selektiver Realitätskonstruktionen bzw. kommunikativer Absichten aufgefasst werden sollte. Anschließend arbeiten wir die konzeptuellen und methodologischen Herausforderungen der Erfassung von Frames mittels Computational Methods heraus. Um die Korrelation von Berichterstattungs-Themen und Frames zu verhindern, stellen wir einen synthetischen Datensatz zur Ermittlung von Frames in multithematischen Nachrichteninhalten bereit. Hierfür greifen wir auf das Konzept der generischen Frames zurück. Mit Hilfe dieses Datensatzes vergleichen wir erstmals verschiedene Methoden der manuellen sowie automatisierten oder semi-überwachten Ermittlung von Frames. Auf der Basis erster Benchmark-Ergebnisse liefert die vorliegende Studie Hinweise darauf, dass die Erfassung generischer Frames sowohl mit Hilfe manueller Codierung als auch mittels automatisierter Methoden möglicherweise nicht genau genug ist.

Schlagwörter: Frame, unüberwachte Methoden, Themenmodell, semi-überwachte Methoden, Validität.

1. Introduction

The goal of this study is to synthesize a dataset for evaluating generic frames found through different methods. In order to achieve this goal, we first review the concept of (generic) frames in communication research and then highlight the conceptual and methodological issues of (generic) frame identification. Then we outline our approach to synthesize such a dataset, analyze the dataset, and report the lessons learned. The most important lesson: Identification of (generic) frames from news content written by someone else is an incredibly difficult task, even for experts.

2. Entmanian frame and the tacit aspect of communicative intention

The notion of (media) frame is probably one of the central concepts in communication research. As of writing, a simple keyword search of "Frame" returned 1278 results from *Journal of Communication* alone. In several journals of the field, special issues have been published to solely interrogate this central concept (e.g., *Journal of Communication*, *57*(1); *Media, War & Conflict*, *11*(4)).

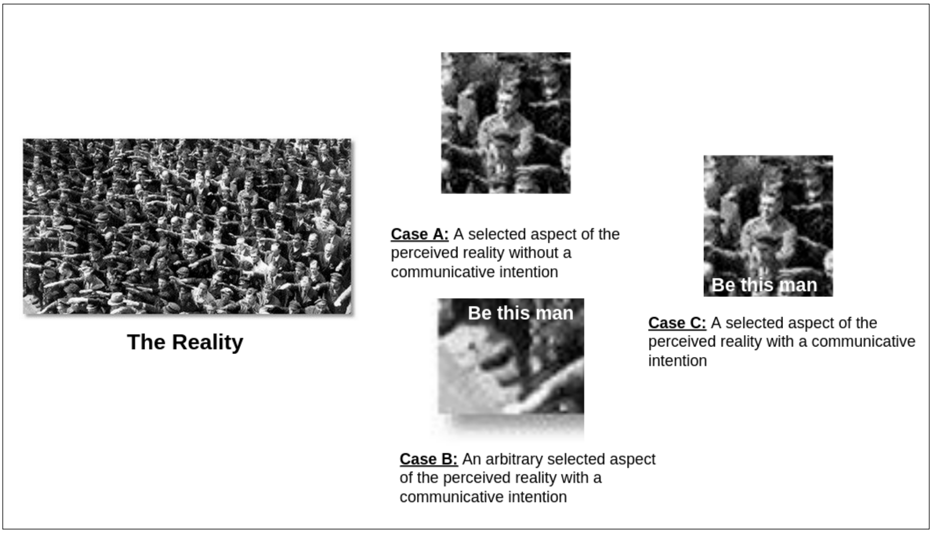
Before the onset of the so-called "Computational Turn" of journalism research (Hase et al., 2022) and the notion of automated content analysis (Boumans & Trilling, 2015), methodological controversies surrounding the detection of frames have already been a greatly discussed topic even in the context of traditional manual content analysis. Even the concept itself has been defined and redefined by various experts. The contested (D'Angelo, 2002), but highly cited, definition by Entman (1993) states that framing (an act, i.e. a verb) is "select[ing] some aspects of a perceived reality and mak[ing] 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" (p. 52, emphasis added). By using "to promote" in active voice indicates that the act of framing in Entman's sense is conducted by the communicators, not the recipients.

We restate the Entmanian definition of *framing* (verb) to define the noun *frame*: A frame is the result of an act of selecting and making salient certain aspects of a perceived reality by a communicator whose intention is to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation. Our restatement makes explicit the tacit aspect of *communicative intention* in the original definition. Our restatement is also compatible with Scheufele and Tewksbury (2006)'s "three models of political communication," which differentiate between framing, agenda setting, and priming. In their models, framing refers to "modes of presentation that journalists and other communicators use to present information in a way that resonates with existing underlying schemas among their audience" (p.12). The underlying communicative intention is then to "resonate with existing underlying schemes among their audience." Similarly, Baden (2015)'s notion of "interpretative frame" also focuses on strategic and constructive purposes. In psychology, the research on framing also deals with the communicative intention of influencing choices and decisions through different ways to represent the same information (e.g., Tversky & Kahneman, 1981).

With this argument we do not mean to imply that communicators, journalists in particular, always think about which frame to choose and consciously adopt a specific frame for a story. Within framing theory, the framing process is understood as allowing for actors to unconsciously adapt frames that have been communicated by other actors. For example, the strategic framing of issues by politicians can lead to journalists adopting (consciously or unconsciously) a specific frame within their reporting (Matthes, 2014, pp. 14–19). A journalist, for instance, can pick up the economic frame on climate change and then frame their reporting in terms of the costs of mitigating climate change, without reflecting upon the fact that other ways to perceive the problem exist. Still, we argue, for a story to communicate a frame, it must communicate the intention contained within the frame, or the frame is not communicated. The journalist in this example would have to communicate the intention that climate change should be seen as an economic problem, even if they do not intentionally rule out other frames. Simply put: A frame contains communicative intention even if it does not express authorial intent.

Making the tacit aspect of communicative intention explicit raises several questions about frame detection. The most obvious is: What exactly is the act of detecting frames? Should it be judging which aspects of a perceived reality have been selected and emphasized by a communicator? Or judging the original communicative intention of the communicator from the text? We propose that frame detection is an act of detecting both (selected aspect of a perceived reality and communicative intention) and we can't tell a frame from texts by just detecting either one. We explain this problem by a visual metaphor (Figure 1).

Figure 1. A metaphor for frame detection


The Reality
Case A: A selected aspect of the perceived reality without a communicative intention
Be this man
Case B: An arbitrary selected aspect of the perceived reality with a communicative intention
Be this man
Case C: A selected aspect of the perceived reality with a communicative intention

Suppose the crowd giving the Nazi salute is the reality. If we detect the man who does not give the Nazi salute (a selected aspect of a perceived reality) and claim that to be a “frame” (Case A), it might resonate with the audience but with an ambiguous communicative intention: a *Gestapo* officer might select this aspect with the communicative intention to ask other citizens to hunt for this man; or a resistance fighter might select this aspect with a communicative intention to save this man or to ask other citizens to be like this man.

