Dipl.-Ing. Markus Reiter-Haas, BSc

Computational Framing Analysis for Polarized Topics Online

PhD Thesis
to achieve the university degree of
Doctor of Technical Sciences (Dr. techn.)
equivalent to the PhD
Doctoral programme: Computer Science

submitted to

Graz University of Technology
Institute of Interactive Systems and Data Science

Supervisor
Assoc.Prof. Dipl.-Ing. Dr.techn. Elisabeth Lex

Co-Supervisor: Univ.-Prof. Mag. Dr.rer.soc.oec. Markus Hadler

External Assessor: Prof. Dr. Martin Potthast

Graz, May 2024

Affidavit

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present doctoral thesis.

Date

Signature

Research is creating new knowledge.

— Neil Armstrong —

Abstract

Framing, which elevates pieces of information in communication to make them more noticeable, affects people's perceptions and choices. Given that framing influences behavior, it is a cardinal consideration for society, especially on polarized topics online, such as COVID-19 and climate change. Although the analysis of frames is crucial, frames are exceptionally challenging to conceptualize. Additionally, annotated framing data is only scarcely available. Due to that, I aim to leverage the breakthroughs in natural language processing to advance computational framing research on two fronts. First, I developed framing detection algorithms for three distinct exploratory levels, i.e., frame labels, frame dimensions, and frame structure. My work shows trade-offs between the validation of established and the exploration of novel frames using a multi-perspective approach. Second, I studied the relations between content and users concerning the prevalence of the frames employed in online systems. A substantial interplay between user behavior and the framing of content in information systems is revealed in a research direction yet to be explored. At its core, my research integrates social science research with computational approaches, broadening the field and revealing several new research directions. Besides fostering an increased understanding of framing, I developed novel methodologies for framing analysis and released their artifacts for public use, e.g., as open-source tools. My findings can inform the design of future information systems to balance the users' online behavior regarding the framing diversity of the content.

Kurzfassung

Framing, bei dem Informationen in der Kommunikation hervorgehoben werden, um sie auffälliger zu machen, beeinflusst die Wahrnehmungen und Entscheidungen von Menschen. Da Framing das Verhalten beeinflusst, ist es für die Gesellschaft von zentraler Bedeutung, insbesondere bei polarisierenden Online-Themen, wie etwa COVID-19 und dem Klimawandel. Obwohl die Analyse von Frames von entscheidender Bedeutung ist, ist es außerordentlich schwierig, Frames zu konzeptualisieren. Zusätzlich sind annotierte Framing-Daten nur in geringem Umfang verfügbar. Aus diesem Grund will ich die Durchbrüche in der Verarbeitung natürlicher Sprache nutzen, um die computergestützte Framing-Forschung an zwei Fronten voranzutreiben. Erstens habe ich Algorithmen zur Erkennung von Framing auf drei verschiedenen Untersuchungsebenen entwickelt, nämlich Frame-Bezeichnungen, Frame-Dimensionen und Frame-Struktur. Durch unseren multiperspektivischen Ansatz finden wir Abwägungen zwischen der Validierung etablierter Frames und der Erforschung neuartiger Frames. Zweitens untersuche ich die Beziehungen zwischen Inhalten und Nutzer:innen in Bezug auf die Prävalenz der in Online-Systemen verwendeten Frames. In dieser bisher unerforschten Forschungsrichtung zeigen meine Arbeiten eine wesentliche Wechselwirkung zwischen dem Nutzungsverhalten und dem Framing von Inhalten in Informationssystemen. Im Kern integriert meine Forschung sozialwissenschaftliche Forschung mit computergestützten Ansätzen, wodurch das Feld erweitert und mehrere neue Forschungsrichtungen aufgezeigt werden. Neben der Förderung eines besseren Verständnisses von Framing habe ich neuartige Methoden für die Framing-Analyse entwickelt und ihre Artefakte zur öffentlichen Nutzung freigegeben (z. B. als Open-Source-Tools). Die Erkenntnisse meiner Forschung können in die Gestaltung künftiger Informationssysteme einfließen, um das Online-Verhalten der Nutzer:innen im Hinblick auf die Framing-Diversität der Inhalte zu verbessern.

Acknowledgements

First and foremost, I would like to thank my supervisor, Elisabeth Lex, for her invaluable guidance throughout my PhD studies. She kept motivating me in my scientific endeavors, gave me the freedom to pursue my ideas, and was always available when needed. Therefore, I am very grateful for her continuous support during this rollercoaster ride of my life.

Next, I would like to thank my collaborators, without whom the thesis would not have been possible in its current form. Notably, I want to thank co-supervisor Markus Hadler and colleague Beate Klösch from the University of Graz for their fruitful cooperation in the interdisciplinary Route63 project[1]. I highly appreciate their insights, which exceeded the materials in my academic studies. Moreover, it was a pleasure to work with them, exemplified by their willingness to use LaTeX, which is much appreciated. I would also like to thank my student Alexander Ertl for his enthusiasm during the SemEval participation and subsequent research. It was a rewarding experience working with him and supporting him in conducting his Master's thesis.

Additionally, I would like to thank my office colleagues from both the two of my offices (i.e., Finger Four at the Data House and FAIR-AI at the Know Center), for their helpful tips, feedback, and general discussions: Tomislav Đuričić, Niklas Hopfgartner, Kevin Innerebner, Simone Kopeinik, Dominik Kowald, Emanuel Lacić, Thorsten Ruprechter, and Alexander Steinmaurer.

Last but not least, I would like to thank my family and friends, for whom I would often not dedicate enough time. Thank you for your support and believing in me. One special thanks to my parents Christina and Herbert and brother Robert for always being there for me, and assisting me throughout my whole studies. And finally, one very special thanks to my girlfriend Helena, for always standing by my side and the emotional support whenever needed.

[1] I want to thank the rectorates of Graz University of Technology and University of Graz for their funding of this research project.

Chapter Overview

Chapter 1 - Introduction motivates the importance of the conducted research and presents the research questions and contributions of the work.

Chapter 2 - Background provides the necessary background knowledge. Theoretical underpinnings are outlined, followed by a description of the leveraged methodology of using NLP for text understanding. Furthermore, related work of the emerging but still sparse research of computational framing analysis is discussed.

Chapter 3 - Publications includes the core publications of this thesis and describes their specific findings. Besides the core publications, supplementary publications conducted during the PhD studies are also listed and described.

Chapter 4 - Conclusion summarizes the most significant findings and discusses their main implications. Besides, it critically reflects on the conducted research by addressing open problems and limitations. Finally, it outlines potential future research directions that have emerged from the thesis.

Contents

List of Figures

List of Tables

The best way to predict the future is to invent it.

(Alan Kay)

1 Introduction

Understanding how people interact with information is crucial for society at large on many topics, such as health information during a pandemic. In this regard, Soroya et al. (2021) describe that heightened social media exposure, such as during the COVID-19 pandemic, can lead to information overload and subsequent information anxiety and avoidance. Moreover, White and Hassan (2014) showed that the retrieved content from health searches tends to be biased towards showing positive results for medical interventions, which is a deviation from reality. Biases could also stem from users themselves, e.g., with confirmation bias where users select information supporting their established beliefs (Del Vicario et al., 2017). Such cognitive biases have been intertwined with polarization in hundreds of research articles (Xing et al., 2024). To that end, Westerwick et al. (2017) studied the influence of message content on confirmation bias and related it to attitude polarization. Hence, understanding the relationship between users and content is an essential aspect of information systems.

There are many ways to study and understand content, with sentiment analysis (Pang, Lee, et al., 2008) being a prominent example. Besides, there are more nuanced aspects, such as the framing of the content. While framing is a fragmented term in literature, a common definition deals with the salience of certain aspects within communication, which implicitly suggest certain solutions to a problem (Entman, 1993). For instance, immigration can be framed economically by emphasizing gains and losses (Card et al., 2015). However, while the analysis of framing is well established in social sciences, it still remains underexplored from the computational perspective and the analyzed conceptualizations vary in literature (see Ali & Hassan, 2022).

Nowadays, approaches tend to use neural networks, typically relying on Transformers (Vaswani et al., 2017) – a specialized architecture for framing detection. For instance, in the SemEval 2023 framing detection task (Piskorski et al., 2023), many of the best-performing teams employed Transformer-based approaches (Liao et al., 2023; Reiter-Haas et al., 2023a; Wu et al., 2023). Besides predicting frames in a supervised setting, there are several unsupervised approaches, e.g., using topic models (DiMaggio et al., 2013) or embeddings (Kwak et al., 2021).

In this thesis, I use text understanding methods based on embeddings and Transformers for computational framing analysis. Specifically, my focus lies on polarized topics in the online space. The study of polarized topics is especially important, as the opinions of people differ, but this fact may only be expressed in very subtle ways. For instance, two divergent stances on COVID-19 measures could both be rooted in fear, where one side is fearful of the virus itself while the other side fears too much governmental control. Therefore, one side could emphasize *preventing the spread of the virus*, while the other side could advocate for *fighting for their liberties* (as depicted in Figure 1.1). Moreover, the emphasis on prevention in the first argument could similarly be reframed to *preventing individual liberty*, thus shifting the focus from community effort to personal responsibilities. Detecting such differences is challenging, and even more so when trying

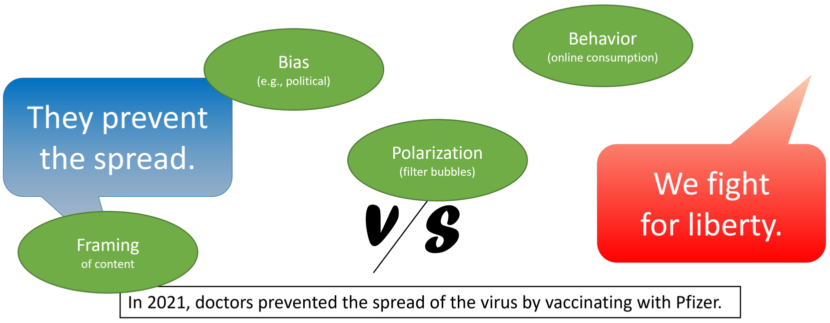


Figure 1.1: Schematic Depiction of the Motivation. The framing of the example sentence is shown in blue, which is likely to also reflect implicit biases. The topic has the potential to be polarized with alternative framings shown in red, which could affect the behavior of people.

to understand how the framing influences online discourses due to user behavior.

With the problem of detecting the framing in polarized topics online in mind, I tackle the task from two angles. First, I investigate how to improve the detection techniques at distinct exploratory levels. Second, I study the relationship between frames and the corresponding behavior in online media. Therefore, I aim to expand the knowledge of computation framing analysis and the influence of framing on behavioral patterns in selected polarizing topics. The specific topics that I study comprise COVID-19 measures, climate change, health conspiracies, gun violence, political social media posts, misinformation and disinformation, and online news. To that end, I combine methods of natural language understanding with framing theory for computational framing analysis, supported by research in opinion polarization and behavior modeling.

1.1 Research Questions

The thesis deals with two primary research questions (RQs). The first research question (RQ1) is methodological-centric and the focus of the thesis, while the other (RQ2) is observational and cross-cutting. The first question is divided into three sub-questions, each dealing with a different aspect and increasing in exploratory levels.

RQ1: How to detect differences in the framing of online content at various exploratory levels?

Detecting the framing of texts is challenging, as frames tend to be very nuanced. Moreover, framing detection is more complex than other detection targets, such as topics, which also limits the available data. Therefore, framing detection cannot rely on large framing datasets that contain annotations. Instead, framing detection must exploit the information in the low amount of yet available data. The amount of suitable data also depends on the kind of frames that are being analyzed, as some frames can be broader and some are narrower defined. As such, computationally assigning a single frame to a document has been criticized compared to the higher-level conclusions derived in the social sciences (Vallejo et al., 2023).

The aim is to extract frames at various exploratory levels. To that end, I consider framing detection from multiple perspectives and use pretrained embeddings and Transformers for transfer learning. I first employ the framing detection approaches at each exploratory level individually before incorporating them into one complete solution. I split the framing approaches into frame labels, frame dimensions, and frame structure. Frame labels are used when annotation data is available. Whereas, I consider frame dimensions when annotations are absent but antagonistic pole descriptions can be leveraged. Finally, the frame structure is analyzed without using explicit training data for exploratory purposes. A schematic overview of the approaches is provided in Figure 1.2.

