A benchmark dataset for detecting frames in multi-topical news content

Abstract

Frames are a central concept in communication research. Based on our literature review, we propose that frame detection is an act of detecting selected reality and communicative intention. We then highlight the conceptual and methodological issues of frame detection using computational methods. Due to the correlation between topics and frames, we provide a synthetic benchmark dataset for evaluating frames found in multi-topical news content. For the first time, the current study is able to benchmark manual coding and various automatic and semi-supervised methods using this synthetic benchmark dataset. Based on the benchmark results, this study casts doubt on the validity of frame detection using automatic inductive methods such as Structural Topic Models (STM) and Analysis of Topic Model Networks (ANTMN).

Keywords: frame, unsupervised method, topic model, semi-supervised method, validity

Word count: 5876

A benchmark dataset for detecting frames in multi-topical news content

Introduction

The goal of this study is to provide a benchmark dataset for evaluating frames found through any method (e.g. Latent Dirichlet Allocation or Structural Topic Models). In order to achieve this goal, we first review the concept of frames in communication research and then highlight the conceptual and methodological issues of the notion of automatic detection of frames. Then we outline our approach to produce such a dataset, analyze the dataset, and report the preliminary results.

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).

Even before the onset of the so-called "Computational Turn" of journalism research (Hase, Mahl, & Schäfer, 2022) and the notion of automated content analysis (Boumans & Trilling, 2015), detection of frames has 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." (pp.52, emphasis added) Purely from this definition, the part on "to promote" (in active voice) indicates that the act of framing in Entman's sense is from the perspective of the communicators.

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." (pp.12) The underlying communicative intention is then to "resonate with existing underlying schemes among their audience." Similarly, Baden (2015)'s definition 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 example, can pick up the economic frame on climate change and then frame their reporting in terms of the costs of mitigating climate change, without being aware 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.

With the tacit aspect of **communicative intention** making explicit, it raises several questions about frame detection. The most obvious question is: What exactly is detecting frames? Should it be judging which aspects of a perceived reality have been selected 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).

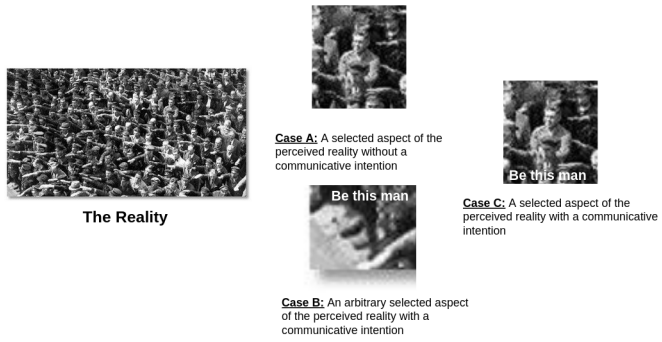

Figure 1. A metaphor for frame detection

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 a "frame" (Case A), it might resonate with the audience but with an ambiguous communicative intention: a *Gestapo* officer might select this aspect with a 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 know the communicative intention of the communicator but this communicator selects an arbitrary aspect to convey their intention (Case B). It can't

be a "frame" either because the selected aspect cannot 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).

The chaotic state of detecting 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." (pp.217) A more harsh 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." (pp.151) There is no wonder 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 them are referencing the same 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. The major focus of this paper is the computational approaches. 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 generation of computational approach is to apply unsupervised machine learning techniques to inductively detect frames. As of writing, we are able to find several methods papers suggested that these unsupervised machine learning techniques can be used to detect frames (Burscher, Vliegenthart, & de Vreese, 2016; DiMaggio, Nag, & Blei, 2013; Greussing & Boomgaarden, 2017; Nicholls & Culpepper, 2020; 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 justification of these unsupervised machine learning techniques for detecting frames, it would be a better idea to revisit what unsupervised machine learning techniques do. Unsupervised machine learning techniques 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 retain the maximum information from the original data. Clustering analysis, which the methods proposed by DiMaggio et al. (2013), Nicholls and Culpepper (2020), Burscher et al. (2016), and Walter and Ophir (2019) are based on, attempts to find groups within the high-dimensional data, yet those members of the same group has the minimum variance among them. Yet, all unsupervised methods attempt to find potentially meaningful clusters of words through either maximizing the information or reducing the internal variance among members. All methods do not involve any labeled data, thus these methods are fully inductive. Although not always, the findings from these unsupervised techniques are *posthoc* validated against human-coded data to show that those clusters are indeed meaningful. Also, these inductive methods originally weren't developed to capture frames, but mostly to identify topics within documents. Topic modeling is the alternative name for the unsupervised methods which DiMaggio et al. (2013), Nicholls and Culpepper (2020), and Walter and Ophir (2019) are based on.

But in what way can these word clusters found by unsupervised machine learning techniques be used to detect frames? These papers have given different justifications. 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)" (pp.578), which might be conflating the two concepts. Other papers justify it by

saying those word clusters 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." (pp.1755) Similarly, Walter and Ophir (2019) claim that their community detection of topical network approach (Analysis of topic model networks, ANTMN) maps closer to Entman (1993)'s conceptualization of frame. But still, the argument is based on the clustering of semantic meaning of various topics.

With these papers showing the perceived validity of these unsupervised methods for detecting frames, these approaches are now extremely popular in empirical journalism research. But there are also authors who do not agree with this trend. Jacobi, Atteveldt, and Welbers (2016), in their guide on LDA in journalism research, caution that word clusters (topics) are not "interpretive packages". Similarly, Guo et al. (2022) 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 approach do not match what are expected from the existing media framing literature. Guo et al. (2022) 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. (2022) argue that inductive methods are only capable of detecting "topic-like frames". Hase et al. (2022) use how the output from these unsupervised methods are 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 thought through using the visual metaphor in Figure 1. We can say that the case A has a clear semantic meaning (there is a man who doesn't give the Nazi salute). But the communicative intention remains unknown. Applying the Scheufele and Tewksbury

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

(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 need for separating topics from frames

Nicholls and Culpepper (2020) evaluate different 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 is the distinction between Issue-specific frames and Generic Frames. 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." (pp.54) The finding by Nicholls and Culpepper (2020) effectively bars STM or relative inductive methods from detecting generic frames. But nonetheless, these methods have still been applied in multi-topical news content to identify generic frames (e.g. Walter, Ophir, Pruden, & Golan, 2022).