In another direction, we might know the communicative intention of the communicator. However, this communicator selects an arbitrary aspect to convey their intention (Case B). This cannot be a “frame” either because the selected aspect does not convey their communicative intention. Only after both the selected aspect of a perceived reality and the communicative intention are detected can we unambiguously say what the frame is (Case C).

3. The many approaches of identifying frames empirically/computationally

There has been criticism on how researchers detect frames empirically. Carragee and Roefs (2004) criticize that some researchers “reduce frames to story topics, attributes, or issue position” (p. 217). A harsher criticism from Reese (2007) is for framing researchers to “give an obligatory nod to the literature before proceeding to do whatever they were going to do in the first place” (p.151). Against this background, it comes as no surprise that the systematic review by Matthes (2009) identifies a great variety of operationalization and reporting standards in empirical framing research, despite the fact that most of the considered studies are referencing the same conceptual framing literature.

In the realm of content analysis, Matthes and Kohring (2008) suggest there are five different approaches for the identification of frames: hermeneutic approach, linguistic approach, manual holistic approach, computer-assisted approach, and deductive approach. It is important to note that the approach outlined in Matthes and Kohring (2008) is computer-*assisted* approach and the exemplar models are dictionary-based approaches such as Miller (1997).

A relatively new approach is to apply unsupervised machine learning techniques to find frames through induction. As of writing, we are able to find several methods papers suggesting that these unsupervised machine learning techniques can be used to find frames (Burscher et al., 2016; DiMaggio et al., 2013; Eisele et al., 2023; Greussing & Boomgaarden, 2017; Nicholls & Culpepper, 2021; Walter & Ophir, 2019). Not surprisingly, all, except DiMaggio et al. (2013), have given the “obligatory nod” – in Reese (2007)’s sense – to Entman (1993).

Before diving into the methodological rationale behind frame detection with unsupervised machine learning techniques, we revisit what unsupervised machine learning does. Unsupervised machine learning can be divided into three categories: dimensionality reduction, clustering, and density estimation. The first two are focused here because the aforementioned frame detection techniques do not utilize the last approach. Text data represented in the traditional bag-of-word method has a high dimensionality in the feature space. Dimensionality reduction, which the method proposed by Greussing and Boomgaarden (2017) is based on, attempts to reduce the dimensionality in the feature space, yet retains a maximum amount of information from the original data. Clustering analysis, which the methods evaluated or proposed by DiMaggio et al. (2013), Nicholls and Culpepper (2021), Burscher et al. (2016), and Walter and Ophir (2019) are based on, attempts to find groups within high-dimensional data, yet members of the same group have a minimum of variance between each other. These differences notwithstanding, all unsupervised methods attempt to find potentially meaningful clusters of words through either maximizing the information or reducing the internal variance among members of a cluster. All methods do not involve any labeled data, thus these methods are fully automatic and inductive. These methods are referred to as "automatic inductive methods" in the rest of this article. Often, however not always, the clustering solution from these automatic inductive techniques are *posthoc* validated against human-coded data to show that the resulting solution is indeed meaningful. Also, it is noteworthy that these inductive methods originally were not developed to capture frames, but mostly to identify topics within documents. Therefore, topic modeling is an alternative name for the inductive methods which DiMaggio et al. (2013), Nicholls and Culpepper (2021), and Walter and Ophir (2019) are based on. Here, already the question is: How are the patterns that we detect here – clusters of words that commonly occur together – related to intent in the sense we ascribe to the framing process?

The papers referenced above have given different answers. DiMaggio et al. (2013) suggest that "many topics may be viewed as frames (semantic contexts that prime particular associations or interpretations of a phenomenon in a reader)" (p. 578), which might be conflating the two concepts. Other papers argue that the word clusters that result from unsupervised clustering have semantic meanings. Greussing and Boomgaarden (2017) suggest that word clusters "are networks of co-occurring words, constituting the semantic patterns in which words are used, and capturing the underlying structures that provide meaning to a text" (p. 1755). Walter and Ophir (2019) find "several reasons to believe topics could *not* be conceptualized as news frames" (p. 5, emphasis added). They propose a solution to "connect topics into larger themes or meta-topics" (p. 5). In order to do that, they propose their Analysis of Topical Model Networks (ANTMN) approach, a three-step process of topic modeling, network analysis, and community detection. On the question of whether or not the "larger themes or meta-topics" found through their approach can be interpreted as frames, they argue that "the methodological steps ANTMN follows are consistent with the conceptualization of frames as news patterns" (p. 14). But the definition of framing they cite from Entman is news patterns that "repeatedly invoke the same objects and traits, using identical or synonymous words and symbols in a series of similar communications that are concentrated in time", a broader definition that would include patterns that do not promote a specific problem definition and are agnostic towards authorial intent (Walter & Ophir, 2019, p. 261).

There are also authors who do not agree with the interpretation of word clusters as frames. Jacobi et al. (2016), in their guide on LDA in journalism research, caution that word clusters (topics) are not "interpretive packages." Similarly, Guo et al. (2023) maintain that word clusters are not equivalent to frames. Specifically, their criticism to Walter and Ophir (2019)'s approach is that the so-called "frames" identified with the authors' method do not match the required criteria of what constitutes a frame according to the existing media framing literature. Guo et al. (2023) make a distinction between "topic-like frames," e.g., "Safety of nuclear plants" in Burscher et al. (2016), and media frames from a constructive perspective, e.g., generic frames in Semetko and Valkenburg (2000). Guo et al. (2023) argue that inductive methods are only capable of detecting "topic-like frames." Hase et al. (2022) use how the output of these methods being interpreted as an example of "trivialization of theories/concepts"[1] in computational communication science.

The (in)sufficiency of semantic meanings as evidence of frame detection can also be illustrated using the visual metaphor in Figure 1. We can say that case A has a clear semantic meaning (there is a man who does not give the Nazi salute). But the communicative intention remains unknown. Applying Scheufele and Tewksbury (2006)'s "three models of political communication", those "topic-like frames" should rather be subsumed under issues or agendas, not frames. Or more appropriately: subtopics of a news topic. The question we therefore raise is: Can the categories we inductively generate from a text corpus capture the authorial intent to promote a specific problem definition, or do they capture something else?

4. Generic frame detection as a test case and ensuring independence from topics

For this paper we decided to use generic frames as a test case to measure how well different methods can reproduce authorial intent from a given corpus. Generic frames, as defined by De Vreese (2005), are frames that "transcend thematic limitations and can be identified in relation to different topics, some even over time and in different cultural contexts" (p. 54). This has two advantages: First, we can rely on an available definition of the frames to instruct authors to write articles for us that contain a specific generic frame. We therefore have access to the ground truth of authorial intent.