On the one hand, data availability is a critical challenge, even in the supervised setting with frame labels. A standard classification pipeline would not suffice, as even the ratio between data points to the number of labels tends to be extremely low. On the other hand, in a completely exploratory setting for frame structure, the validation is the main issue, as no ground truth is available. The frame dimensions fall in the middle as a special case, where there is a specific kind of data available, potentially with test data.

In sum, I find that there are several trade-offs to consider regarding validation and exploration. The more data and information is available for a particular frame, the better validation can be conducted. Conversely, to find novel frames in an explorative manner, more observational approaches are needed.

RQ1a: How to extract Framing Labels with limited annotated data?

Data scarcity is a prevalent issue in framing detection. In general, there are only a few datasets with labeled frames, as the annotation process is very labor-intensive. Moreover, framing data tends to be imbalanced, as the prevalence of frames varies significantly. Furthermore, in certain scenarios, no data may exist in the target domain at all. These issues lead to traditional classification methods to be infeasible. Alternative methods, such as few- and zero-shot learning, are thus more important. Besides, there are multiple methods to effectively exploit the small amount of data available for framing detection. The problem can still be modeled as a classification task to predict labels (discrete) or their probabilities (continuous).

In Reiter-Haas et al. (2023a), our approach uses contrastive learning in combination with a multilingual multi-stage pipeline. With the multi-label-aware contrastive learning procedure, we optimize the embedding space. We find that by using a weighted contrastive loss, we can improve the embedding space regarding uniformity and alignment. In particular, samples that have more labels in common appear closer together compared to mostly dissimilar labels. This in turn improves subsequent classification and thus accuracy.

RQ1b: How to extract Framing Dimensions in an unsupervised manner?

In some scenarios, labeled training data can be completely missing. In such cases, we can use pretrained transformer models to capture semantic information in the textual data. By combining them with models from the social sciences, such as the moral foundations theory, we can interpret the underlying frames within the texts. This approach allows us to uncover and analyze the moral dimensions present in the data without the need for labeled examples.

In Reiter-Haas et al. (2021b), we consider the embedding space of pretrained models to measure their alignment with antagonistic poles using the FrameAxis approach (Kwak

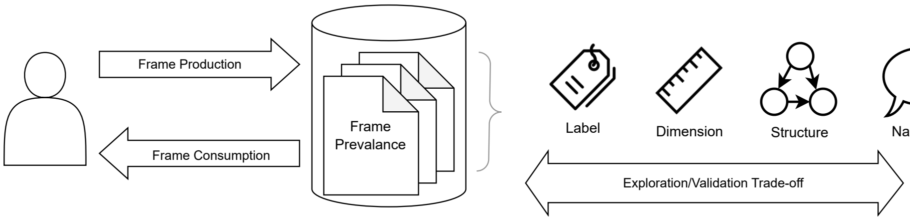


Figure 1.2: Overview of Framing as discussed in the thesis. On the left, I depict the framing in online behavior, while the right part depicts the framing types for detection approaches.

et al., 2021). These poles thus form axes from which we can derive a one-sided leaning (i.e., frame bias) and variation (i.e., frame intensity). Both measures are on a continuous scale, with frame biases being signed and frame intensity being an absolute value. The approach is particularly fitting when aggregating results, e.g., of multiple users. Moreover, we find that the extracted frame biases are topic-dependent. For instance, the moral framings of COVID-19 tweets in Austria oppose the established alignment. In particular, conservatives emphasize care rather than authority, as would be expected.

RQ1c: How to extract Framing Structure without prior conceptualization?

As the amount of frames is unbound, the most important frames to research might not be known in advance. Therefore, the frames need to be discovered first rather than predefined. In Reiter-Haas et al. (2024c), we transform texts into semantic graphs and mine structural patterns. The mined patterns allow us to find simple but predominant narrative elements without any prior information. To that end, the approach relies on frequency information of substructures. While our approach allows for distinct patterns to be found, it still requires expert knowledge to interpret the potential frames. Therefore, the approach is explorative in nature based on mixed methods, i.e., combining quantitative and qualitative analysis.

RQ2: How does framing relate to online information behavior?

Framing from a cognitive perspective is known to affect human behavior. However, the relation between frames and online information behavior is not clear. Therefore, the question arises whether we observe this influence of framing in online systems too. In this thesis, I consider the influence from two sides, how content is produced with respect to its wording (i.e., *frame production*) and how frames influence the content that is consumed (i.e., *frame consumption*). This is unlike RQ1, which deals only with the presence of frames in media (i.e., *frame prevalence*).

For the production, I simplify the problem by considering opinion polarization regarding the expressed sentiment and its relation to opinions offline. This should answer whether people repeat similar sentiments (as a proxy for framing) when posting online compared to their offline opinions. Therefore, it deals with potential influences from the outside, like opinion formation. For the consumption, I observe the user behavior in terms of their content consumption regarding framing and compare it against topics. In that regard, in the thesis, I focus on news information behavior specifically. I find that users tend to show similar online and offline characteristics. Moreover, users tend

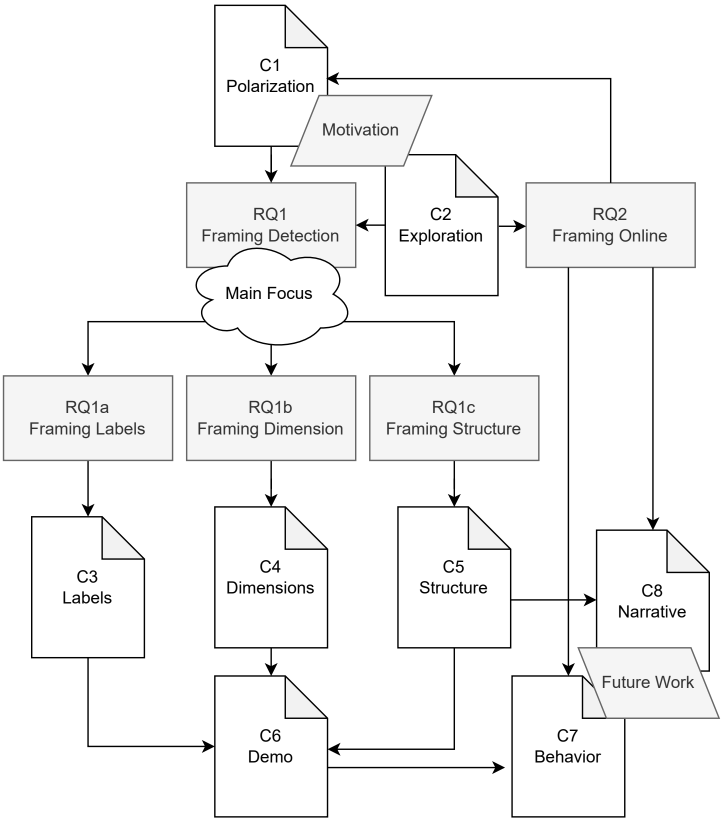


Figure 1.3: Overview of RQs and how they are related, as well as answered by the included publications.

to repeatedly consume similarly framed news online. Finally, I also regard the overlap between narratives and frames as a promising future research direction.

1.2 Scientific Contributions

This section discusses the scientific contribution of this thesis in three subsections. First, I list all relevant contributions of the doctoral project. Then, I describe the relation between the publications and the stated research questions. And finally, I briefly summarize the main contributions.

1.2.1 Overview of Contributions

In this subsection, I provide a brief overview of the published works, while Chapter 3 describes the publications in detail. Table 1.1 lists the core publications, which are referenced by their key (with hyperlinks to the articles) and short names. Each core publication is associated with a respective research question. These publications thus constitute the main part of the thesis.

Besides, several additional publications were published that go beyond the main topic of the thesis. These are typically the result of collaborations that span related topics, as listed in Table 1.2. Each of the supplementary publications is associated with one of the three theoretical foundations of the thesis, i.e., information *behavior*, opinion *polarization*, *framing* theory.

Besides tackling the stated research questions (Section 1.1) and the resulting publications, I provide a brief list of additional contributions as part of the PhD research in Table 1.3. These contributions comprise datasets, conference abstracts, and posters, as well as a non-peer-reviewed article in a special journal issue.

1.2.2 Relation between Publications and Research Questions

Here, I briefly discuss the relations between the research questions and the core publications, as visually depicted in Figure 1.3. C1 (Polarization) is set as the root, which investigates how sentiment and opinion are related. Therefore, it partially answers RQ2 with sentiment as a proxy and nicely motivates the importance of RQ1. C2 (Exploration) then provides an overview of the research as a whole and serves for the formulation of both RQs. The three sub-research questions are each tackled in one core publication, i.e., C3 (Labels) for RQ1a, C4 (Dimensions) for RQ1b, and C5 (Structure) for RQ1c. All three are consolidated in C6 (Demo), which is the open tool for framing research. Hence, publications C3 until C6 answer RQ1 and are the main focus of the thesis. Afterward, C7 (Behavior) uses C6 to answer the other part of RQ2. Finally, C8 (Narrative) opens a novel direction by using the insights gained from C5 and how they relate to novel directions regarding RQ2.

1.2.3 Summary of the Main Contributions

I summarize my contributions as follows:

1. My research is positioned in the emergent field of **exploring content bias in online media**. To that end, I use computational framing analysis to uncover latent aspects that potentially influence readers.
2. I employ **state-of-the-art approaches of deep learning to new problem settings**. Specifically, I employ the Transformer architecture for content analysis in online media.
3. In my research, I developed **novel computational methods to detect and investigate differences in framing**. Consequently, I provide new insights into how certain topics are framed online.
4. My interdisciplinary research **bridges strands of the social sciences and computer science**. Notably, I use tools of natural language processing to advance framing research.
5. My research **fosters advancements in the scientific community** by participating in shared tasks and releasing artifacts, such as open-source tools.

In sum, this thesis makes several relevant contributions to a specific area, i.e., computational framing research, but has implications for many related areas. Specifically, I see the design of information systems and the impact of framing on public opinion as the most noteworthy points of connection.

Core Publications

Key (ShortName)	Reference	Type (Subtype)	RQs
C1 (Polarization)	Reiter-Haas et al., 2023b	Journal	RQ2
C2 (Exploration)	Reiter-Haas (2023)	Conf. (Symposium)	RQ1, RQ2
C3 (Labels)	Reiter-Haas et al. (2023a)	WS (Shared Task)	RQ1a
C4 (Dimensions)	Reiter-Haas et al. (2021b)	Conf. (Short)	RQ1b
C5 (Structure)	Reiter-Haas et al. (2024c)	in Review (w/ preprint)	RQ1c
C6 (Demo)	Reiter-Haas et al. (2024b)	Conf. (Demo)	RQ1
C7 (Behavior)	Reiter-Haas and Lex (2024)	WS (Full)	RQ2
C8 (Narrative)	Reiter-Haas et al. (2024a)	WS (Position)	RQ2

Table 1.1: Overview of the Core Publications (ordered as presented in Figure 1.3). Type of contribution with potential subtypes, such as the track of conferences (Conf.) or type of workshop (WS).

Supplementary Publications

Key (ShortName)	Reference	Type (Subtype)	Topic
S1 (Relistening)	Reiter-Haas et al. (2021c)	Conf. (LBR)	Behavior
S2 (Bridging)	Reiter-Haas et al. (2020)	Workshop	Polarization
S3 (Glue)	Hadler et al. (n.d.)	Journal (in Review)	Framing
S4 (Humanize)	Kowald et al. (2024)	Book Chapter	Behavior
S5 (ACT-R+CF)	Moscati et al. (2023)	Conf. (Short)	Behavior
S6 (Desirability)	Klösch et al. (2022)	Conf. (Extended)	Polarization
S7 (Willingness)	Hadler et al. (2022)	Journal	Polarization
S8 (Role)	Klösch et al. (2023)	Journal	Polarization

Table 1.2: Overview of Supplementary Publications, approximately ordered by level of importance to the thesis (i.e., a combination of relevance and involvement).