However, the methodological issue is even more complicated because these issue-transcending generic frames are usually associated with news topics in real-life news coverage. One purpose for studying these generic frames is to study the distribution of these generic frames across news topics. Iyengar (1994), for example, proposes his binary classification of framing (episodic versus thematic) across different political issues. 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 possible to shoehorn in these two topics found in a multi-topical corpus as a reasonably accurate indicator of episodic and thematic frames. 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. It is because a single news topic usually has subtopics, e.g. Islamist terrorism, left-wing terrorism, and right-wing terrorism. For example, the shoehorned-in 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 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.

Our approach: a "platinum standard" benchmark dataset

After we have given the "obligatory nod" (Reese, 2007) to the framing literature, we propose our approach. In order to test whether a method can sufficiently detect frames, we need to have a dataset where the frames and topics are completely independent. But this dataset does not exist in reality. Also, there is no guarantee that manual coding, the so-called "gold standard" of frame detection (Nicholls & Culpepper, 2020), can actually "reverse-engineer" communicative intentions as an audience member. Therefore, we need to synthesize a counterfactual 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. The synthetic approach has been used by Clever, Frischlich, Trautmann, and Grimme (2020) and Frischlich, Clever, Wulf, Wildschut, and Sedikides (2022) 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 ($\chi^2 = 17.7$, df = 16, p = 0.34).

Table 1

Distribution of topics and frames

	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 frame-topic pairs to four authors (political science master students with prior knowledge concerning framing theory and generic frames 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.

Through this process we generated a multi-topical corpus of 100 news articles with independent frames and topics. In reference to the so-called "gold-standard" of manually coded frames (Nicholls & Culpepper, 2020), we refer to these 100 items as "platinum standard" since the ground truth of frames contained within these 100 multi-topical news articles are known even without manual coding.

Benchmarking

We selected different methods and attempted to detect the frames in those 100 multi-topical news articles.

Automatic inductive methods

All automatic methods that have been claimed of being able to detect 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), STM (Nicholls & Culpepper, 2020), and Topic Model Networks (Walter & Ophir, 2019). The number of clusters to find (k) was five.

Semi-supervised methods

We also investigated semi-supervised methods. This consists of Seeded-LDA (Watanabe & Zhou, 2020) and Keyword Assisted Topic Model (keyATM, Eshima, Imai, & Sasaki, 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. These methods were included 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.

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). 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.

Exploratory analysis: Expert coding. This part of the analysis has not been pre-registered and was planned after the above "gold standard" coding. 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, Van der Velden, & Boukes, 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 pre-registered, 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 possible influence the correctness.

Evaluation: Multiverse analysis

We tested three preregistered hypotheses:

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

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.

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, Meer, & Vliegenthart, 2022). We preregistered all possible combinations of analytical steps and benchmark these methods with all possible combinations of methodological decisions using multiverse analysis (Pipal, Song, & Boomgaarden, 2022). 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.

"Best case" Correct Classification Rate. Suppose y to be ground truth and $\hat{y}$ to be the output from a method. To assess the accuracy of each method, we used the correct classification rate (CCR), where $CCR = Pr(y = \hat{y})$.

As most of these methods are inductive in nature, $\hat{y}$ is a vector of frame indicators, f_k where $k = 1, 2, ..., 5$. However, there is no way to tell which frame indicator corresponds to which actual frame in 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 don't want the variation in these 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}). In this analysis, we generate all possible permutations of all possible values of k, i.e. $\epsilon f_1, f_2, ..., f_k$. 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 in those topics as frames. This expected CCR_{max} value should also be 0.2 theoretically, when frames and topics are randomly assigned. 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. In the figures below, we indicate both null values. One can 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".

We reported also the 89% confidence interval of CCR_{max} (McElreath, 2020).

Results

Automatic inductive methods

K-Means + TF-IDF. Figure 2 shows the result of multiverse analysis of the frame detection method proposed by Burscher et al. (2016). There is not enough evidence to support that the method can deliver significantly better or worse performance than null in 100% of the combinations of methodological decisions.

Principal Component Analysis + TF-IDF. Figure 3 shows the result of multiverse analysis of the frame detection method proposed by Greussing and Boomgaarden (2017). There is not enough evidence to support that the method can deliver