Meanwhile, detection of generic frames is challenging for automated/automatic methods: Nicholls and Culpepper (2021) evaluate different automatic inductive methods and show that STM (Roberts et al., 2014) is capable of detecting frames in a corpus with a narrow scope. But the method extracts topics rather than frames in another corpus with a broad scope. Using the typology by De Vreese (2005), this reflects the distinction between issue-specific frames and generic frames. The finding by Nicholls and Culpepper (2021) effectively bars STM or relative inductive methods from detecting generic frames. As a matter of fact, all examples used in the original methods papers were limited to one news topic (e.g., art in DiMaggio et al., 2013; refugee coverage in Greussing & Boomgaarden, 2017; and financial news in Nicholls & Culpepper, 2021). On the other hand, inductive methods have been applied to multi-topical news corpora before with the goal to identify generic frames (e.g., Walter et al., 2022).

A second advantage is that generic frames, in theory, can be combined with any topic, and we can combine them without introducing correlation. In real-world settings, frames (both issue-specific frames and generic frames) often corre-

[1] Originally in German: Banalisierung von Theorien/Konzepten.

late with topics. For example, Iyengar (1994) observes for episodic versus thematic framing that the episodic frame is applied more frequently in crime stories but not in terrorism (by foreign actors and left-wing perpetrators) stories. The opposite is true when it comes to the thematic frame.

There is no doubt that automatic inductive methods can distinguish crime stories from terrorism stories. However, it is also possible to shoehorn these two topics found in a multi-topical corpus into a reasonably accurate indicator of episodic and thematic frames. Frame detection would therefore be incidental to what the method actually measures. It is like measuring the consumption of chocolate within a country as an indicator of scientific advancement. The indicator might be associational (since chocolate consumption rises with economic prosperity, and economic prosperity correlates with scientific advancement), but not causal. This problem also manifests itself in a single-topic situation, e.g., terrorism, which is usually discussed in media discourse using thematic framing. A single news topic usually has subtopics such as Islamist terrorism, left-wing terrorism, and right-wing terrorism. However, the shoehorned indicator breaks when the *episodic* frame is applied in right-wing terrorism stories ("lone wolf") in Western media (Hase, 2021; Zdjelar & Davies, 2021).

Therefore, for a method to sufficiently detect frames – generic or not – this method should be able to detect frames independent of topics. For example, if a method is proposed to detect episodic and thematic frames, this method should be able to really tell the differences between episodic and thematic frames in both right-wing and Islamist terrorism stories. But it should not be picking up the right-wing terrorism stories and then claim them to be episodic.

5. Our approach: A synthetic dataset

After we have given the "obligatory nod" (Reese, 2007) to framing literature, we propose our approach. In order to test whether a method can reliably detect generic frames that correspond to authorial intent, we need to have a dataset in which the frames and topics are completely independent. This distribution is not natural, but can be generated by randomization. Also, there is no guarantee that manual coding, the so-called "gold standard" of frame detection (Nicholls & Culpepper, 2021), can actually "reverse-engineer" communicative intentions as perceived by an audience member. Therefore, we need to synthesize a dataset where frames (the package of selected aspects of perceived reality and communicative intention) are independent of topics all the way back to the communicative intention. This allows us to test whether our frame induction methods are able to find the intended frames without – for human coders – relying on intuitions about which frames commonly occur within certain topics or – for automatic methods – picking up frames incidentally as a by-product of actually generating distinct topics. Such a synthetic approach has previously been used by Clever et al. (2020) and Frischlich et al. (2023) to solve a similar problem (evaluation of nostalgia detection).

We randomly assigned 100 pairs of topics and frames (Table 1). The topics were "Ukraine," "corona," "tech companies," "climate," and "any topic." For the frames we used the generic frames following Semetko and Valkenburg (2000):

"Attribution of responsibility," "Human interest," "Conflict," "Morality," and "(Economic) consequences." The topics and frames are independent ($\chi^2 = 17.70$, $df = 16$, $p = 0.34$). In the subsequent sections, these are called "ground truth topics" (z) and "ground truth frames" (y) respectively.

Table 1. Distribution of topics and frames

topic	Conflict	Conseq.	Hum. Int.	Morality	Resp.
Climate	6	1	7	2	4
Corona	5	3	4	5	3
Joker	1	5	5	6	3
Tech	4	7	2	3	4
Ukraine	5	6	2	2	5

We gave these ground truth frame-topic pairs to four authors (political science master students with prior knowledge concerning framing theory and generic frames as proposed by Semetko & Valkenburg, 2000) as stimuli and instructed them to write news articles containing the assigned topics and frames. These authors were also randomly paired up to edit the articles written by their peers to ensure the articles were actually conveying the assigned topics or frames. The instruction (in verbatim) given to the four authors regarding the criterion for an article containing a frame when editing articles is as follows: "You would check at least one item of a specific frame (see Semetko & Valkenburg, 2000)." In other words, the four authors wrote and edited the articles with a very specific communicative intention of framing the topics in a specific generic frame such that at least one item of the codebook (Semetko & Valkenburg, 2000) would be checked.

Through this process we generated a multi-topical corpus of 100 news articles with orthogonal ground truth frames and topics. The ground truth of frames contained within these 100 multi-topical news articles are known even without manual coding.

6. An application: A preregistered preliminary analysis of generic frame identification methods[2]

As a use case of the synthetic dataset, we selected different methods and attempted to identify frames in those 100 multi-topical news articles. As there are many methodological limitations in this benchmark, please consider this benchmark as preliminary.

6.1 Hypotheses

We tested three preregistered hypotheses:

> *H1: Compared with manual methods, automatic inductive methods are less accurate in detecting frames.*

[2] The preregistration documents can be consulted here: https://doi.org/10.17605/OSF.IO/SY6JX.

> *H2: Compared with semi-supervised methods, automatic inductive methods are less accurate in detecting frames.*

> *H3: Compared with manual methods, semi-supervised methods are less accurate in detecting frames.*

6.2 The "gold standard"

Two coders (two other political science Master students) were instructed to manually code the 100 articles to find the frame elements of each news item using the codebook by Semetko and Valkenburg (2000) (the complete codebook is available in the Online Appendix: https://osf.io/gkft5/). One item (*"Does the story contain visual information that might generate feelings of outrage, empathy-caring, sympathy, or compassion?"*) was omitted because no images are generated in our synthetic approach. These two coders underwent two rounds of pretesting and training, prior to the actual coding. Despite pretesting and training, the inter-coder reliability between these two coders was still low for some items, based on the test coding of 10 articles (see Online Appendix).

6.3 Exploratory analysis: Expert coding

This part of the analysis has not been pre-registered and was planned after the above "gold standard" coding. After observing surprisingly low correct classification rates for student coders, as an exploratory analysis we studied whether the "gold standard" can be improved by using expert coding instead of the traditional two trained coders (Van Atteveldt et al., 2021). Two experts with PhD in communication were invited to repeat the above manual coding task. In addition, two items were added. The first item "F1" asks the frame of article in an exclusionary manner: *"Overall: The frame of this story is"* with five possible generic frames. This item is called "exclusionary item" because this item assumes a story can only have one frame. The second item "F2" asks the confidence of the answer for "F1": *My level of confidence for F1 is:* with a five-point Likert scale from Very Low to Very High.

As this part of the analysis has not been preregistered, the results were not used to test our preregistered hypotheses. Instead, the expert coding was used to study how experience and knowledge can influence the "gold standard;" also the item "F2" was used to study how the confidence level of the coders can possibly influence the correctness.