Other Contributions

Key (ShortName)	Reference	Type (Subtype)	Topic
O1 (MINDFrames)	Reiter-Haas (2024)	Dataset	Framing
O2 (Opinion)	Reiter-Haas et al. (2021a)	Conf. (Poster)	Polarization
O3 (Semantic)	Reiter-Haas et al. (2023c)	Conf. (Poster)	Framing
O4 (Comparison)	Klösch et al. (2021a)	Conf. (Abstract)	Polarization
O5 (Teaching)	Ambros et al. (2023)	Journal (Special)	Polarization
O6 (SurveyData)	Hadler et al. (2021)	Dataset	Polarization
O7 (Insights)	Klösch et al. (2021b)	Conf. (Abstract)	Polarization
O8 (ClimateGlue)	Hadler et al. (2024)	Conf. (Abstract)	Framing
O9 (LastGen)	Wardana et al. (2024)	Conf. (Abstract)	Framing

Table 1.3: Overview of Other Contributions, approximately ordered by level of importance to the thesis (i.e., a combination of relevance and involvement).

Attention Is All You Need

(Title of groundbreaking paper by Vaswani et al.)

2 Background

This chapter provides a review of the relevant background for the thesis. Section 2.1 focuses on the underlying theoretical concepts covered in the thesis: Information behavior (Section 2.1.1) originates from information sciences, whereas polarization (Section 2.1.2) and framing theory (Section 2.1.3) are rooted in the social sciences. In Section 2.2, a review of natural language understanding, which is a subarea of natural language processing rooted in computer science, is provided where I specifically focus on textual data. The section contains the main methodological approaches and is further divided into two subsections. Subsection 2.2.1 surveys the specific advancement of deep learning that powers most of today's artificial intelligence systems, i.e., embeddings and Transformers. Subsection 2.2.2 lists popular content analysis and text mining approaches, which typically use the former models as a basis. Finally, Section 2.3 discusses related work in computational framing research, which is the core of the thesis.

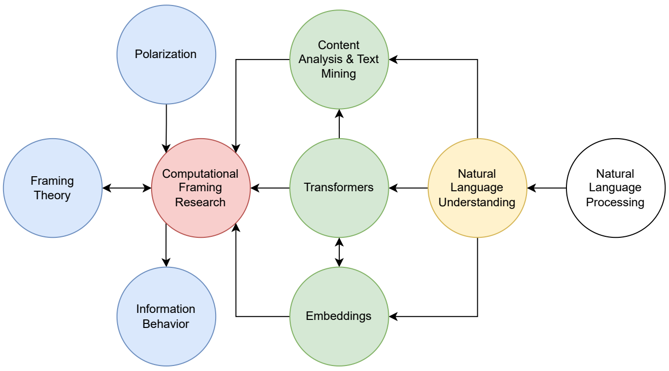

Polarization
Content Analysis & Text Mining
Framing Theory
Computational Framing Research
Transformers
Natural Language Understanding
Natural Language Processing
Information Behavior
Embeddings

Figure 2.1: Conceptual overview of Relation between Background research areas. Blue are the theoretical areas covered. Orange is the main methodological area, with subparts in green. Red is the specific niche of the thesis.

Figure 2.1 visually depicts the areas and their relations. At the core is computational framing research, which is directly related to framing theory in the social sciences. Hence, they mutually benefit each other. In this thesis, I focus on polarized topics as the target of the research while I investigate the relation of framing to information behavior. From a methodological perspective, I use content analysis based on Transformers and embeddings, which are all part of natural language understanding, a subfield of natural language processing.

Each section first provides the relevant background in general and then goes into the specific details required for the thesis. Afterward, I outline highly relevant core publications for the given background section in a gray box, where I also mark the most prototypical example in bold. Finally, each section concludes with a brief statement on how the thesis is related to the described research.

2.1 Theoretical Underpinnings

The thesis lies at the intersection of computational sciences and various theories. Hence, I start by providing the necessary background on the three main theoretical areas on which the practical work is based on, i.e., information behavior, opinion polarization, and framing theory.

Besides the main theories discussed in the following section, I have also used other theories in the course of the thesis, in particular, the moral foundations theory (Graham et al., 2013) and the theory of narrative understanding (Piper et al., 2021). For a detailed description of these theories, I refer the reader of this thesis to C4 (Dimensions) and C8 (Narrative).

Relevance for: C1-C7, **C8**

In every core publication, the theoretical underpinnings are an essential part and thus marked as relevant. C8 (Narrative) is the most theoretical one and ties together the two distinct research strands of computational framing analysis and computational narrative understanding.

2.1.1 Information Behavior

In the information sciences, a model of information behavior was initially conceptualized by Wilson (1981) and mainly relates to information seeking in information systems to satisfy information needs. In Wilson (2000), the distinction between information behavior in general, and subtypes of information seeking (i.e., to satisfy an information need), information searching (i.e., interactions with information systems), and information use (i.e., incorporating the information) is further clarified. The term information behavior is more generally described in Bates (2010) with a focus on information interaction, which comprises both how humans seek and use information.

In this thesis, I focus on a subset of behavior patterns relating to repeat consumption and viewpoint diversity. Regarding repeat consumption, Anderson et al. (2014) showed that users tend to repeatedly consume the same items in several information systems, such as on YouTube. We found similar patterns in one of my supplementary works in the music domain (Reiter-Haas et al., 2021c). In information systems, another important consideration is accounting for the diversity of retrieved content (Clarke et al., 2008). For instance, Draws et al. (2021a, 2021b) assessed the viewpoint diversity of search results and linked them to cognitive biases. Biases are one-sided tendencies, which also affect user interactions on the Web (Baeza-Yates, 2018).

Relevance for: C1, C2, C6, **C7**, C8

The thesis considers framing as a form of bias and relates it to behavior in information systems, with a specific focus on repeat consumption and viewpoint diversity. While this is explicitly established in C7 (Behavior), several other core publications consider parts of information behavior, such as information access in retrieval systems.

2.1.2 Opinion Polarization

Polarization of opinions is characterized by dispersion and bimodality of distributions and is both a state and a process (DiMaggio et al., 1996). Bramson et al. (2017) more explicitly establish nine senses of polarization in distributions, with dispersion being a particular sense and bimodality blurring multiple senses. Polarization has been extensively analyzed on social media regarding its quantification, dynamic process, and potential reduction measures (Garimella et al., 2018). He et al. (2021) use contextualized embeddings to detect polarized topics. Polarization has also been linked to biases, such as confirmation bias (Del Vicario et al., 2017), emerging through behavior in agent-based modeling (Sikder et al., 2020).

In this thesis, I consider polarization as a state and first establish that polarization is similar offline and online using statistical measures such as the bimodality coefficient (Ellison, 1987). Besides, I use polarization for topic selection and polarized data sources as the focus of my doctoral research. Moving from a first initial study on studying polarization in public opinion using sentiment analysis (Reiter-Haas et al., 2023b), my later work focuses on frame detection rather than sentiment analysis as a form of opinion mining (Pang, Lee, et al., 2008).

Relevance for: **C1**, C2, C4-C6, C8

Only in C1 (Polarization), but in several supplementary publications, is polarization the target of analysis. Nevertheless, polarized topics are used as the specific focus of the thesis due to the divergence of opinions. In this regard, it has been shown that there is also a linguistic divergence in polarized online media (Karjus & Cuskley, 2024). Hence, I base my choice on the hypothesis that more polarized topics should be a suitable target to find more pronounced differences in the framing of textual data.

2.1.3 Framing Theory

Framing has long been considered a fractured paradigm (Entman, 1993). Sullivan (2023) argues that framing operates on three levels: semantic frames, cognitive frames, and communicative frames. The semantic frames of Fillmore et al. (1976) appear as an intrinsic part of language. Many subsequent computational linguist works build upon this notion, with FrameNet (Baker et al., 1998) and PropBank (Kingsbury & Palmer, 2002) being two notable examples of language resources. Using such resources, frames build semantic relations, such as dependencies, between various elements of texts (Fillmore & Baker, 2001). Cognitive frames affect our thoughts, e.g., through using metaphors (Lakoff, 2014). Such phenomena are well explored in psychology, where Tversky and Kahneman (1981) showed that changing the formulation of a problem affects the choices of people. Communicative frames deal with salience, i.e., to promote certain pieces of information, which influences the reader's perception (Entman, 1993). Scheufele (1999) further distinguishes these communicative frames in a four-cell typology of *media vs. individual frames* as *dependent vs. independent variables*. For instance, the frames of Entman (1993) belong to *media frames as independent variables* due to their concern with the influence on the audience perception.

In this thesis, my focus lies on media frames as a form of communicative frames. Nevertheless, all three types are present in the various publications. For instance, I also use PropBank (Kingsbury & Palmer, 2002) for extracting and analyzing semantic frames. Besides, I relate framing to other social theories, such as the moral foundations theory (Graham et al., 2013).

Relevance for: C2-C4, **C5**, C6-C8

Except for one publication, where we use sentiment as a proxy for simplicity, the framing theory is an essential part of all core publications. Specifically, the conceptualizations for framing detection depend on the theory with a focus on one or more types of frames being analyzed.

2.2 Natural Language Processing for Text Understanding

From a methodological perspective, I draw from research findings from the domain of natural language understanding (NLU). NLU is a subfield of natural language processing (NLP) concerned with capturing the meaning, while natural language generation deals with the opposite of verbalization (Pilehvar & Camacho-Collados, 2020). Both areas tend to focus on text data and can go hand-in-hand, e.g., by first analyzing a text before generating a response.

For the thesis, I mainly focus on NLU for textual data by applying embeddings, e.g., as a representation of text like Word2Vec (Mikolov et al., 2013), and Transformers (Vaswani et al., 2017). Embeddings and Transformers are two cornerstones of state-of-the-art NLP solutions and are also used for content analysis. In the following subsections, I first discuss the former two (in Subsection 2.2.1) before considering various content analysis and text mining approaches (in Subsection 2.2.2).

Relevance for: C1-C5, **C6**, C7, C8

NLU is an essential part of computational framing analysis, which is why I have resorted to NLU (and partially NLG) techniques for all core publications.

2.2.1 Embeddings and Transformers

In this section, I jointly discuss embeddings and Transformers, as their application and history are interwoven. To that end, I clarify the use of the terminology of the two terms and their evolution.

Embeddings, as used in this thesis, describe distributed semantic representations based on low-dimensional vectors (Pilehvar & Camacho-Collados, 2020). Before embeddings, distributed representations of documents based on the vector space model (Salton et al., 1975) were common, e.g., for information retrieval. Word embeddings, such as Word2Vec (Mikolov et al., 2013) and Glove (Pennington et al., 2014), expanded the idea of vector-based representations to words grounded in the distributional hypothesis, such that words appearing in similar contexts should have similar vector representations. Hence, each word gets assigned a particular static representation. Contextualized representations, such as ELMo (short for Embeddings from Language Models; Peters et al., 2018), create dynamic representations by considering the context that words appear in, and, as the name suggests, might rely on language models.

Language models can be defined by functions based on the probability measure of strings created from a vocabulary, with the simplest form being the unigram model as a result of choosing terms independently (Manning et al., 2008). We can incorporate sequential information by conditioning on the previous terms as a generalization for n-gram models (e.g., bigram if just the single previous term is used). This gave rise to various deep learning methods for language modeling, such as RNNs (Rumelhart et al., 1985), LSTMs (Hochreiter & Schmidhuber, 1997), and GRUs (Cho et al., 2014). More recent models rely on the attention mechanism (Bahdanau et al., 2014) to dynamically focus on the relevant parts of a text, which was originally used for machine translation. In a similar vein, the Transformer architecture (Vaswani et al., 2017), following an encoder-decoder approach, uses this idea as a fundamental component in both parts. Various subsequent architectures and models have since been created, with BERT (Devlin et al., 2018) and variants such as RoBERTa (Y. Liu et al., 2019) being the most noteworthy encoder models, which use masked language modeling (e.g., via randomly masking pieces). These encoder models convert pieces of text (i.e., tokens) to contextualized embeddings that can afterward be used for downstream tasks like classification. One approach to using these embeddings for text representations is to pool them (e.g., by averaging), and the similarity of such representations can even be used as training objectives for better representations, such as in SBERT (Reimers & Gurevych, 2019). Besides numerical representations, we can use encoder-decoder models like BART (Lewis et al., 2020) to convert text from one domain to another. As an example, a text can be converted to a graph representation, such as abstract meaning representations (Banarescu et al., 2013), which simplifies text analysis.