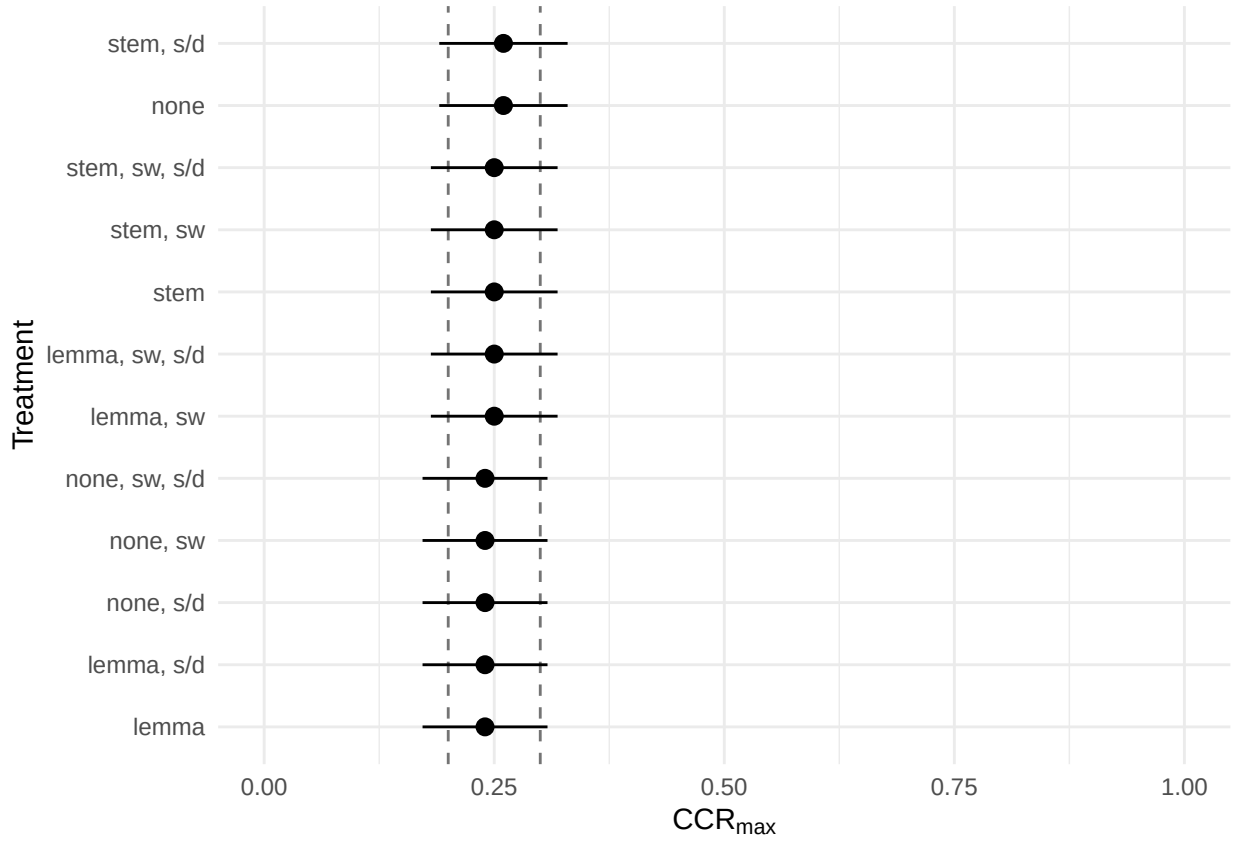

Figure 2. Multiverse analysis of K-Means

significantly better or worse performance than null in 91% of the combinations of methodological decisions.

LDA. Figure 4 shows the result of multiverse analysis of the frame detection method proposed by DiMaggio et al. (2013). There is not enough evidence to support that the method can deliver significantly better or worse performance than null in 47% of the combinations of methodological decisions.

STM. Figure 5 shows the result of multiverse analysis of the frame detection method proposed by Nicholls and Culpepper (2020). There is not enough evidence to support that the method can deliver significantly better or worse performance than null in 91% of the combinations of methodological decisions.

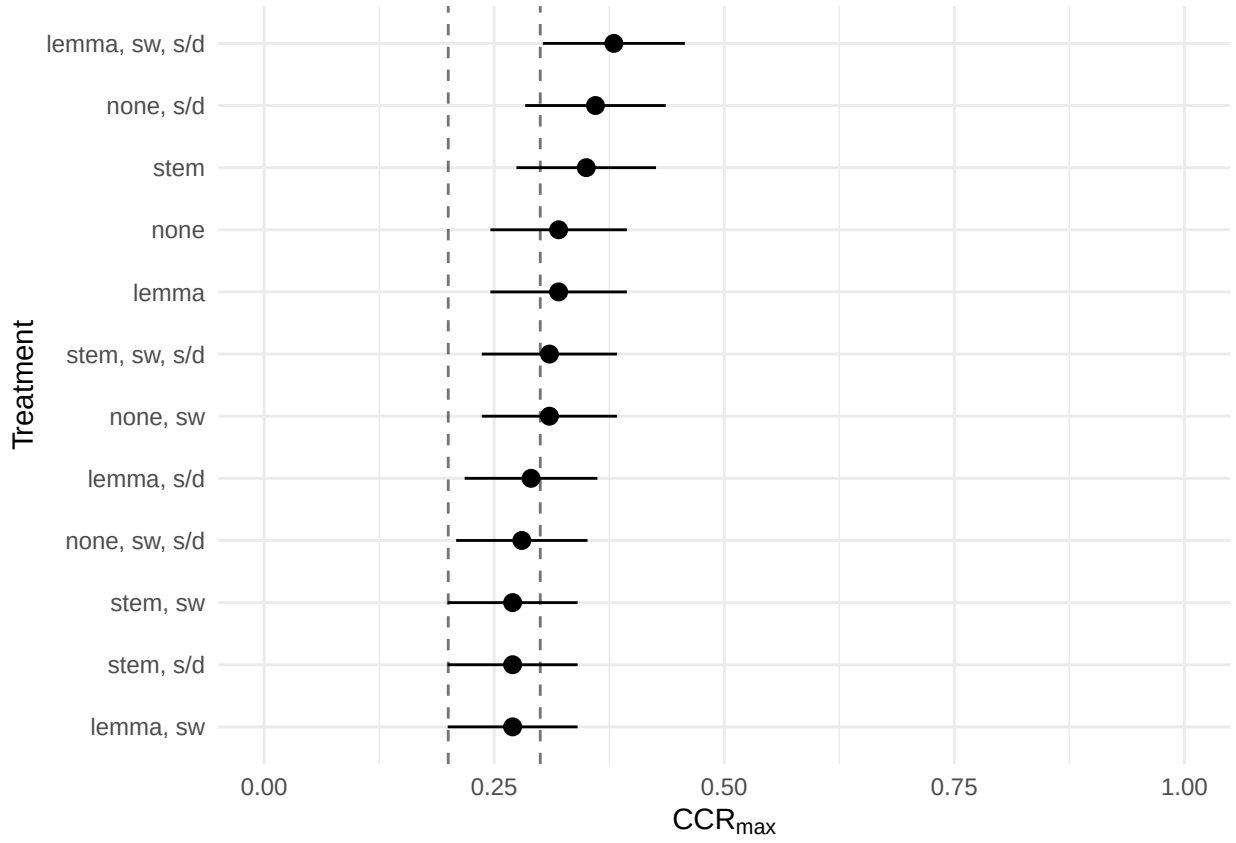

Figure 3. Multiverse analysis of PCA

ANTMN. Figure 6 shows the result of multiverse analysis of the frame detection method proposed by Walter and Ophir (2019). There is not enough evidence to support that the method can deliver significantly better or worse performance than null in 86% of the combinations of methodological decisions.

Semi-supervised methods

Seeded-LDA. Figure 7 shows the result of multiverse analysis of Seeded-LDA (Watanabe & Zhou, 2020). There is not enough evidence to support that the method can deliver significantly better or worse performance than null in 47% of the combinations of methodological decisions.