A further exploratory study of "ground truth contestion" was conducted to study whether the two experts agree with the ground truth, given their codings. That is, they were asked to assess whether an article matched the generic frame it was intended to convey after this frame was revealed to them.

6.4 Automatic inductive methods

All automatic methods that have been claimed of being able to induce frames were investigated. This includes k-Means with TF-IDF (Burscher et al., 2016), Principal Component Analysis with TF-IDF (Greussing & Boomgaarden, 2017), LDA (DiMaggio et al., 2013; evaluated by Eisele et al., 2023), STM (Nicholls & Culpepper, 2021), and Topic Model Networks (Walter & Ophir, 2019). The number of clusters to find (*k*) was five. See Online Appendix for an overview of all included methods.

6.5 Semi-supervised methods

We also investigated semi-supervised methods. This consists of Seeded-LDA (Watanabe & Zhou, 2020) and Keyword Assisted Topic Model (keyATM, Eshima et al., 2020). It is important to clarify that the authors of these methods do not claim that their semi-supervised methods can be used for detecting frames. But both methods are claimed to be able to measure theoretical constructs through the provision of theory-driven dictionaries. Eisele et al. (2023) applied KeyATM to detect frames but found it unfit for the task. Therefore, semi-supervised methods were included in this study for exploratory purposes only. To apply these methods, we needed dictionaries that should be able to find the five generic frames. Before data collection, we surveyed two experts of journalism studies and pre-registered the dictionaries they suggested.[3]

6.6 Evaluation: Multiverse analysis

For automatic inductive methods, many methodological decisions need to be made: There are different ways to preprocess the text data (Maier et al., 2018). Even for the "gold standard," there are many methods to combine the frame elements into frames, despite the standard codebook (e.g., averaging by Dirikx & Gelders, 2010; factor analysis by d'Haenens & Lange, 2001; binary categorization by Kroon et al., 2022). We preregistered all possible combinations of analytical steps and benchmark these methods with all possible combinations of methodological decisions using multiverse analysis (Pipal et al., 2023; Steegen et al., 2016). For example, STM was applied using all combinations of possible preprocessing steps: 1) stemming vs lemmatization vs no processing, 2) removal of stopwords or not, 3) removal of sparse and dense words or not, 4) different levels of α.

[3] With the authors of the articles and the human coders working deductively – based on predefined categories derived from theory – and computational methods working inductively, human coders should already be at an advantage, of course: They know what characteristics of the text to look for and know what frames to distinguish. Meanwhile, inductive methods generate the categories based on common patterns within the text corpus – so they have the disadvantage that they could find other, albeit also meaningful, patterns in the text. This of course opens up additional methodological questions (e.g.: What would be the ground truth for a category that is derived inductively from the text?). For this paper, we confine ourselves to the more limited question to answer "Do our methods find the same frames that the authors consciously put into the text?" and will bracket the more sophisticated question "If inductive methods find something else, is that something else also meaningful?"

The original developers of Topic Model Networks have provided a careful justification of text preprocessing and model parameters (e.g., Walter & Ophir, 2019). For it, we highlight the prescribed preprocessing suggestions in the multiverse analysis.

6.7 "Best-case" correct classification rate

To assess the accuracy of each method, we used the correct classification rate (CCR). Let y be the ground truth frame vector and $\hat{y}$ to be the output from a method. $\hat{y}$ is a vector of frame indicators, f_{k} where $k = 1,2,...,5$. CCR is defined as $CCR = Pr(y = \hat{y})$. However, there is no way to tell which frame indicator f_{k} corresponds to which actual frame in the ground truth y. The usual practice is for a human rater to evaluate the topic words or visualization such as LDAvis (Sievert & Shirley, 2014) and map f_{k} to the specific frame in y accordingly (Maier et al., 2018). Several of these automatic inductive methods suggest human intervention at this stage. Walter and Ophir (2019)'s method, marketed as an "inductive mixed-method," also has this mapping of detected clusters from multi-topical news content to the generic frames (Walter et al., 2022). There have been concerns about the validity of this approach (e.g., Chan & Sältzer, 2020) and we do not want the variation in these manual mapping decisions to influence our benchmark results.

Inspired by the calculation of best case complexity in the analysis of algorithms, we calculated what we called the "best case" CCR (CCR_{max}) using exhaustive search. In this analysis, we generate all possible permutations of all possible values of k, i.e. $\epsilon f_1,f_2,...,f_{\mathrm{k}}$. For $k = 5$, there are 120 possible permutations. For each of these 120 possible permutations, we calculated the CCR. From these 120 possible values, we selected the highest value, i.e. CCR_{max} , to represent the best-case scenario. This analysis is "unrealistic" in the sense that the ground truth is never known in real life. But this "best-case" analysis ensures that the real-life performance of these methods is equal to or in most cases worse than the CCR_{max} reported, but never better. Therefore, the findings from this paper cannot be defended by the lack of human interpretation or any intervention. We have assumed that there were a *divinus* who can always perform the best in this mapping task.

The null value for CCR_{max} is 0.2, when $k = 5$ (Krippendorff, 2011). It is also possible to calculate the same null CCR_{max} value when a method can perfectly tell topics and then shoehorn those topics into frames. This expected CCR_{max} value should also be 0.2 theoretically, when frames and topics are randomly assigned and the sample size is large. But due to the small sample size and idiosyncrasy of randomness, the *de facto* null value of CCR_{max} is 0.3 in our 100 frame-topic pairs. This value was found using the same exhaustive search technique by using the ground truth topics z as $\hat{y}$ to map into the ground truth y.[4] In the figures below, we indicate both null values. One can

[4] One can also look at Table 1 and use the so-called Greedy Algorithm to derive the de facto CCR_{max}. We first go after the largest numbers in the table: there are two cells with seven and use that as the initial mapping (mapping "Climate" to "Human Interest;" and "Tech Companies" to "(Economic) Consequences"); we left with three topics: ("Joker," "Ukraine," and "Corona") and three frames ("Conflict," "Morality," "Responsibility"). We repeat the mapping of the largest numbers until all frames and topics are mapped. In the end, we have a complete mapping and the numbers are 7,7,6,5,5. Their sum is 30 and dividing it by the total number of articles 100 gives 0.30.