Note that I focus on text understanding and thus do not discuss decoder-only models like GPT (Radford et al., 2018) and its variants. Nevertheless, the pre-training aspect is still relevant, and fine-tuning was popularized by works like ULMFiT (Howard & Ruder, 2018). Additionally, several other training paradigms like knowledge distillation (e.g., in distillBERT Sanh et al., 2019 or MiniLM W. Wang et al., 2020) and contrastive learning have become state-of-the-art. The former aims to create student models with similar performance but smaller size than the teacher model, while the latter optimizes the embedding space for the alignment of positive samples and uniformity in the embedding space (T. Wang & Isola, 2020). Several solutions aim at transfer learning in low-data scenarios, such as libraries like SetFit (Tunstall et al., 2022) by contrastively optimizing the body with a classification head for few-shot learning or a textual entailment approach for zero-shot learning (Yin et al., 2019).

Relevance for: C2, **C3**, C4-C6

I use various of the discussed concepts and models for my doctoral research. Concretely, I use word embeddings based on Word2Vec, encoder models like MiniLM, encoder-decoder models like BART, training methodologies like contrastive learning, and complete solutions like SetFiT for framing detection.

2.2.2 Content Analysis and Text Mining Approaches

In this section, I focus on computational approaches for content analysis and mining of textual data, whereas other approaches (e.g., Mayring et al., 2004, from the social sciences) can be part of the individual publications but are beyond the scope of this section. In this context, sentiment analysis (Pang, Lee, et al., 2008) and topic modeling (e.g., LDA by Blei et al., 2003 being a notable example) are two of the most common techniques. In the former, a polarity score/label is assigned to individual texts or parts of it and is sometimes equated to opinion mining, while the latter identifies suitable topics based on commonalities in the data. Besides, argument mining for automatic reasoning, i.e., a premise leading to a conclusion, is closely related to other content analysis approaches (Lawrence & Reed, 2020). For instance, both sentiment and topic models can support the task at hand, which suffers from a shortage of annotated data (Lawrence & Reed, 2020). Furthermore, texts can be analyzed by their writing style, i.e., stylometry, which allows for differentiating between authors, among other tasks (Neal et al., 2017). In a similar vein, Potthast et al. (2018) show that style features enable distinguishing certain types of texts (e.g., hyperpartisan vs. mainstream news) but have shortcomings in other areas (e.g., for fake news detection).

Both embeddings and Transformers have also heavily influenced this research area. Hence, embeddings can be used to represent documents and the content within, with subsequent techniques like clustering or visualizations being applied. For instance, embedding spaces can be further reduced for plotting using well-established approaches like PCA, t-SNE (Van der Maaten & Hinton, 2008), or UMAP (McInnes et al., 2018). Approaches like BERTopic (Grootendorst, 2022) can even use Transformer directly. Besides, entities within a text can be analyzed using semantic role labeling (SRL), or the complete text can be converted to a graph representation, e.g., abstract meaning representation (AMR; Banarescu et al., 2013). Moreover, probabilistic measures like the log-odds-ratio can be used to identify important words (Monroe et al., 2008).

Relevance for: C1, **C4**, C5, C6

During my doctoral research, I have used various approaches to content analysis, from simple ones like sentiment analysis over the visualization of embedding space to abstract meaning representations for text mining. Furthermore, extracting the framing of content using the tools established in this thesis can be seen as another type of content analysis similar to sentiment analysis.

2.3 Computational Framing Research

There is a wide range of approaches used and frames analyzed in computational framing analysis (see Ali and Hassan, 2022 for an overview). The researched frames often relate to polarized topics such as war (Wicke & Bolognesi, 2020), terrorism (Demszky et al., 2019), morality (Mokhberian et al., 2020), or blame (Shurafa et al., 2020). Accordingly, frames are typically equated to topics and deviate from the established definitions in social sciences (Ali & Hassan, 2022). Instead, framing can be defined as “how” a text is presented rather than “what” is apparent (Ali & Hassan, 2022). In terms of approaches, the range spans from topic models (e.g., with LDA in DiMaggio et al., 2013), neural networks (e.g., BERT in S. Liu et al., 2019), the FrameAxis approach (Kwak et al., 2021), and analyzing semantic relations (Jing & Ahn, 2021). Approaches for modeling frames often overlap with other areas, such as computationally argumentation (Ajjour et al., 2019). There are only a few framing datasets with annotations available for computational framing research, with the media frame corpus (Card et al., 2015) and gun violence frame corpus (S. Liu et al., 2019) being two prominent examples. Furthermore, the SemEval 2023 Task 3 SubTask 2 (Piskorski et al., 2023) also deals with framing in both few- and zero-shot scenarios. In particular, the subtask aims to identify multi-class, multi-label media frames from multilingual news articles where the number of samples per class and language is imbalanced.

Besides the already mentioned works, several others resemble my research in certain aspects. Many SemEval teams have similar solutions to ours (Reiter-Haas et al., 2023a), the two top-performing solutions (i.e., Wu et al., 2023 and Liao et al., 2023) used Transformers with either a pre-training procedure or a contrastive loss, respectively. Opitz and Frank (2022) combine AMR with embeddings for improved interpretability, while Bonial et al. (2020) use AMR for semantic matching on COVID-19. Bhatia et al. (2021) also released an open-source tool for computational framing analysis.

Relevance for: **C2**, C3-C8

The present thesis is related to and advances computational framing research in several aspects (I depict the basis of each approach in **bold**, the frame type in *italics*, and underlined the topics). From a methodological perspective, I developed a **multilingual contrastive learning** model for *media frame* prediction on a variety of topics like the Ukraine war, analyzed the *moral frames* in political discussion on social media using **FrameAxis**, and explored the *semantic frames* of health narratives with **abstract meaning representations**. Besides, I demonstrated the usefulness of a **multi-perspective** approach to the gun violence frame corpus with all three types on the developed open-source tool. Similarly, I investigated the influence of framing on online information behavior, which has been a mostly unexplored research area so far.

Simple is better than complex.
Complex is better than complicated.

*(Excerpt from PEP 20 –
The Zen of Python)*

3 Publications

The chapter is split into three parts. First, a brief analysis of the publication embedding space and their relations is provided. Then, the full list of publications is provided with the corresponding description and contribution statement. Finally, the core publications are included in the thesis, each within their own corresponding section.

3.1 Visualization of Publication Embeddings

For a better understanding of the publications and how they are related, I provide a brief analysis using BERTopic (Grootendorst, 2022)[1]. The embedding space of the core and supplementary publications based on their abstract is shown in Figure 3.1.

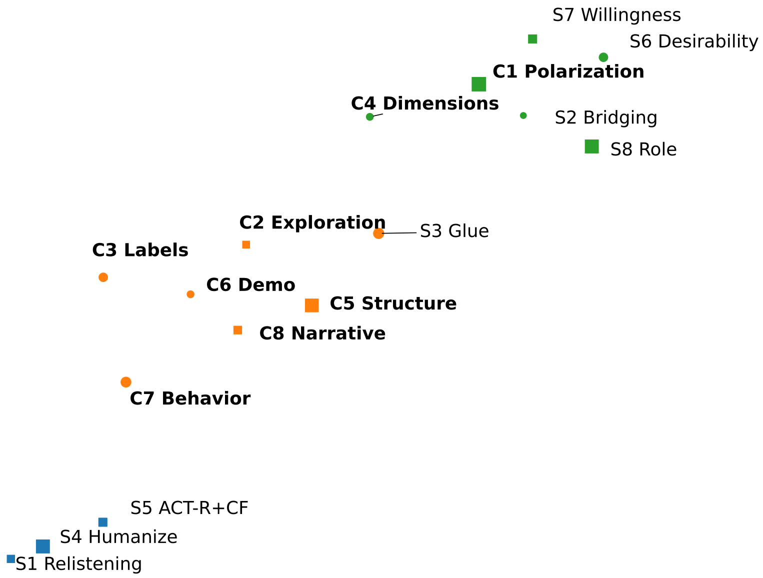


Figure 3.1: Positioning of the research. The plot consists of UMAP reduced, HDBScan clustered embeddings of paper abstracts. The research forms three main clusters that can be summarized as (i, orange) framing research, (ii, green) survey and polarization research, and (iii, blue) psychology-informed recommender systems and behavior modeling research. In the broader sense, these three clusters align with the three main communities of my research, i.e., (i) natural language processing, (ii) computational social sciences, and (iii) information retrieval and web science.

After initial observation, I empirically set the seed words as "polarization", "framing", and "behavior" to guide the process. I scale the point of each contribution by their magnitude (approximated using of total number of pages, neglecting the difference

[1] Following the practices of open science, the code is publicly available at: https://iseratho.github.io/thesis/phd-thesis-bertopic.ipynb

between single and double-column format). I mark the core contribution in bolt and use a square rather than a dot for the top three most representative documents per cluster. The word lists for the three topics are as follows:

Framing: framing, analysis, frame, frames, content, media, news, detection, narrative, using
Polarization: polarization, social, survey, data, measures, media, opinions, covid, twitter, 19
Behavior: behavior, music, act, tracks, memory, user, relistening, human, models, model

3.2 Publication Details

The following table provides a detailed list of publications (both core and supplementary), including a brief description and the contribution statements, both in free text for all collaborators and my own contribution according to CRediT (Allen et al., 2014):

Table 3.1: Table containing the list of publications as part of the PhD. Each publication is accompanied by a brief description and core contributions.

Key	Citation + Description + Contribution
	Core Publications
C1	**Reiter-Haas, M.***, Klösch, B.*, Hadler, M., & Lex, E. (2023). Polarization of Opinions on COVID-19 Measures: Integrating Twitter and Survey Data. *Social Science Computer Review 41 (5), 1811-1835.* **About Polarization.** We compare the polarization in Twitter data, survey data, and a small integrated dataset created from survey users who consented to data collection and provided their Twitter handles. We use an agreement scale in the survey data and sentiment scores in the Twitter data as a proxy for our analysis. We find a similar tendency regarding polarization on several COVID-19 prevention measures. **Contribution Statement.** I was responsible for the Twitter analysis and statistical measures. Beate Klösch was responsible for the survey analysis and qualitative analysis of the tweets. All authors were involved in the design, discussion, and interpretation of the results, as well as in the writing process. **My CRediT.** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization
C2	**Reiter-Haas, M.** (2023). Exploration of Framing Biases in Polarized Online Content Consumption. In *Companion Proceedings of the ACM Web Conference 2023, 560–564. Presented at the Austin, TX, USA.* **About Exploration.** The paper provides an introductory analysis of framing in relation to research questions that form the basis of the doctoral work. Hence, it lays the foundation of the three types of framing analysis used throughout this thesis. Subsequently, it explores the trade-offs between the framing labels, framing dimensions, and framing structure.

Contribution Statement. I was responsible for every aspect of the paper comprising mainly design, plots, analysis, and writing.

My CRediT. Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing - Original Draft, Writing - Review & Editing, Visualization

C3 **Reiter-Haas, M.***, Ertl, A.*, Innerebner, K., & Lex, E. (2023). mCPT at SemEval-2023 Task 3: Multilingual Label-Aware Contrastive Pre-Training of Transformers for Few- and Zero-shot Framing Detection. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 941–949, Toronto, Canada. Association for Computational Linguistics.*

About Labels. The paper is our contribution to a shared task on framing detection. We created a system using contrastive learning together with a multi-stage pretraining and fine-tuning procedure. We achieved notable placements in the nine languages while submitting the best-performing contribution for Spanish zero-shot framing detection.

Contribution Statement. I had the idea of participating, conceptualizing, and conducting the initial analysis and led the group research. Besides, I was responsible for the SetFit baseline, the open-source repository, and communication with the organizers. Alexander Ertl created the main experimental setup, as well as refined the idea by selecting the loss function and developing the multi-stage training procedure. Kevin Innerebner supported the programming and was responsible for the embedding space analysis. Elisabeth Lex supervised the student team and coordinated the research group. All authors participated in discussions and the writing of the paper.

My CRediT. Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Writing - Original Draft, Writing - Review & Editing, Supervision, Project administration

C4 **Reiter-Haas, M.**, Kopeinik, S., & Lex, E. (2021). Studying Moral-based Differences in the Framing of Political Tweets. In *Proceedings of the International AAAI Conference on Web and Social Media, 15(1), 1085-1089.*

About Dimensions. We applied the FrameAxis approach for unsupervised moral framing detection with politically associated accounts on Twitter. We find that followers of Austrian politicians tend to morally frame the COVID-19 discourse. For instance, followers associated with the ruling conservative party use care as a framing device, which is also represented in the government spread prevention campaign.