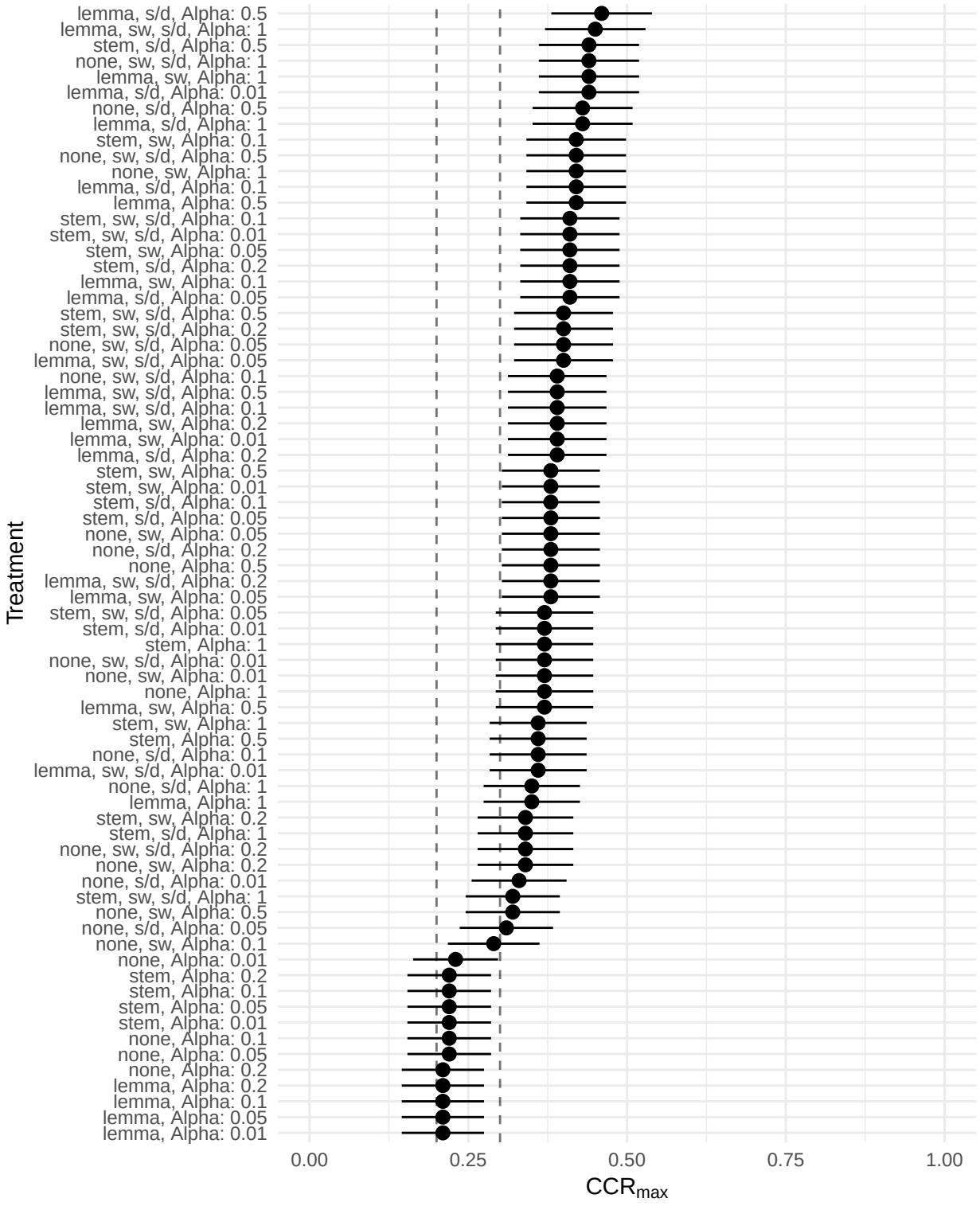

Figure 4. Multiverse analysis of LDA

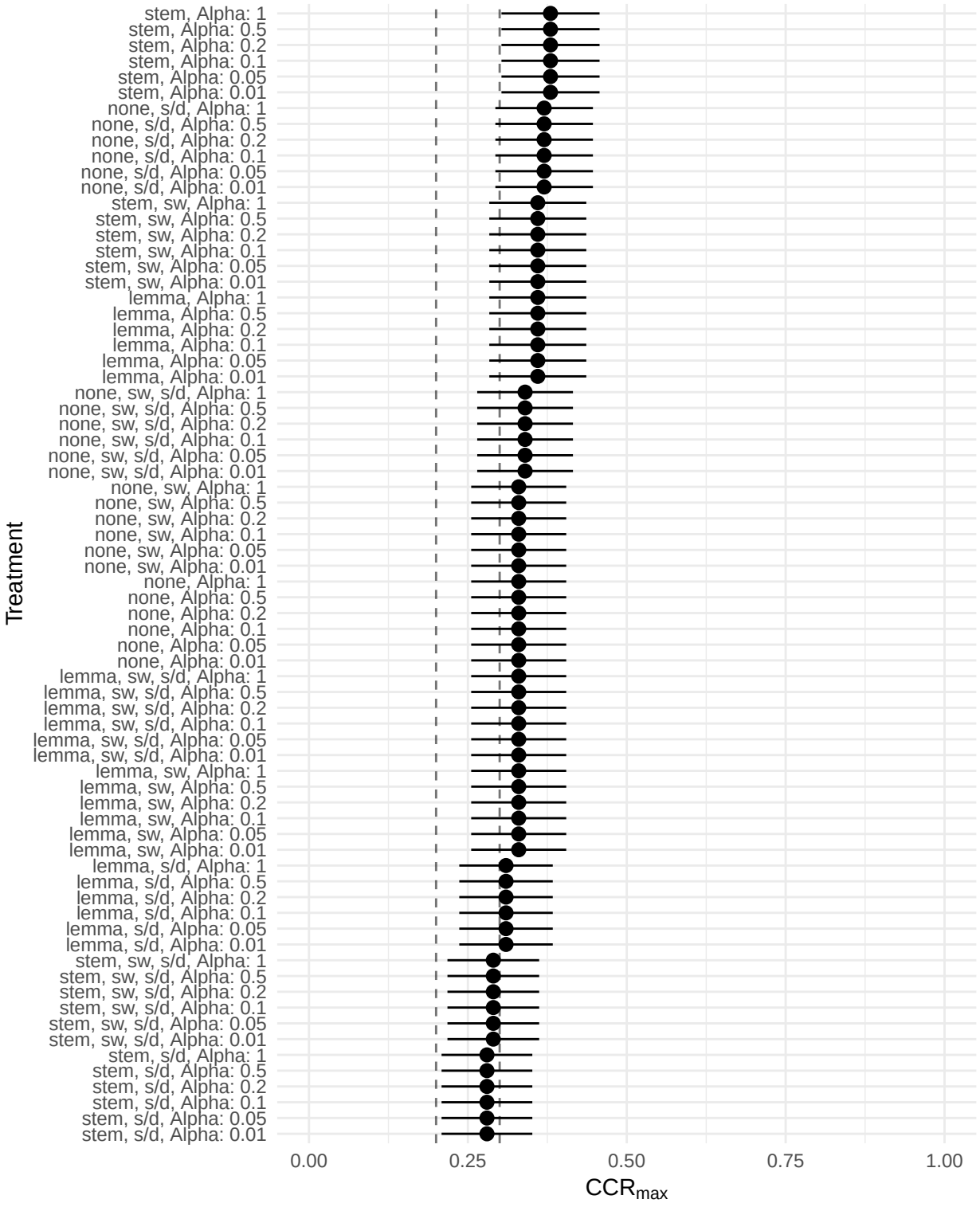

Figure 5. Multiverse analysis of STM

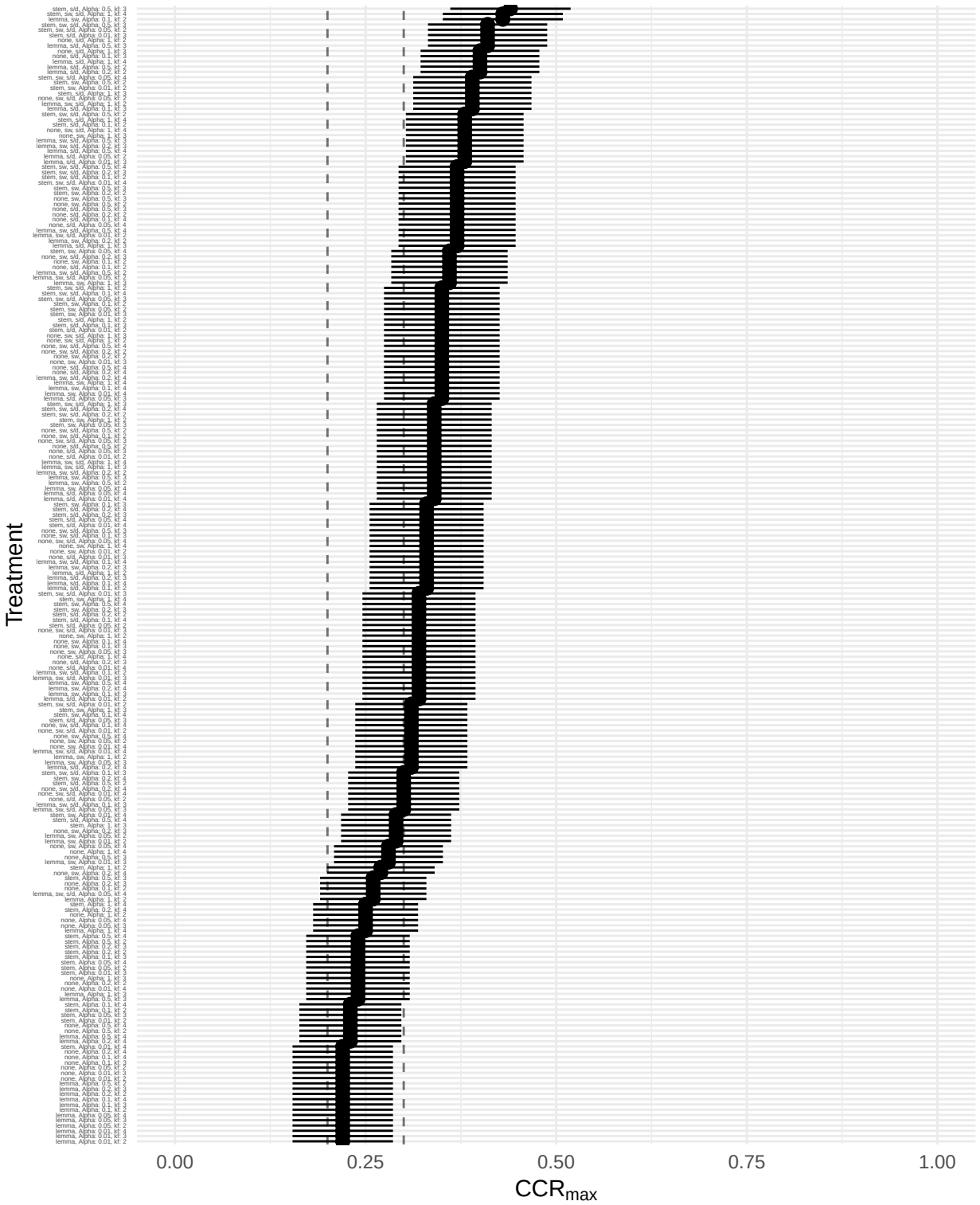

Figure 6. Multiverse analysis of ANTMN