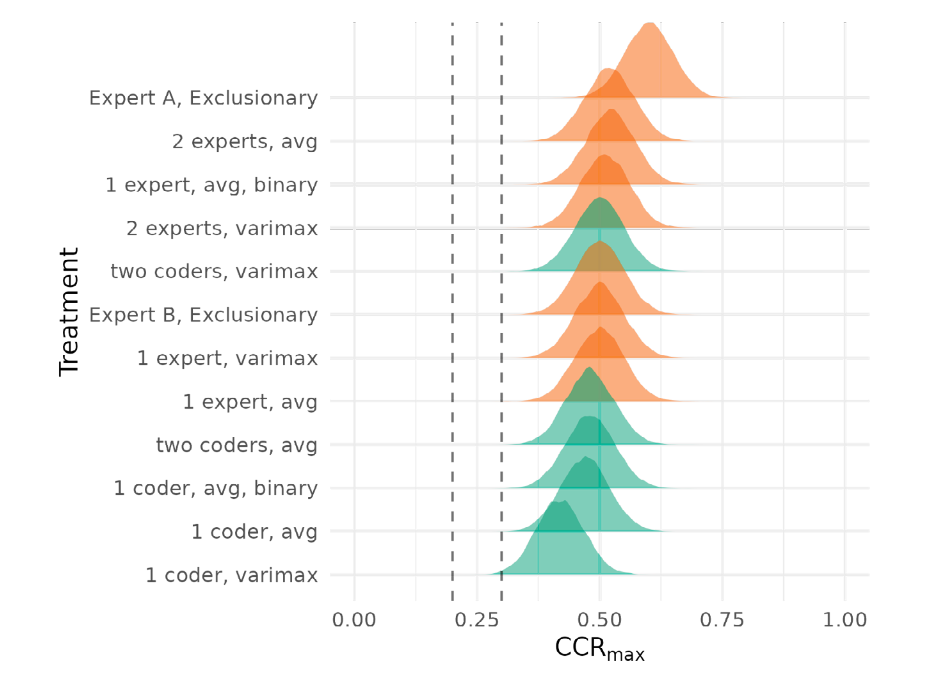
think about the two null values as “can tell neither topics nor frames from the data” and “can tell topics but not frames from the data” respectively.

7. Results

7.1 “Gold Standard”

Figure 2 shows the result of multiverse analysis of the “Gold Standard.” In the ridgeline figure (Wilke, 2022), the entire distribution of CCR_{max} (based on binomial distribution) is shown. The multiverse analysis suggests that the “Gold Standard” can detect frames in the multi-topical news content better than null in almost all situations, regardless of methodological decisions, as indicated by the extremely little overlapping between the distributions of CCR_{max} and the two null values. However, the performance is not as high as one would expect. The best of the best value of CCR_{max} is only around .5.

Figure 2. Multiverse analysis of the “Gold Standard” (Orange: Not pre-registered, coding from expert(s); dotted lines: nulls)



This analysis also reveals that there is not enough evidence to show that using experts instead of trained coders increases the accuracy. The same can be said about coding frame elements and coding frame as a single item. In the Online Appendix, a comparison of confidence level of the expert coders between correct and incorrect answers is presented. Experts could give incorrect answers confidently.

7.2 Analysis of incorrect answers from two experts

Using the exclusionary item and the ground truth frames, we conducted a non-preregistered exploratory analysis of inaccurate answers from the two experts. Table 2 shows the *CCR* values across all ground truth frame categories. Both experts were relatively good at identifying the “economic consequences” frame. However, the two experts were relatively worse at identifying the morality frame.

Table 2. Correct classification rates of two experts by ground truth frames

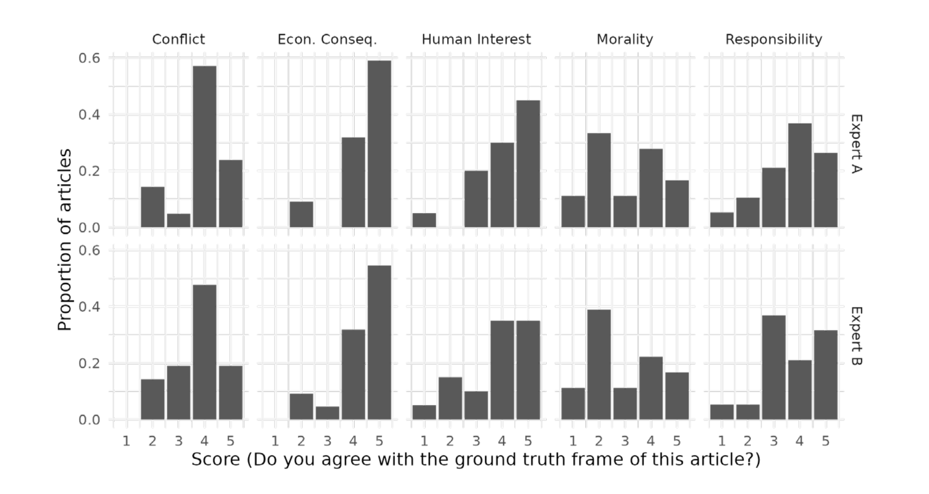
Frame (ground truth)	Overall	Expert A	Expert B
Morality	0.36	0.28	0.44
Responsibility	0.50	0.53	0.47
Conflict	0.55	0.67	0.43
Human interest	0.57	0.65	0.50
Economic consequences	0.73	0.82	0.64

In a *posthoc* manner, we also attempted to explain the disagreement between the ground truth and the expert judgment by asking the experts to comment on whether or not the ground truth is justified in the synthetic corpus. The two experts were given γ and their $\hat{y}$ and were asked whether they agree with the ground truth using a 5-point Likert score (1 = Strongly disagree, 5 = Strongly agree). They were also asked to provide comments about the articles in an open-ended question.

Figure 3 displays distributions of scores from the two experts. In general, the two experts agree with the ground truth of the economic consequences, conflict, and human interest frames, but not the morality and responsibility frame. The modal values from the two experts for morality articles tend towards disagreement in general. Expert B’s modal value for responsibility articles also indicates disagreement.

In open-ended responses, the two experts expressed two main concerns: (1) For articles with the ground truth morality frame, they found the moral message – “religious tenets or moral prescription” in the original definition by Semetko and Valkenburg (2000) – is unclear; and (2) they found elements of multiple frames in one article.

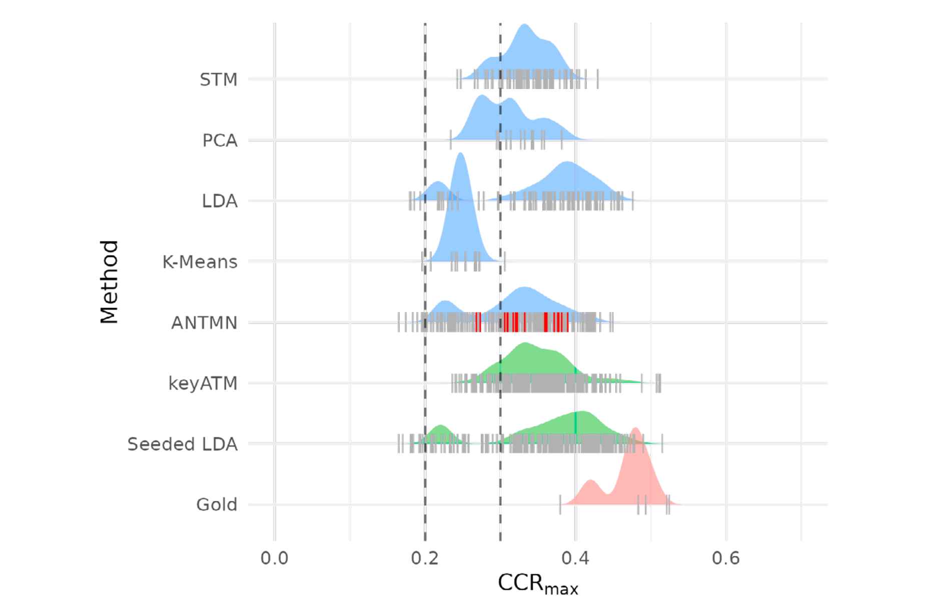
Figure 3. Distribution of ground truth agreement scores from two experts



7.3 Preliminary comparison of methods

Individual results from the multiverse analysis for each method are available in the Online Appendix. Figure 4 displays the overall results of the multiverse analysis. The results from different methodological decisions are combined as density masses.