Contribution Statement. I set up the design and experiment, as well as conducted the analysis. Simone Kopeinik provided theoretical input, while Elisabeth Lex advised the development of the experiment. All authors were involved in discussing and interpreting the results, as well as writing the paper.

My CRediT. Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization

C5 **Reiter-Haas, M.**, Klösch, B., Hadler, M., & Lex, E. (2024). Framing Analysis of Health-Related Narratives using Semantic Graphs: Conspiracy versus Mainstream Media. *arXiv preprint arXiv:2401.10030.*

About Structure. We designed an explorative approach for the framing detection of health narratives using abstract meaning representations. By comparing mainstream and conspiracy media, we find well-established differences in the framing, such as a science-oriented framing compared to a belief-oriented framing. Moreover, we also identified more subtle differences, like a stronger focus on immediacy in conspiracy media.

Contribution Statement. I designed the experiment and implemented the approach, as well as took the lead in writing. All authors contributed in writing, surveying related work, discussion, and interpretation of the results.

My CRediT. Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization

C6 **Reiter-Haas, M.**, Klösch, B., Hadler, M., & Lex, E. (2024). FrameFinder: Explorative Multi-Perspective Framing Extraction from News Headlines. In *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval (pp. 381-385).*

About Demo. We created a tool for multi-perspective framing extraction and analysis. We demonstrated the tool on a well-established framing corpus and uncovered novel insights, such as health being noticeably absent as a frame regarding gun violence. Besides, we discussed the implications of the tool for social science research and its inclusion in information systems.

Contribution Statement. I created the library and online demo, as well as designed and conducted the experiments. Beate Klösch and Markus Hadler tested the tool and discussed its relation to social science research. Elisabeth Lex supervised the project. All authors discussed, reviewed, and edited the manuscript.

My CRediT. Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Writing - Original Draft, Writing - Review & Editing, Visualization

C7 **Reiter-Haas, M.** & Lex, E. (2024). The Framing Loop: Do Users Repeatedly Read Similar Framed News Online? In *Proceedings of the 7th HUMANIZE Workshop.*

About Behavior. We analyze user behavior concerning media, moral, and semantic frames in an information system dataset. Specifically, we investigate repeat consumption and users' viewpoint diversity concerning frames. We find that users repeatedly consume similar frames, which information systems can counteract.

Contribution Statement. I designed and conducted the experiment, analyzed the results, and wrote the draft of the manuscript under the supervision of Elisabeth Lex. Both authors discussed the results and edited the paper.

My CRediT. Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization

C8 **Reiter-Haas, M.**, Klösch, B., Hadler, M., & Lex, E. (2024). Computational Narrative Framing: Towards Identifying Frames through Contrasting the Evolution of Narration. In *Proceedings of the Text2Story'24 Workshop, Glasgow (Scotland), 24-March-2024.*

About Narrative. In the position paper, we discussed the relation of the research in computational framing analysis and computational narrative understanding. We exemplify how the investigation of the temporal evolution in narrative structure belonging to competing narratives could improve the understanding of framing in climate change discourse. Hence, we argue for a convergence of their research directions.

Contribution Statement. I wrote the manuscript based on insights gained from discussions with Markus Hadler, Beate Klösch, and Elisabeth Lex. All authors reviewed and approved the manuscript.

My CRediT. Conceptualization, Methodology, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization

Supplementary Publications

S1 **Reiter-Haas, M.**, Parada-Cabaleiro, E., Schedl, M., Motamedi, E., Tkalcic, M., & Lex, E. (2021, September). Predicting music relistening behavior using the ACT-R framework. In *Proceedings of the 15th ACM Conference on Recommender Systems (pp. 702-707).*

About Relistening. We predict the music-relistening behavior of users using a psychology-inspired approach for recommender systems. In this article, the best results are achieved by considering a combination of recency and frequency, co-occurrences, and familiarity. Besides, we also consider the similarity of tracks and randomness in relistening behavior, as well as analyze the power law distribution of relistening patterns.

Contribution Statement. I was responsible for the setup and main experiments, as well as the lead in writing the manuscript. Emilia Parada-Cabaleiro created the content representations. Markus Schedl was responsible for the dataset. Both Elham Motamedi and Marko Tkalcic provided theoretical support. Elisabeth Lex coordinated and supervised the research. All authors were involved in designing the idea, discussing the results, and writing the manuscript.

My CRediT. Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Writing - Original Draft, Writing - Review & Editing, Visualization, Project administration

S2 **Reiter-Haas, M.**, Klösch, B., Hadler, M., & Lex, E. (2020). Bridging the Gap of Polarization in Public Opinion on Misinformed Topics. *Challenging Misinformation: Exploring Limits and Approaches, workshop co-located with Social Informatics'20.*

About Bridging. We provide an initial data analysis on survey data and social media data for subsequent research. Our dataset is based on a survey conducted in the German-speaking DACH region on views about COVID-19 and climate change. In this study, we find that Twitter users are less likely to believe in the non-natural origin of COVID-19 compared to the overall sample.

Contribution Statement. I was responsible for the social media analysis and led the writing of the draft. Beate Klösch was responsible for the data preparation and survey analysis. All authors were involved in the design and discussion of the research, as well as the writing process.

My CRediT. Conceptualization, Software, Validation, Formal analysis, Investigation, Writing - Original Draft, Writing - Review & Editing

S3 Hadler, M., Ertl, A., Klösch, B., **Reiter-Haas, M.**, & Lex, E. (in Review). The climate gluing protests: Analyzing their development and framing in media since 1986 using sentiment analyses and frame detection models.

About Glue. The paper tracks the evolution of gluing protests with a focus on climate in media. We employ sentiment analysis and framing detection models for the task. To that end, we find mostly negative media reports and a limited prevalence of prognostic frames.

Contribution Statement. I mainly advised Alexander Ertl in using computational tools for framing detection and sentiment analysis. Markus Hadler organized the data collection, conducted the primary analysis, and wrote the initial draft, supported by Beate Klösch. Elisabeth Lex coordinated and supervised the research. All authors discussed and reviewed the manuscript.

My CRediT. Methodology, Validation, Formal analysis, Investigation, Data Curation, Writing - Review & Editing, Supervision, Funding acquisition

S4 Kowald, D., **Reiter-Haas, M.**, Kopeinik, S., Schedl, M., & Lex, E.(2024). Transparent Music Preference Modeling and Recommendation with a Model of Human Memory Theory. In *A Human-Centered Perspective of Intelligent Personalized Environments and Systems (pp. 113-136). Cham: Springer Nature Switzerland*

About Humanize. The book chapter discusses how a model of human memory theory can provide transparent recommendations for music genre preferences. The focus lies on two components of that model which are also supported by the underlying theory. It also discusses potential extensions. The research shows the efficacy of the approach on three user groups of low, medium, and high mainstreaminess.

Contribution Statement. I was responsible for the section of the model extensions. Dominik Kowald had the writing lead, based on his prior experiments together with Markus Schedl and Elisabeth Lex. Simone Kopeinik was responsible for the theoretical underpinnings. All authors discussed, edited, and verified the final version of the manuscript.

My CRediT. Methodology, Investigation, Writing - Original Draft, Writing - Review & Editing

S5 Moscati, M., Wallmann, C., **Reiter-Haas, M.**, Kowald, D., Lex, E., & Schedl, M. (2023, September). Integrating the ACT-R framework with collaborative filtering for explainable sequential music recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems (pp. 840-847).*

About ACT-R+CF. We create four hybrid recommendation algorithms combining the cognitive architecture ACT-R with collaborative filtering. The combination enables better explainability in sequential music recommendations while also improving their novelty and diversity. Moreover, we investigated the salience of components (e.g., current vibes), which could be used to tune the user's needs in future work.

Contribution Statement. I provided technical knowledge for using ACT-R in music relistening modeling. Marta Moscati had the lead in the design of the experiment and in the writing of the manuscript. Christian Wallmann was responsible for the implementation of the algorithms. All authors discussed the results and reviewed and edited the manuscript.

My CRediT. Methodology, Validation, Writing - Review & Editing, Project administration

S6 Klösch, B., Hadler, M., **Reiter-Haas, M.**, & Lex, E. (2022). Social Desirability and the Willingness to Provide Social Media Accounts in Surveys. The Case of Environmental Attitudes. *4th International Conference on Advanced Research Methods and Analytics.*

About Desirability. The paper investigates whether social desirability influences the willingness to provide social media accounts in surveys and whether the opinions expressed on social media are congruent with the survey data. This article shows that Facebook users who oppose environmental measures oppose sharing their accounts, while this effect is absent for Twitter users, who also show congruent tendencies with their survey responses.

Contribution Statement. I was responsible for creating the original plots and verifying the data. Beate Klösch had the lead in writing the manuscript. All authors reviewed and discussed the manuscript.

My CRediT. Methodology, Software, Validation, Investigation, Writing - Review & Editing, Visualization

S7 Hadler, M., Klösch, B., **Reiter-Haas, M.**, & Lex, E. (2022). Combining Survey and Social Media Data: Respondents' Opinions on COVID-19 Measures and Their Willingness to Provide Their Social Media Account Information. *Frontiers in Sociology, 7.*

About Willingness. The paper investigates sampling bias in the willingness to share social media accounts due to specific views for survey research. In this study, we find that survey respondents with more positive attitudes towards COVID-19 measures are more willing to share the account handles.

Contribution Statement. I was responsible for accessing the accounts and providing insights into the process. Markus Hadler drafted the paper, which all authors then discussed and reviewed.

My CRediT. Methodology, Software, Validation, Investigation, Writing - Review & Editing

S8 Klösch, B., Hadler, M., **Reiter-Haas, M.**, & Lex, E. (2023). Polarized opinions on Covid-19 and environmental policy measures. The role of social media use and personal concerns in German-speaking countries. *Innovation: The European Journal of Social Science Research, 1-24.*

About Role. The paper compares the polarization patterns between COVID-19 and environmental measures on different platforms while also considering the free-rider problem as an explanation for a one-sided polarization. We also observe divergent and generational effects, e.g., that older generations tend to support COVID-19 measures more strongly.

Contribution Statement. I provided insights for discussing the platforms. Beate Klösch wrote the manuscript and conducted the experiments. All authors reviewed and approved the manuscript.

My CRediT. Writing - Review & Editing

3.3 Core Publications

The following pages include the core publications of the thesis. The order of the publication (see also Figure 1.3) should enable a natural flow as follows:

C1 We establish that polarization, measured by agreement in survey data, is similar to sentiment in social data regarding COVID-19 prevention measures. In this study, we use sentiment as a proxy for opinion for comparison's sake, which nicely bridges two strands of research. This study can be seen as a rudimentary form of frame production, i.e., whether users use positive or negative framing when writing their tweets. However, we also suggest that more advanced NLP methods would be beneficial. Hence, it serves as a nice motivation for the importance of studying advanced methods of content analysis such as framing.

C2 We concretize the pressing research questions in understanding the framing in textual data, both for its detection and its interplay with user behavior. This paper establishes the three kinds of framing detection that we explore in this thesis, i.e., frame labels, frame dimensions, and frame structure. Similarly, it identifies framing behavior as an unexplored research direction.

C3 We present our contribution to the SemEval shared task, which follows the classical approach of frame label detection. Thus, we created and trained a neural network for predicting the frame labels of the annotated corpus. Nevertheless, a special loss function and training procedure were used to exploit label similarities, thus improving the prediction performance. Hence, this experiment already shows the issues in framing detection due to data sparsity that become even more challenging in the other types of frame detection.

C4 We consider moral dimensions, which can be antagonistically defined, as frame dimensions. The frame dimensions can be used for prediction purposes but also for understanding differences between groups. In this study, we compared the established characteristics of Republicans and Democrats, but also the differences in the Austrian multi-party system regarding COVID-19. Frame dimensions lie between frame labels that have rigorous validation and frame structure that allows for a high amount of exploration.

C5 We explore health narratives using frame structures based on AMR graphs. We show that our approach allows us to find differences between how mainstream and conspiracy media frame their articles that are not established in existing computational frame conceptualizations. This study marks the explorative end of the developed frame detection approaches.

C6 We demonstrate how the detection of frame labels, frame dimensions, and frame structure complement each other and allow for a more holistic frame understanding of gun violence. We open-source the tool for subsequent frame detection research.