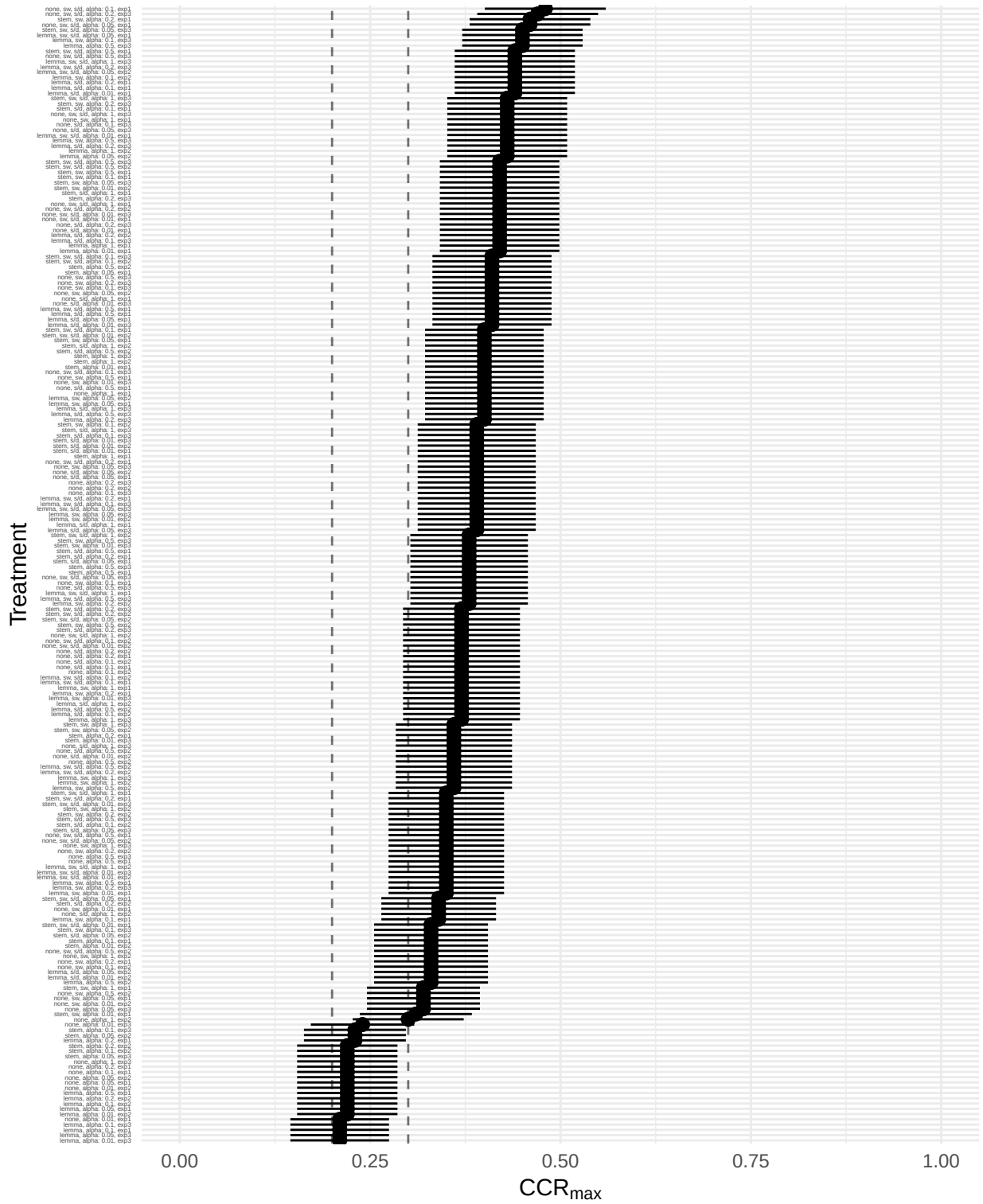

Figure 7. Multiverse analysis of seeded-LDA

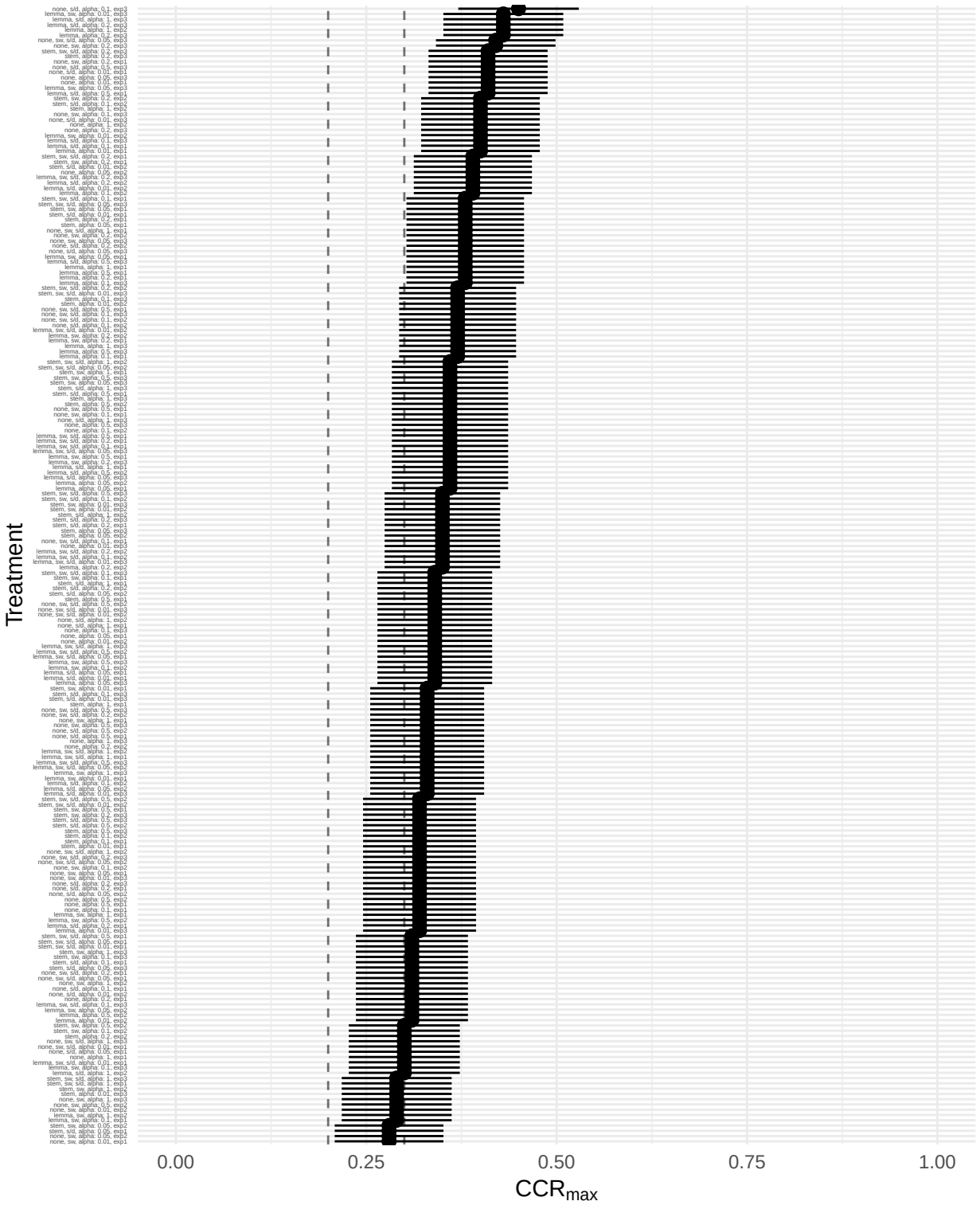

Figure 8. Multiverse analysis of keyATM

keyATM. Figure 8 shows the result of multiverse analysis of keyATM (Eshima et al., 2020). There is not enough evidence to support that the method can deliver significantly better or worse performance than null in 75% of the combinations of methodological decisions.