Figure 4. Distribution of best-case correct classification rates by methods (dotted lines: nulls)



K-means, PCA, STM have no to extremely little overlap in density mass with human coding. ANTMN, keyATM, LDA, and Seeded LDA (in that order) have a relatively higher overlap in density mass with human coding.

By considering the entire distribution, H1 (automatic methods perform worse than human coding) appears to be supported, although LDA and ANTMN might be comparatively better with some methodological decisions. There is not enough evidence to support H2 (automatic methods perform worse than semi-supervised methods) and H3 (semi-supervised methods perform worse than human coding).

8. Discussion

Based on our review of the framing literature, we provide a synthetic benchmark dataset for detection of generic frames where frames and topics are independent. As an application of this synthetic dataset, we evaluated the "gold standard" (Semetko & Valkenburg, 2000), fully-automated clustering methods that are claimed to be able to detect frames inductively (Burscher et al., 2016; Greussing & Boomgaarden, 2017; Nicholls & Culpepper, 2021; Walter & Ophir, 2019), and two semi-supervised methods (Eshima et al., 2020; Watanabe & Zhou, 2020).

Using the multiverse analytic approach (Pipal et al., 2023; Steegen et al., 2016), we exhaustively studied all methods, irrespective of methodological choices, and reported their "best-case" performance.

In the following paragraphs, we will summarize the surprising findings of our analyses, particularly regarding the human "gold standard" coding. Then we discuss possible reasons for our results, first on the level of article generation, second concerning measurement, then we discuss potential conceptual issues. Lastly, we note limitations of our current approach and give suggestions for further research.

For the first time, the current study is able to benchmark the so-called "gold standard:" manual content analysis for frame identification. Although the "gold standard" performs significantly better than null, the performance is not superb (Figure 2). There have been concerns about the reliability of Semetko and Valkenburg (2000)'s codebook (e.g., Kroon et al., 2022). But the less-than-superb performance cannot be explained by intercoder variations alone because the multiverse analysis has considered both the single- and double-coder scenarios. This low performance is not what we expected and also thought-provoking: Given the fact that the "gold standard" can only detect 50% of frames correctly and the state-of-the-art supervised classifiers classify frames at around 60% accuracy, should we trust the supervised frame classifiers trained on the so-called "gold standard" data (e.g., Kroon et al., 2022; Eisele et al., 2023; Kwak et al., 2020; Liu et al., 2019)? We do not have an empirical answer to this question, because we did not study supervised methods in this paper.

Our findings do point to one important fact: Identification of (generic) frames from news content written by someone else is an incredibly difficult task, even for experts (Figure 2). This task is so complex because we need to evaluate the semantics of the selected reality as well as the communicative intention of a third party. Prediction of a communicator's intention from their written text is similar to the goal of detecting a communicator's psychological state by counting their usage of pronouns (Tausczik & Pennebaker, 2009). Individuals tend to assume they can tell someone else's communicative intention (or someone else's psychological state), and therefore no validation seems necessary. But when this assumption is crosschecked (as in the present study), results indicate that the assumption that we can tell someone else's communicative intention might very well be biased. Even experts can confidently make incorrect guesses (see the Online Appendix).

That being said, it is our assumption that the synthetic dataset itself captures the four authors' communicative intentions. Our approach has an important weakness because it creates a "turtles all the way down" situation of who can tell whether the four authors' communicative intentions have been adequately expressed in the content they produced.

As an exploratory analysis, we went down one layer of turtles and allowed the two experts to contest the ground truth. From this ground truth contestation exercise, we learned several lessons about the detection of generic frames and the notion of generic frames in general which could explain our results.

We found that the original conceptualization of generic frames by Semetko and Valkenburg (2000) contributes to the difficulty of the whole endeavor of creating articles with a specific frame and finding one frame in a specific article. First, our experts heavily contested articles with the ground truth morality frame. This comes as no surprise as the detection of the morality frame also has the worst performance. The extraction of latent moral information from text is a highly subjective task (Weber et al., 2018) to begin with and this subjectivity contributes not only to the low reliability of traditional content analytic approaches, but also hampers the ability of our coders and experts to determine whether or not any moral message is conveyed (an item in the original codebook). Two experts mainly found that the moral message either does not exist or is implicitly communicated in many of the morality-aimed texts within our corpus. This implicit aspect of morality framing has been expressed in the definition by Semetko and Valkenburg (2000), which assumes journalists usually communicate a moral message indirectly (such as raising questions, see Neuman et al., 1992) due to the professional norm of objectivity.[5] An example given by Semetko and Valkenburg (2000) is using the view of an interest group to raise questions about sexually transmitted diseases (rather than what one should do to not get sexually transmitted diseases). In the dataset, there are similar cases where the student authors simply raised questions about morally charged topics, e.g., artificial intelligence, while the moral prescriptions are not clearly communicated. Should we count that as a morality frame? This question remains unanswered in the literature.

[5] Another weakness of this paper is the fact that the communicators employed for this study are not professional journalists. Journalists might have a better ability to communicate their intentions than our four student authors. Therefore, future research will need to try to replicate this study with journalists. However, as the communicators (student authors) in this study were instructed to write as if they were journalists, the indirectness in moral message expression should also apply to them.

Second, the two experts found that an article can actually contain frame elements of multiple frames. The original operationalization by Semetko and Valkenburg (2000) actually allows that as their frame measurement is not one score but five scores. In empirical research, however, all the frame elements considered are usually consolidated into just one per news item $\hat{y}$ (e.g., Dirikx & Gelders, 2010; d'Haenens & Lange, 2001; Kroon et al., 2022). The ramification for both manual and automatic coding tasks was that the method to resolve the dominant frame also contributed to accuracy. One can foresee that the performance of any detection method under consideration would be much better if a generic frame is communicated exclusively, without any use of frame elements that belong to one of the other four generic frames.

We addressed these two issues raised by the two experts with two sensitivity analyses: (1) We evaluated the performance of all methods without all morality articles, and (2) we evaluated the performance of all methods but using a performance metric that does not assume a dominant frame in each article. These two sensitivity analyses are provided in the Online Appendix. H1 appears to be less supported when all morality articles are removed. But as with any measurement problem, without further research, it is hard to delineate where the problem originates – whether the student authors did not communicate the frame clearly, whether coders did not pick them up properly, or, more worryingly, whether there is a conceptual problem with generic frames at the root. We observe that, for example, Burscher, Odijk, Vliegenthart, Rijke, and De Vreese (2014) report low intercoder reliability for individual frame elements in real-world news data, and note that while in framing theory (e.g., Matthes, 2014, p. 15) the assumption is that communicators' frames feed into the news frames in the texts they write, we rarely validate measured frames against authorial intent.