C7 We consider the three types of framing in user behavior in information systems. Specifically, we investigate the frame consumption behavior regarding repeat consumption and viewpoint diversity. We find patterns that suggest that framing, indeed, plays a vital role and should be considered as future work.

C8 We discuss the overlap between narrative understanding and framing analysis. The study identifies that considering complete narratives is a (potentially even more explorative) direction for framing detection.

An overview of the used data is provided in Table 3.2.

	Topic(s)	Data Source(s)	Frame Type(s)/Concept(s)	Language(s)	Main Method	# Samples	
C1	COVID-19 prevention measures	survey + Twitter	agreement + sentiment (proxy)	German	Statistics (bimodality)	98,549	combined tweets and survey responses
C2	polarized topics in general	web content	all three/exploration	- (English)	Transformers (various)	-	(example)
C3	various (e.g., COVID-19, climate change, migration, war)	SemEval'23 Task 3 (annotated news and web articles)	labels/media frames	9 languages (6 few-, 3 zero-shot)	Supervised Contrastive Pretraining	2049	train, dev, and test
C4	politics, COVID-19	Twitter	dimensions/moral frames	English + German	FrameAxis	1,410,403	US and Austrian tweets
C5	health (i.e., COVID-19, diseases, pharmacology)	LOCO (news websites of mainstream and conspiracy)	structure/semantic frames	English	Abstract Meaning Representations	33,648	documents
C6	gun violence in the US	GVFC (news headlines)	all three/(media, moral, semantic)	English	FrameFinder (Transformers)	2990	headlines
C7	news in general	MIND-small (news recommender dataset)	all three/(media, moral, semantic)	English	Behavior Analysis (FrameFinder)	6,690,694	sequences
C8	climate change	news article	narrative (semantic) frame	- (English)	Theoretical	-	(example)

Table 3.2: Data Overview

Check for updates

Article

Social Science Computer Review
2023, Vol. 41(5) 1811–1835
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/08944393221087662
journals.sagepub.com/home/ssc
Sage

Polarization of Opinions on COVID-19 Measures: Integrating Twitter and Survey Data

Markus Reiter-Haas[1,†], Beate Klösch[2,†], Markus Hadler[2], and Elisabeth Lex[1]

Abstract

Polarization of public opinion is a major issue for societies, as high levels can promote adverse effects such as hostility. The present paper focuses on the polarization of opinions regarding COVID-19 prevention measures in survey data and on Twitter in the German-speaking regions of Germany, Austria, and Switzerland. The level of polarization is measured by dispersion and bimodality in the opinions based on the sentiment in Twitter data and the agreement in the survey data. Our paper, however, goes beyond existing research as we consider data from both sources separately and comparatively. For this purpose, we matched individuals' survey responses and tweets for those respondents who shared their Twitter account information. The analyses show that vaccination is more polarizing compared to mask wearing and contact tracing in both sources, that polarization of opinions is more pronounced in the survey data compared to the Twitter data, but also that individuals' opinions about the COVID-19 measures are consistent in both sources. We believe our findings will provide valuable insights for integrating survey data and Twitter data to investigate opinion polarization.

Keywords

interdisciplinary research, opinion polarization, surveys, twitter, social media, integrating data sources, COVID-19 measures

[1]Graz University of Technology, Graz, Austria
[2]University of Graz, Graz, Austria

[†]These authors have contributed equally to this work.

Corresponding Authors:
Markus Reiter-Haas, Institute of Interactive Systems and Data Science, Graz University of Technology, Inffeldgasse 13/V, Graz 8010, Austria.
Email: reiter-haas@tugraz.at

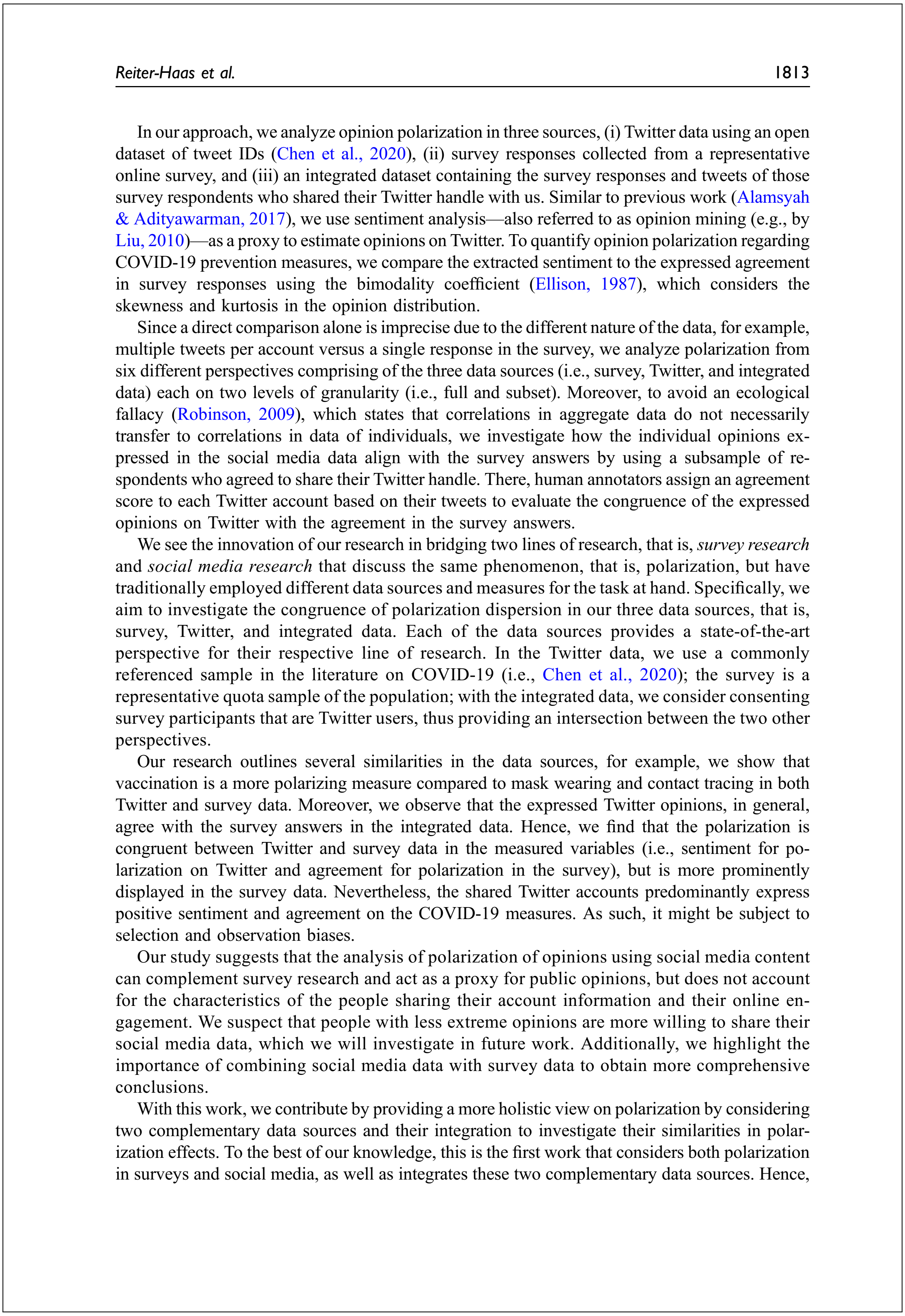
Elisabeth Lex, Institute of Interactive Systems and Data Science, Graz University of Technology, Inffeldgasse 13/V, 8010 Graz.
Email: elisabeth.lex@tugraz.at

Opinion polarization is a major issue for a society as it leads to adverse effects such as the spread of misinformation (Bessi et al., 2015; Del Vicario et al., 2016). For instance, the opinions on COVID-19 are polarized, as people disagree whether the virus is of natural origin or was created artificially (Reiter-Haas et al., 2020). Opinion polarization is characterized by extreme positions (Stroud, 2010) and can be defined as a state in terms of dispersion and modality of opinions (DiMaggio et al., 1996). Neither is a high dispersion of opinions negative (e.g., personal preferences like opinions on taste or weather) nor is a bimodality in itself harmful (e.g., whether people prefer cats or dogs). Even high polarization on both dispersion and bimodality might not be harmful, for example, the perception of whether a dress is black and blue or white and gold. Nevertheless, research has shown that polarization in terms of sentiment and emotion, that is, affective polarization, can lead to hostility in societies, for example, alongside partisanship (Tucker et al., 2018). As a consequence, research on polarization (e.g., Baldassarri & Bearman, 2007; Borge-Holthoefer et al., 2015; Conover et al., 2011; Fiorina & Abrams, 2008; Garimella & Weber, 2017; J. Jiang et al., 2020) and related issues, such as group polarization (Sunstein, 1999), selective exposure (Knobloch-Westerwick & Meng, 2009), and echo chambers (Garrett, 2009), has been a longstanding research focus.

From a methodological perspective, polarization of public opinion over controversial topics has typically been analyzed via surveys (e.g., Bramson et al., 2017; Hetherington, 2001). In surveys, data about opinions and attitudes is primarily collected from a representative group of respondents to gain insights into the drivers of polarization. In addition, users increasingly exchange opinions and share their attitudes and beliefs via online social media platforms, making them an alternative source for public opinion. Thus, extensive research has been conducted on polarization in various online platforms using user-generated content and digital behavioral data (e.g., Adamic & Glance, 2005; An et al., 2013; Bakshy et al., 2015; Bessi et al., 2014; Conover et al., 2011; Darwish, 2019; Garcia et al., 2012).

While research on polarization at the intersection of surveys and online social media is still scarce, recent work has recognized the potential of linking social media and survey data to measure public opinion (Stier et al., 2020). Nevertheless, it is unclear if similar or different opinion dynamics can be observed in both sources. Moreover, Al Baghal et al. (2021) outline the asymmetry between survey and Twitter data, such as the differences in the quantity and information content, as well as its variability. Generally speaking, Twitter data is more abundant and provides longitudinal insights, whereas it typically lacks socio-demographic information and does not directly probe for the opinions of people, which is in turn provided by survey data. Hence, these two data sources are complementary to each other and taken together provide valuable insights into the opinions of people toward certain topics.

In this paper, we aim to study the relations of opinion polarization between survey responses and social media content with respect to three COVID-19 prevention measures, that is, vaccination, mask wearing, and contact tracing. Our study analyzes the polarization in the German-speaking DACH region (D-Germany, A-Austria, and CH-Switzerland) at the beginning of August 2020, when the first wave of COVID-19 was over, and Austria, Germany, and Switzerland were almost entirely open. Yet, in this period, the number of COVID-19 cases started to rise again due to holiday traffic, which kept the public engaged in discussions of the COVID-19 prevention measures analyzed in the present work. We focus on COVID-19 as this topic is highly polarized and an emerging societal issue (Allcott et al., 2020; Bruine de Bruin et al., 2020; Dohle et al., 2020; Hart et al., 2020). Its societal relevance is exemplified, for instance, in the rise of dark web marketplaces for medical products, for example, personal protective equipment and hydroxychloroquine, that were in short supply (Bracci et al., 2021). We deem the study of polarization on COVID-19-related topics as crucial since a high level of polarization can lead to biased reasoning in humans, which in turn may hinder public pandemic mitigation strategies (Van Bavel et al., 2020).



In our approach, we analyze opinion polarization in three sources, (i) Twitter data using an open dataset of tweet IDs (Chen et al., 2020), (ii) survey responses collected from a representative online survey, and (iii) an integrated dataset containing the survey responses and tweets of those survey respondents who shared their Twitter handle with us. Similar to previous work (Alamsyah & Adityawarman, 2017), we use sentiment analysis—also referred to as opinion mining (e.g., by Liu, 2010)—as a proxy to estimate opinions on Twitter. To quantify opinion polarization regarding COVID-19 prevention measures, we compare the extracted sentiment to the expressed agreement in survey responses using the bimodality coefficient (Ellison, 1987), which considers the skewness and kurtosis in the opinion distribution.