"Gold Standard"

Figure 9 shows the result of multiverse analysis of the "Gold Standard". The multiverse analysis suggests that the "Gold Standard" can detect frames in the multi-topical news content better than null, regardless of methodological decisions. However, the performance is not as high as one would expect. The best of the best is only around .5.

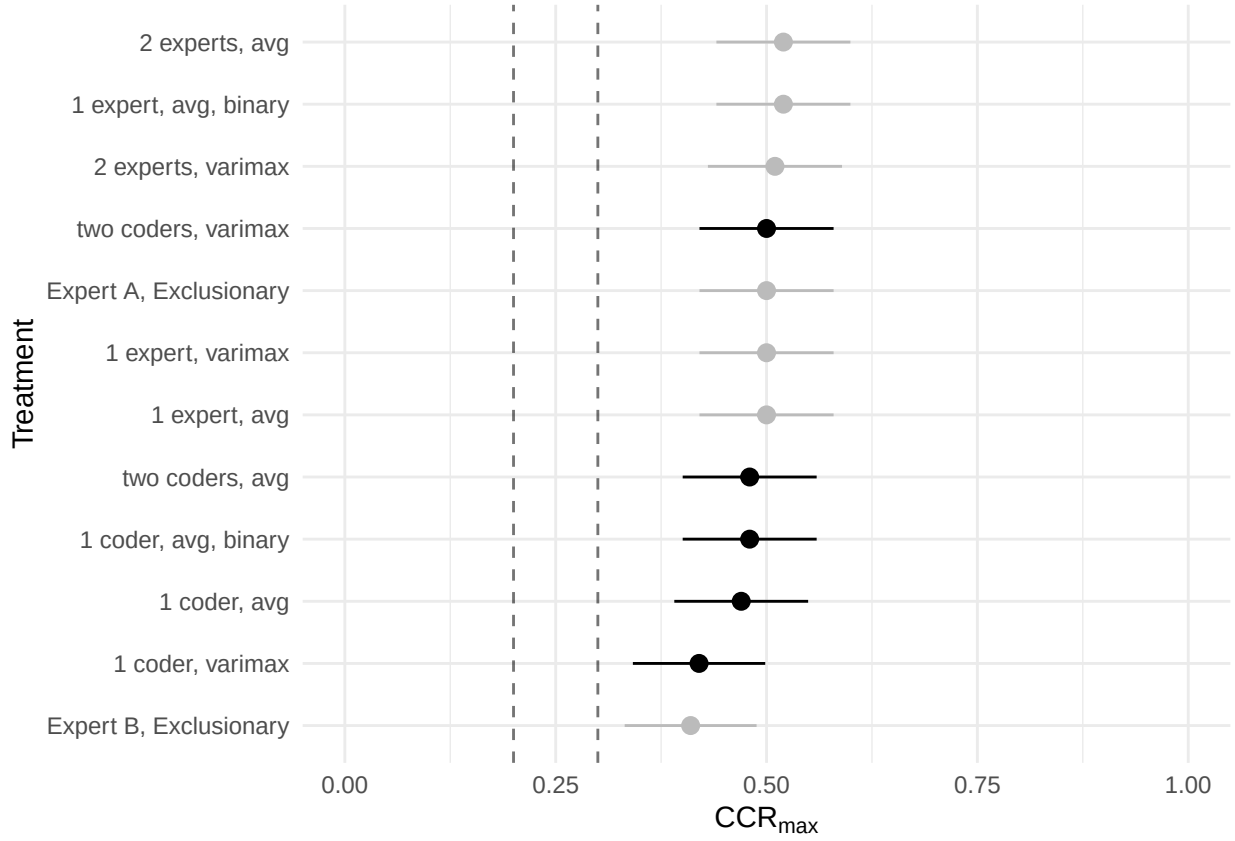

Figure 9. Multiverse analysis of the "Gold Standard" (Grey: Not pre-registered)

The above analysis also reveals that 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.

Comparison of methods

We modeled the CCR_{max} from different methodological combinations of all methods using Bayesian multilevel regression (Bürkner, 2017). Figure 10 shows the robust conditional effects of the gold standard, semi-supervised, and automatic inductive methods on CCR_{max} at 89% credibility (McElreath, 2020). The regression coefficients are available in the online appendix. We repeated here that the not pre-registered expert coding has been excluded in this analysis.

The regression model suggests that the automatic methods have a significantly lower accuracy than the "gold standard". This supports H1. However, there is no meaningful difference between the automatic methods and semi-supervised methods, as well as between semi-supervised methods and "gold standard". There is not enough evidence to support H2 and H3.

Discussion

Based on our review of the framing literature, we provide a synthetic benchmark dataset for frame detection where frames and topics are independent. Using what we called the "platinum standard" dataset, we evaluated methods claimed to be able to detect frames inductively (Burscher et al., 2016; Greussing & Boomgaarden, 2017; Nicholls & Culpepper, 2020; Walter & Ophir, 2019), as well as two semi-supervised methods (Eshima et al., 2020; Watanabe & Zhou, 2020) and the "gold standard" (Semetko & Valkenburg, 2000).