This opens up two conceptual questions that can be raised if measurement problems like the ones we observe persist: a) We can ask whether generic frames are indeed too generic. As the expert coders observed, overlap between frames did happen when the student authors incidentally used frame elements of another frame – for example, responsibility attribution could be such a normal part of human communication that it becomes difficult to assess when if happens coincidentally (for example in the lead into a conflict story) or whether it is authorial intent to *highlight* that part of a story. b) We can ask how stringent the assumption of authorial intent in framing is – or whether a constructivist view of communication which locates the meaning of text in the interpretation of the reader is more appropriate (Luhmann, 1997, p. 72). That way, the "ground truth" would not be the infinite regress of author's intent, but a quality of audience reception. In that case, though, we also have to relax the assumption of framing as a linear process from journalist to audience.

Answering these questions is beyond the scope of this paper. Additional research will have to assess whether methodological or theoretical adjustments are necessary, or whether our corpus included too many outlier articles. We do hope that future research will adapt the construction of synthetic datasets with available ground truth to help find answers or further conceptualize where we expect the ground truth to lie.

9. Limitations

One limitation of the paper is that by focusing on generic frames, we have consciously chosen a specifically tough case for automatic (and, it appears: human) methods of frame detection. We suggest further research with issue-specific frames, following a similar setup to ours: Problem definitions would have to be predefined and assigned orthogonally to sub-topics to test if inductive methods can detect authorial intent.

Another limitation is the sample size of 100. This limited sample size can have two different implications: (1) whether we have enough variety in articles (e.g., variations in vocabulary, stylistic clues, angles) to use any of the automatic or semi-supervised methods and (2) statistical power of the analysis and/or whether we have enough articles to use any of the automatic or semi-supervised methods. Statistically speaking, increasing the sample size does not always increase variety (if variety means variance, increasing the sample size tends to decrease the variance). In our opinion, a more reliable way to increase the variety of articles is not just to increase the sample size but also to increase the content categories (e.g., topics) or/and authors. With our current data, it is not possible to simulate the possible effect of increasing variety and this warrants further studies.

For the second implication, this study has no say about the equivalence among methods and null. We can only check our superiority hypotheses (H1, H2, and H3). We refrained from concluding that a method is equivalent to null. The kind of equivalence conclusions can only be drawn with a different study design (see a primer by Weber & Popova, 2012). Suppose one would need to test the equivalence hypothesis, the required sample size would be 13,708 (null value of 20%; equivalence limit of 2%; α: 0.05; β: 0.2). To give a perspective to this sample size, the New York Times publishes around 230 new articles per day. The cost to produce this number of articles using our synthetic approach would be equivalent to asking journalists to write two-month worth of news content.

Another issue with the sample size is that automatic and semi-supervised methods studied in this benchmark might not work well with a sample size of 100, as these methods were not designed to work with this relatively small sample size. In the Online Appendix, an analysis is presented to simulate the possible impact of increasing the sample size on the three hypotheses. Our simulation points to the direction that both H1 and H2 are more likely to be supported when the sample size increases. H3, however, might be less likely to be supported. Therefore, automatic methods are more likely to detect *topics* rather than generic frames when the sample size increases. But we maintain that this finding needs to be confirmed with actual data in a new empirical study.

Finally, an important drawback of our approach of calculating CCR_{max} is that the search space of possible permutations grows factorially, i.e. $k!$. Therefore, the search space is unbearably large when k is just slightly larger than 5 (e.g., 10!=3628800). It also means that we cannot increase k from 5 to allow for the so-called boilerplate topics (Maier et al., 2018) or providing flexibility in frame mapping (Nicholls & Culpepper, 2021; Walter & Ophir, 2019). A smarter heuristic algorithm for the calculation of CCR_{max} is needed.

10. Concluding remarks

By synthesizing a dataset for the validation of the measurement of generic frames and benchmarking manual and automatic methods with the dataset, we came across problems measuring these frames that, we hope, ignite further discussion on the methodological and theoretical assumptions underlying framing research. Even though the present endeavor might have raised awareness for more issues than it could help resolve, it still provides a number of insights both for framing research and for validity testing in content analysis more broadly.

Using the example of generic frames, we have demonstrated that in a scenario where different media content features are systematically correlated (in our case, frames and topics) it might be better to work with synthetic corpora that randomize and, thereby, decorrelate, these features to assess the actual validity of detection methods. Our study has also shown that this solution, despite ruling out a number of problems that working with a sample of real-life data has, still comes along with specific issues that are inherent to human content production (but might, likewise, also occur with texts generated by large language models). In our case, this refers to the occurrence of several generic frame elements within one text as well as the journalistic tendency to express morality framing in a way which might be too subtle to be reliably detected in human or machine coding. These shortcomings notwithstanding we believe that the general approach of constructing randomized, synthetic news corpora should be followed upon in future validation studies. Researchers in the field of frame detection should be aware of the limitations addressed here when working with or building upon the corpus published alongside this article.

For framing research in particular, the present findings pose a number of important questions that future research will have to address: Is it fair to assume that individual news items express one dominant or even exclusionary (generic) frame? Is the concept of generic frames actually too generic to be validly detected in content analysis? How can a threshold level for explicitness be defined that allows to assign a specific frame label to a text (considering that Morality frame seems to be expressed mainly implicitly)? And, finally: What are the specific strengths and weaknesses of studying a deductive concept such as frames by using inductive methods such as fully automated clustering?

Author note

The link to the online appendix of this submission: https://osf.io/gkft5/. Preregistration: https://doi.org/10.17605/OSF.IO/SY6JX. Code and data: https://github.com/chainsawriot/frame. The authors would like to thank 1) Valerie Hase (Ludwig-Maximilians-Universität München) for providing keywords to be used in this study, 2) Open Science Office, University of Mannheim, for the financial support, 3) Sandra Krais for language editing, and 4) student assistants: Filippo Borsato, Hannah Erb, Hyosun Jang, Fatih Ozhasar, Zeynep Özgülec, and Jonathan Vincent.