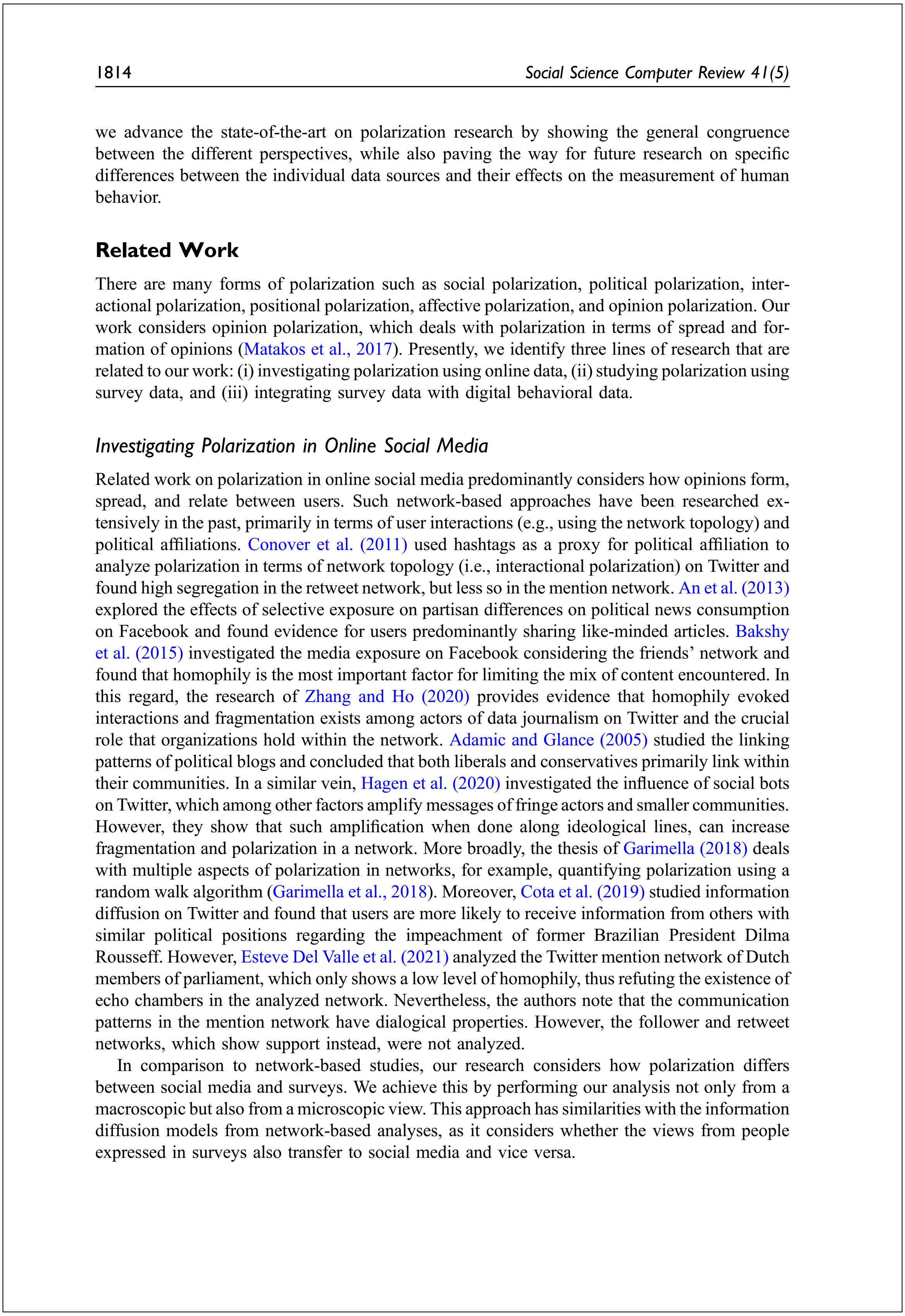
Since a direct comparison alone is imprecise due to the different nature of the data, for example, multiple tweets per account versus a single response in the survey, we analyze polarization from six different perspectives comprising of the three data sources (i.e., survey, Twitter, and integrated data) each on two levels of granularity (i.e., full and subset). Moreover, to avoid an ecological fallacy (Robinson, 2009), which states that correlations in aggregate data do not necessarily transfer to correlations in data of individuals, we investigate how the individual opinions expressed in the social media data align with the survey answers by using a subsample of respondents who agreed to share their Twitter handle. There, human annotators assign an agreement score to each Twitter account based on their tweets to evaluate the congruence of the expressed opinions on Twitter with the agreement in the survey answers.

We see the innovation of our research in bridging two lines of research, that is, *survey research* and *social media research* that discuss the same phenomenon, that is, polarization, but have traditionally employed different data sources and measures for the task at hand. Specifically, we aim to investigate the congruence of polarization dispersion in our three data sources, that is, survey, Twitter, and integrated data. Each of the data sources provides a state-of-the-art perspective for their respective line of research. In the Twitter data, we use a commonly referenced sample in the literature on COVID-19 (i.e., Chen et al., 2020); the survey is a representative quota sample of the population; with the integrated data, we consider consenting survey participants that are Twitter users, thus providing an intersection between the two other perspectives.

Our research outlines several similarities in the data sources, for example, we show that vaccination is a more polarizing measure compared to mask wearing and contact tracing in both Twitter and survey data. Moreover, we observe that the expressed Twitter opinions, in general, agree with the survey answers in the integrated data. Hence, we find that the polarization is congruent between Twitter and survey data in the measured variables (i.e., sentiment for polarization on Twitter and agreement for polarization in the survey), but is more prominently displayed in the survey data. Nevertheless, the shared Twitter accounts predominantly express positive sentiment and agreement on the COVID-19 measures. As such, it might be subject to selection and observation biases.

Our study suggests that the analysis of polarization of opinions using social media content can complement survey research and act as a proxy for public opinions, but does not account for the characteristics of the people sharing their account information and their online engagement. We suspect that people with less extreme opinions are more willing to share their social media data, which we will investigate in future work. Additionally, we highlight the importance of combining social media data with survey data to obtain more comprehensive conclusions.

With this work, we contribute by providing a more holistic view on polarization by considering two complementary data sources and their integration to investigate their similarities in polarization effects. To the best of our knowledge, this is the first work that considers both polarization in surveys and social media, as well as integrates these two complementary data sources. Hence,



we advance the state-of-the-art on polarization research by showing the general congruence between the different perspectives, while also paving the way for future research on specific differences between the individual data sources and their effects on the measurement of human behavior.

Related Work

There are many forms of polarization such as social polarization, political polarization, interactional polarization, positional polarization, affective polarization, and opinion polarization. Our work considers opinion polarization, which deals with polarization in terms of spread and formation of opinions (Matakos et al., 2017). Presently, we identify three lines of research that are related to our work: (i) investigating polarization using online data, (ii) studying polarization using survey data, and (iii) integrating survey data with digital behavioral data.

Investigating Polarization in Online Social Media

Related work on polarization in online social media predominantly considers how opinions form, spread, and relate between users. Such network-based approaches have been researched extensively in the past, primarily in terms of user interactions (e.g., using the network topology) and political affiliations. Conover et al. (2011) used hashtags as a proxy for political affiliation to analyze polarization in terms of network topology (i.e., interactional polarization) on Twitter and found high segregation in the retweet network, but less so in the mention network. An et al. (2013) explored the effects of selective exposure on partisan differences on political news consumption on Facebook and found evidence for users predominantly sharing like-minded articles. Bakshy et al. (2015) investigated the media exposure on Facebook considering the friends' network and found that homophily is the most important factor for limiting the mix of content encountered. In this regard, the research of Zhang and Ho (2020) provides evidence that homophily evoked interactions and fragmentation exists among actors of data journalism on Twitter and the crucial role that organizations hold within the network. Adamic and Glance (2005) studied the linking patterns of political blogs and concluded that both liberals and conservatives primarily link within their communities. In a similar vein, Hagen et al. (2020) investigated the influence of social bots on Twitter, which among other factors amplify messages of fringe actors and smaller communities. However, they show that such amplification when done along ideological lines, can increase fragmentation and polarization in a network. More broadly, the thesis of Garimella (2018) deals with multiple aspects of polarization in networks, for example, quantifying polarization using a random walk algorithm (Garimella et al., 2018). Moreover, Cota et al. (2019) studied information diffusion on Twitter and found that users are more likely to receive information from others with similar political positions regarding the impeachment of former Brazilian President Dilma Rousseff. However, Esteve Del Valle et al. (2021) analyzed the Twitter mention network of Dutch members of parliament, which only shows a low level of homophily, thus refuting the existence of echo chambers in the analyzed network. Nevertheless, the authors note that the communication patterns in the mention network have dialogical properties. However, the follower and retweet networks, which show support instead, were not analyzed.

In comparison to network-based studies, our research considers how polarization differs between social media and surveys. We achieve this by performing our analysis not only from a macroscopic but also from a microscopic view. This approach has similarities with the information diffusion models from network-based analyses, as it considers whether the views from people expressed in surveys also transfer to social media and vice versa.

Other studies focus more on polarization toward given events, which often contains a temporal dimension as the subject of analysis. Several recent works consider the effects of online conversations on polarization toward given events. Demszky et al. (2019) found that the reactions on Twitter to mass shootings are highly polarized and driven by partisan differences in their messages. Yarchi et al. (2020) conducted an over-time analysis of interactional, positional, and affective polarization on Facebook, Twitter, and WhatsApp on the killing of a Palestinian assailant by an Israeli soldier. They concluded that polarization cannot be seen as a unified phenomenon in social media, as the three platforms showed significant differences. J. Jiang et al. (2020) studied the political polarization of conversations on the COVID-19 pandemic on Twitter using Hashtags and found that partisanship correlates with government prevention measures. In a similar vein, our research focuses on affective polarization in tweets regarding the COVID-19 pandemic. We perform our study in the German-speaking Twitter data on three specific prevention measures, that is, vaccination, mask wearing, and contact tracing.

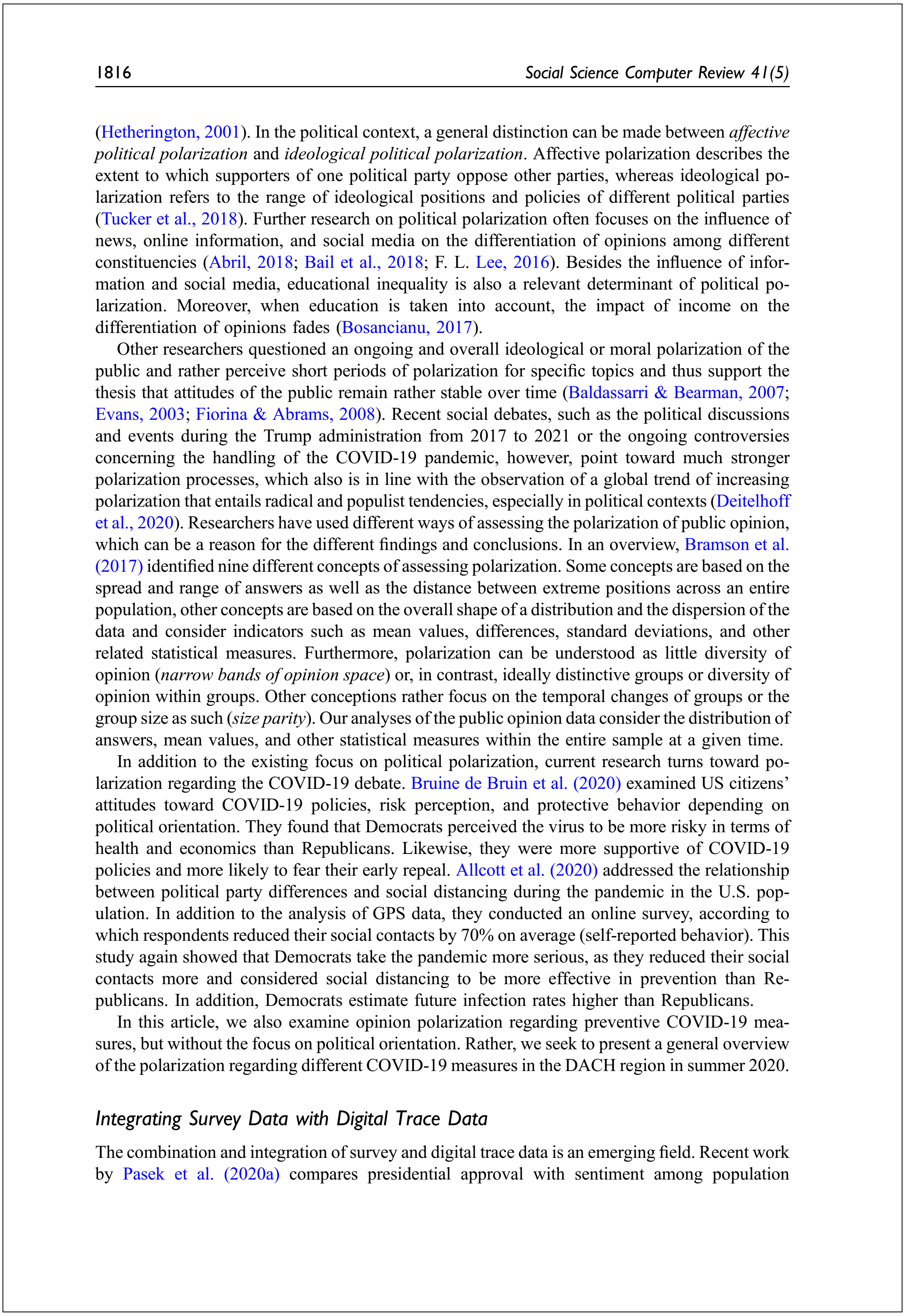
Finally, several approaches deal with the differences in content found online and often consider emotions or similar aspects as proxies to quantify opinions. They often consider *affective polarization*, that is, the emotional reaction of users. Garcia et al. (2012) quantify affective polarization in YouTube videos using likes and dislikes and performed sentiment analysis on comments. Pellert et al. (2020) modeled temporal dynamics of emotions on Facebook using emotional valence, that is, the positivity of emotions, and arousal, that is, the energy of the emotion. They find that both valence and arousal relax exponentially toward a baseline level after stimulation, which is relevant to estimate the actual impact of affect. Alternatively, sentiment can be used to determine affective polarization. Alamsyah and Adityawarman (2017) use the sentiment to label nodes in a network as positive, negative, or neutral for structural analysis in an Indonesian case study on Twitter regarding the reclamation of land through filling ocean waters and found that sentiment reliably captures the polarization process as far fewer conversations happen between the pro and counter reclamation nodes. Affect, that is, emotions and sentiment, can be used to estimate the opinions when considering opinion polarization. Moreover, sentiment analysis is even used interchangeably with the term opinion mining (Liu, 2010).