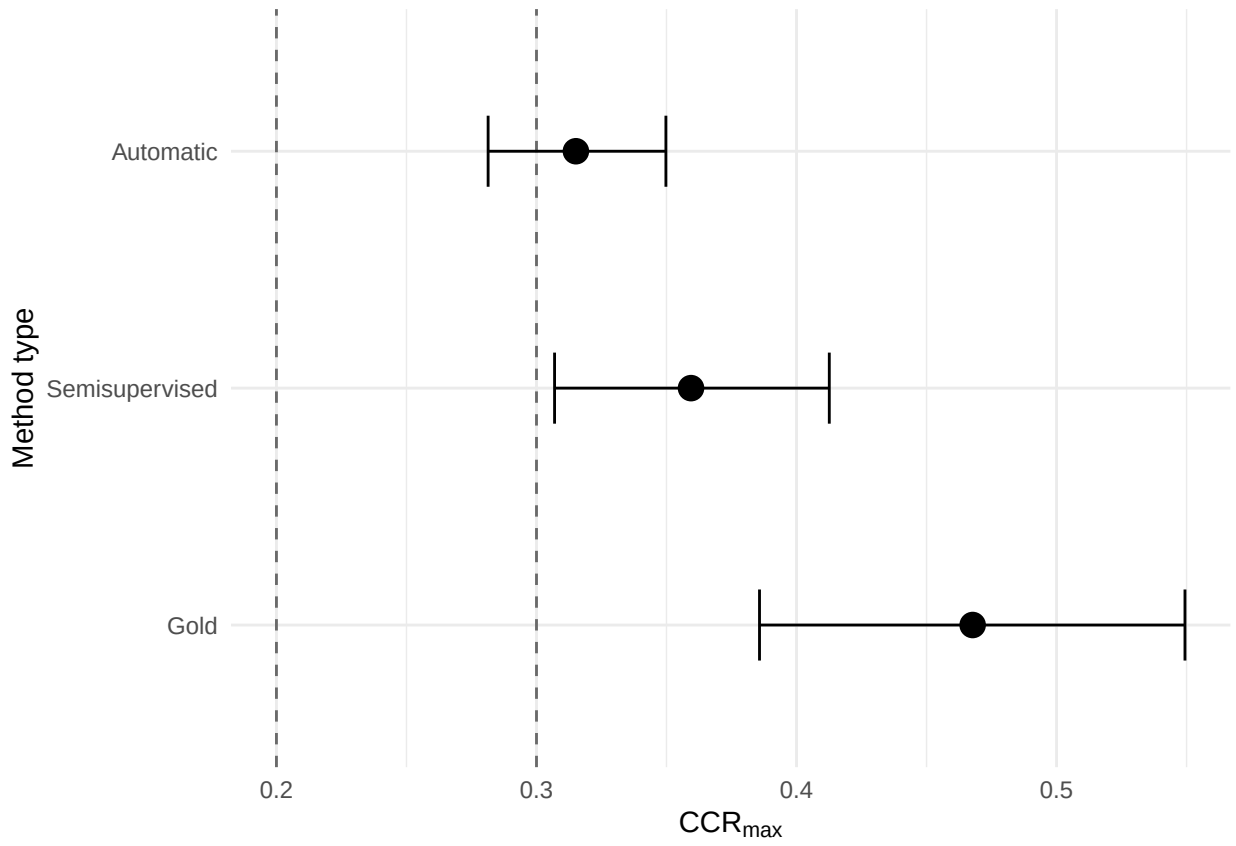

Figure 10. Robust conditional effects from the Bayesian model of the "gold standard", semi-supervised and automatic methods at 89% credibility

Using the multiverse analytic approach (Pipal et al., 2022), we exhaustively benchmarked all methods, irrespective of methodological choices, and reported their "best case" performance. Our benchmark casts doubt on the validity of the so-called "frames" detected by all inductive methods (H1). The conditional effect for all of these inductive methods in figure 10 crosses the *de facto* null value of 0.3, which assumes the detected topics are shoehorned in as an indicator of "frames''. Also, there is no evidence to suggest that the provision of theory-driven dictionaries to those inductive methods can improve the performance (H2) from other automatic inductive counterparts, although the performance of these semi-supervised methods appears to be better than the *de facto* null (Figure 10).

For the first time also, the current study is able to benchmark the "gold standard".

Although the "gold standard" performs significantly better than null, the performance is not superb (Figure 9). There have been concerns about the reliability issues of Semetko and Valkenburg (2000)'s coding scheme (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. But this unexpected finding is 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; Kwak, An, & Ahn, 2020; Liu, Guo, Mays, Betke, & Wijaya, 2019)?

We don't have an answer to the above question, because we didn't study supervised methods in this paper. However, the findings from this paper do point to one important fact: detection of frames from news content written by someone else is an incredibly difficult task, even for experts (Figure 9). This task is incredibly difficult and 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). We always assume we can tell someone else's communicative intention (or someone else's psychological state) and therefore no validation is necessary. But when it is crosschecked, it shows that the assumption that we can tell someone else's communicative intention is well just our hubris. Even experts can confidently make incorrect guesses (see the Online Appendix).

Having said so, it is this paper's authors' hubris that the "platinum standard" can capture the four students' communicator intentions. The approach creates a "turtles all the way down" situation of who can tell whether the four students' communicator intentions have been adequately expressed in the content they produced. We cannot defend this criticism because this is an infinite regress. 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 students. If one has the resource to do so, we strongly recommend replicating this study with journalists.

Another limitation is the sample size of 100. But the 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 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. We can only say that there is not enough evidence to suggest a method is better or worse than null. The former 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. Nonetheless, our apparently small sample size does not affect the confirmed superiority hypothesis (H1).

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 such as ANTMN are more likely to detect **topics** rather than frames when the sample size increases. But we maintain that this finding needs to be confirmed with actual data in a new empirical study.

Despite all the limitations, we provided a benchmark dataset for evaluating frame detection methods and benchmarked various methods with it. The finding from this paper casts doubt on the validity of using automatic inductive methods to detect frames.

References

Baden, C. (2015). *INFOCORE definitions:"Interpretative frame"*. Retrieved from 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

Bürkner, P.-C. (2017). Advanced Bayesian multilevel modeling with the R package brms. *arXiv Preprint arXiv:1705.11123*.

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+ Document Design*, *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

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. (2022). Populists' reliance on nostalgia: A supervised machine learning approach. *International Journal of Communication*.

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., … Wijaya, D. (2022). Proposing an open-sourced tool for computational framing analysis of multilingual data. *Digital Journalism*, 1–22.

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. *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*, 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.

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., … al. (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

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press.

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

Nicholls, T., & Culpepper, P. D. (2020). Computational identification of media frames: Strengths, weaknesses, and opportunities. *Political Communication*, 1–23. https://doi.org/10.1080/10584609.2020.1812777

Pipal, C., Song, H., & Boomgaarden, H. G. (2022). If you have choices, why not choose (and share) all of them? A multiverse approach to understanding news engagement on social media. *Digital Journalism*, 1–21.

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., … 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

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, 1–20. 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*, 1–21. 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., & 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

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