Correspondence concerning this article should be addressed to Chung-hong Chan, Unter Sachsenhausen 6-8, 50667 Cologne, Germany. Contact: chung-hong.chan@gesis.org

References

Baden, C. (2015). *INFOCORE definitions: "Interpretative frame"*. INFOCORE. https://www.infocore.eu/wp-content/uploads/2016/02/def_interpretative_frame.pdf

Boumans, J. W., & Trilling, D. (2015). Taking stock of the toolkit. *Digital Journalism*, *4*(1), 8–23. https://doi.org/10.1080/21670811.2015.1096598

Burscher, B., Odijk, D., Vliegenthart, R., Rijke, M. de, & 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., & De Vreese, C. H. (2016). Frames beyond words. *Social Science Computer Review*, *34*(5), 530–545. https://doi.org/10.1177/0894439315596385

Carragee, K. M., & Roefs, W. (2004). The neglect of power in recent framing research. *Journal of Communication*, *54*(2), 214–233. https://doi.org/10.1111/j.1460-2466.2004.tb02625.x

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

Clever, L., Frischlich, L., Trautmann, H., & Grimme, C. (2020). Automated detection of nostalgic text in the context of societal pessimism. *Lecture Notes in Computer Science*, 48–58. https://doi.org/10.1007/978-3-030-39627-5_5

D'Angelo, P. (2002). News framing as a multiparadigmatic research program: A response to Entman. *Journal of Communication*, *52*(4), 870–888. https://doi.org/10.1111/j.1460-2466.2002.tb02578.x

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

d'Haenens, L., & Lange, M. de. (2001). Framing of asylum seekers in Dutch regional newspapers. *Media, Culture & Society*, *23*(6), 847–860. https://doi.org/10.1177/016344301023006009

DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics*, *41*(6), 570–606. https://doi.org/10.1016/j.poetic.2013.08.004

Dirikx, A., & Gelders, D. (2010). To frame is to explain: A deductive frame-analysis of Dutch and French climate change coverage during the annual UN Conferences of the Parties. *Public Understanding of Science*, *19*(6), 732–742. https://doi.org/10.1177/0963662509352044

Eisele, O., Heidenreich, T., Litvyak, O., & Boomgaarden, H. G. (2023). Capturing a news frame – comparing machine-learning approaches to frame analysis with different degrees of supervision. *Communication Methods and Measures*, *17*(3), 205–226. https://doi.org/10.1080/19312458.2023.2230560

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. *arXiv Preprint arXiv:2004.05964*.

Frischlich, L., Clever, L., Wulf, T., Wildschut, T., & Sedikides, C. (2023). Populists' reliance on nostalgia: A supervised machine learning approach. *International Journal of Communication*, *17,* 2113–2137.

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

Guo, L., Su, C., Paik, S., Bhatia, V., Akavoor, V. P., Gao, G., Betke, M, & Wijaya, D. (2023). Proposing an open-sourced tool for computational framing analysis of multilingual data. *Digital Journalism*, *11*(2), 296–297. https://doi.org/10.1080/21670811.2022.2031241

Hase, V. (2021). What is terrorism (according to the news)? How the German press selectively labels political violence as “terrorism”. *Journalism*, 146488492110170. https://doi.org/10.1177/14648849211017003

Hase, V., Mahl, D., & Schäfer, M. S. (2022). Der „Computational Turn“: ein „interdisziplinärer Turn“? Ein systematischer Überblick zur Nutzung der automatisierten Inhaltsanalyse in der Journalismusforschung [The ‚computational turn‘: An ‚interdisciplinary turn‘? A systematic overview of the use of automated content analysis within journalism research]. *Medien & Kommunikationswissenschaft*, *70*(1–2), 60–78. https://doi.org/10.5771/1615-634x-2022-1-2-60

Iyengar, S. (1994). *Is anyone responsible?: How television frames political issues*. University of Chicago Press.

Jacobi, C., Atteveldt, W. van, & 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

Krippendorff, K. (2011). Agreement and information in the reliability of coding. *Communication Methods and Measures*, *5*(2), 93–112. https://doi.org/10.1080/19312458.2011.568376

Kroon, A. C., Meer, T. van der, & Vliegenthart, R. (2022). Beyond counting words. *Computational Communication Research*, *4*(2), 528–570. https://doi.org/10.5117/ccr2022.2.006.kroo

Kwak, H., An, J., & Ahn, Y.-Y. (2020). A systematic media frame analysis of 1.5 million New York Times articles from 2000 to 2017. *12th Acm Conference on Web Science*, 305–314.

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 US gun violence. *Proceedings of the 23rd Conference on Computational Natural Language Learning (Conll)*, 504–514.

Luhmann, N. (1997). *Die Gesellschaft der Gesellschaft* [The society of the society] (Vol. 2). Suhrkamp Frankfurt am Main.

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer G., Reber, U., Haussler, 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. (2014). *Framing*. Nomos. https://doi.org/10.5771/9783845260259

Matthes, J., & Kohring, M. (2008). 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

Miller, M. M. (1997). Frame mapping and analysis of news coverage of contentious issues. *Social Science Computer Review*, *15*(4), 367–378. https://doi.org/10.1177/089443939701500403

Neuman, W. R., Just, M. R., & Crigler, A. N. (1992). *Common knowledge: News and the construction of political meaning*. University of Chicago Press.

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

Pipal, C., Song, H., & Boomgaarden, H. G. (2023). If you have choices, why not choose (and share) all of them? A multiverse approach to understanding news engagement on social media. *Digital Journalism*, *11*(2), 255–275. https://doi.org/10.1080/21670811.2022.2036623

Reese, S. D. (2007). The framing project: A bridging model for media research revisited. *Journal of Communication*, *57*(1), 148–154. https://doi.org/10.1111/j.1460-2466.2006.00334.x

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, *58*(4), 1064–1082. https://doi.org/10.1111/ajps.12103

Scheufele, D. A., & Tewksbury, D. (2006). Framing, agenda setting, and priming: The evolution of three media effects models. *Journal of Communication*, *57*(1), 9–20. https://doi.org/10.1111/j.0021-9916.2007.00326.x

Semetko, H. A., & Valkenburg, P. M. V. (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

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. https://doi.org/10.3115/v1/w14-3110

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Tausczik, Y. R., & Pennebaker, J. W. (2009). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927x09351676

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458. https://doi.org/10.1126/science.7455683

Van Atteveldt, W., Van der Velden, M. A. C. G., & Boukes, M. (2021). The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures*, *15*(2), 121–140. https://doi.org/10.1080/19312458.2020.1869198

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

Walter, D., Ophir, Y., Pruden, M., & Golan, G. (2022). Watching the whole world: The media framing of foreign countries in US news and its antecedents. *Journalism Studies*, *23*(15), 1994–2014. https://doi.org/10.1080/1461670x.2022.2137838

Watanabe, K., & Zhou, Y. (2020). Theory-driven analysis of large corpora: Semisupervised topic classification of the UN speeches. *Social Science Computer Review*, *40*(2), 346–366. https://doi.org/10.1177/0894439320907027

Weber, R., Mangus, J. M., Huskey, R., Hopp, F. R., Amir, O., Swanson, R., Gordon, A., Khooshabeh, P., Hahn, L., & Tamborini, R. (2018). Extracting latent moral information from text narratives: Relevance, challenges, and solutions. *Communication Methods and Measures*, *12*(2–3), 119–139. https://doi.org/10.1080/19312458.2018.1447656

Weber, R., & Popova, L. (2012). Testing equivalence in communication research: Theory and application. *Communication Methods and Measures*, *6*(3), 190–213. https://doi.org/10.1080/19312458.2012.703834

Wilke, C. O. (2022). *ggridges: Ridgeline plots in 'ggplot2'*. https://wilkelab.org/ggridges/

Zdjelar, V., & Davies, G. (2021). Let's not put a label on it: Right-wing terrorism in the news. *Critical Studies on Terrorism*, *14*(3), 291–311. https://doi.org/10.1080/17539153.2021.1932298