Similarly, we perform sentiment analysis as a proxy for opinions in the analysis of the polarization on Twitter. Additionally, we quantify the results using statistical measures such as the bimodality coefficient, which allows a comparison of those results with the survey responses.

Unlike many other studies (Adamic & Glance, 2005; Conover et al., 2011; Garcia et al., 2012; J. Jiang et al., 2020; Yarchi et al., 2020), we go beyond considering political polarization since we analyze the polarization in all Twitter users and tweets that express their opinions on the prevention measures regardless of their political affiliation. Moreover, instead of relying only on social media data, we concurrently conducted an online survey in the DACH region to contrast the results. Additionally, we combine a subset of the survey participants with their shared Twitter accounts to directly compare and discuss the differences between survey answers and their views expressed in social media. Thus, our approach of analyzing polarization in both surveys and social media also mitigates concerns of Sloan (2017), who showed that the demographic of Twitter is not representative of the population as a whole, and D. Lee et al. (2015), who showed that there is a discrepancy between the opinions expressed offline and online.

Studying Polarization in Survey Data

The polarization of the public has been considered extensively in the United States, with an emphasis on the divide between the two main parties and individuals that identify with them. Some researchers concluded that the polarization of the political elites contributes to the polarization of the mass, at least to ideological polarization of the identifiers with political parties



(Hetherington, 2001). In the political context, a general distinction can be made between *affective political polarization* and *ideological political polarization*. Affective polarization describes the extent to which supporters of one political party oppose other parties, whereas ideological polarization refers to the range of ideological positions and policies of different political parties (Tucker et al., 2018). Further research on political polarization often focuses on the influence of news, online information, and social media on the differentiation of opinions among different constituencies (Abril, 2018; Bail et al., 2018; F. L. Lee, 2016). Besides the influence of information and social media, educational inequality is also a relevant determinant of political polarization. Moreover, when education is taken into account, the impact of income on the differentiation of opinions fades (Bosancianu, 2017).

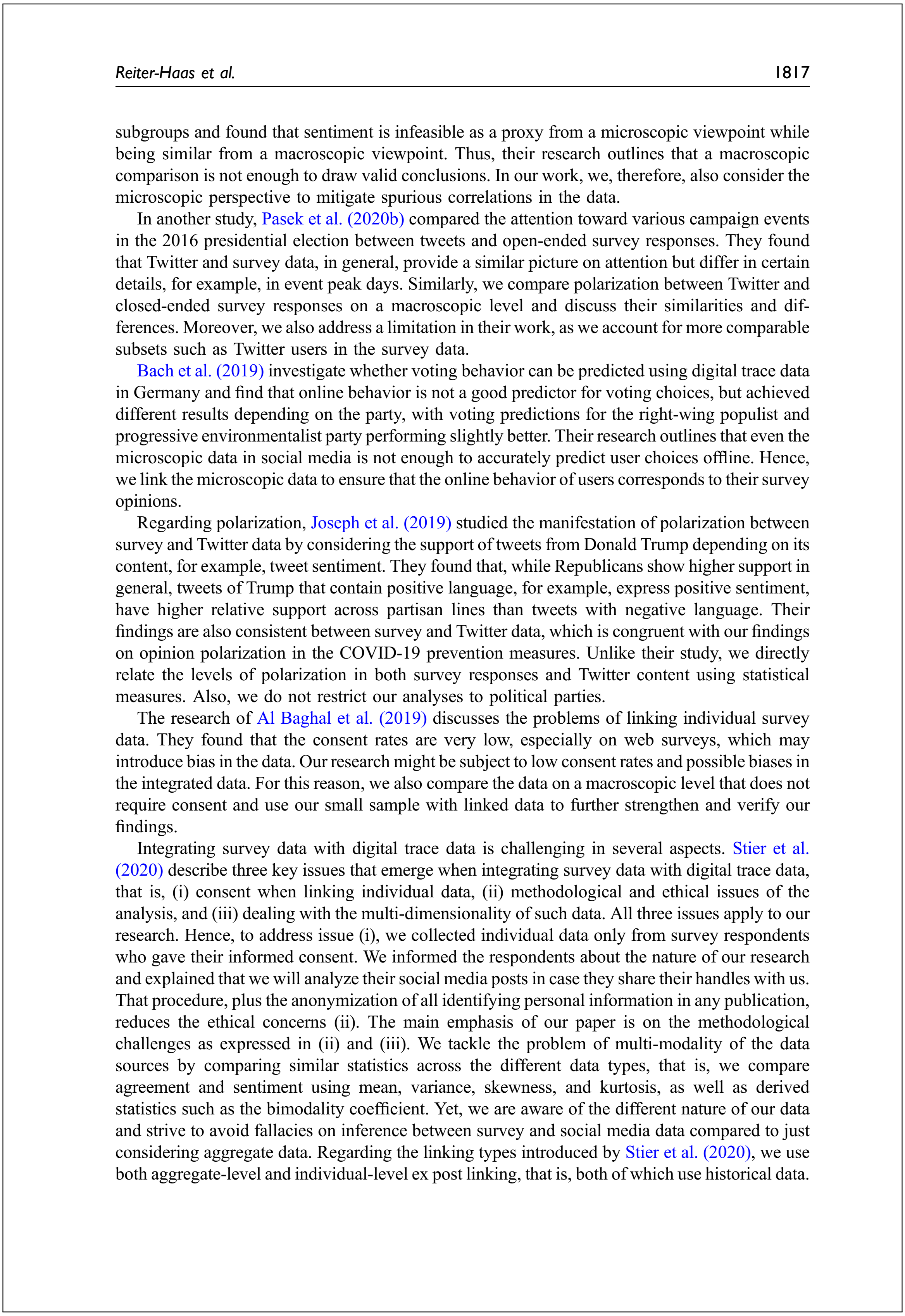
Other researchers questioned an ongoing and overall ideological or moral polarization of the public and rather perceive short periods of polarization for specific topics and thus support the thesis that attitudes of the public remain rather stable over time (Baldassarri & Bearman, 2007; Evans, 2003; Fiorina & Abrams, 2008). Recent social debates, such as the political discussions and events during the Trump administration from 2017 to 2021 or the ongoing controversies concerning the handling of the COVID-19 pandemic, however, point toward much stronger polarization processes, which also is in line with the observation of a global trend of increasing polarization that entails radical and populist tendencies, especially in political contexts (Deitelhoff et al., 2020). Researchers have used different ways of assessing the polarization of public opinion, which can be a reason for the different findings and conclusions. In an overview, Bramson et al. (2017) identified nine different concepts of assessing polarization. Some concepts are based on the spread and range of answers as well as the distance between extreme positions across an entire population, other concepts are based on the overall shape of a distribution and the dispersion of the data and consider indicators such as mean values, differences, standard deviations, and other related statistical measures. Furthermore, polarization can be understood as little diversity of opinion (*narrow bands of opinion space*) or, in contrast, ideally distinctive groups or diversity of opinion within groups. Other conceptions rather focus on the temporal changes of groups or the group size as such (*size parity*). Our analyses of the public opinion data consider the distribution of answers, mean values, and other statistical measures within the entire sample at a given time.

In addition to the existing focus on political polarization, current research turns toward polarization regarding the COVID-19 debate. Bruine de Bruin et al. (2020) examined US citizens' attitudes toward COVID-19 policies, risk perception, and protective behavior depending on political orientation. They found that Democrats perceived the virus to be more risky in terms of health and economics than Republicans. Likewise, they were more supportive of COVID-19 policies and more likely to fear their early repeal. Allcott et al. (2020) addressed the relationship between political party differences and social distancing during the pandemic in the U.S. population. In addition to the analysis of GPS data, they conducted an online survey, according to which respondents reduced their social contacts by 70% on average (self-reported behavior). This study again showed that Democrats take the pandemic more serious, as they reduced their social contacts more and considered social distancing to be more effective in prevention than Republicans. In addition, Democrats estimate future infection rates higher than Republicans.

In this article, we also examine opinion polarization regarding preventive COVID-19 measures, but without the focus on political orientation. Rather, we seek to present a general overview of the polarization regarding different COVID-19 measures in the DACH region in summer 2020.

Integrating Survey Data with Digital Trace Data

The combination and integration of survey and digital trace data is an emerging field. Recent work by Pasek et al. (2020a) compares presidential approval with sentiment among population



subgroups and found that sentiment is infeasible as a proxy from a microscopic viewpoint while being similar from a macroscopic viewpoint. Thus, their research outlines that a macroscopic comparison is not enough to draw valid conclusions. In our work, we, therefore, also consider the microscopic perspective to mitigate spurious correlations in the data.

In another study, Pasek et al. (2020b) compared the attention toward various campaign events in the 2016 presidential election between tweets and open-ended survey responses. They found that Twitter and survey data, in general, provide a similar picture on attention but differ in certain details, for example, in event peak days. Similarly, we compare polarization between Twitter and closed-ended survey responses on a macroscopic level and discuss their similarities and differences. Moreover, we also address a limitation in their work, as we account for more comparable subsets such as Twitter users in the survey data.

Bach et al. (2019) investigate whether voting behavior can be predicted using digital trace data in Germany and find that online behavior is not a good predictor for voting choices, but achieved different results depending on the party, with voting predictions for the right-wing populist and progressive environmentalist party performing slightly better. Their research outlines that even the microscopic data in social media is not enough to accurately predict user choices offline. Hence, we link the microscopic data to ensure that the online behavior of users corresponds to their survey opinions.

Regarding polarization, Joseph et al. (2019) studied the manifestation of polarization between survey and Twitter data by considering the support of tweets from Donald Trump depending on its content, for example, tweet sentiment. They found that, while Republicans show higher support in general, tweets of Trump that contain positive language, for example, express positive sentiment, have higher relative support across partisan lines than tweets with negative language. Their findings are also consistent between survey and Twitter data, which is congruent with our findings on opinion polarization in the COVID-19 prevention measures. Unlike their study, we directly relate the levels of polarization in both survey responses and Twitter content using statistical measures. Also, we do not restrict our analyses to political parties.

The research of Al Baghal et al. (2019) discusses the problems of linking individual survey data. They found that the consent rates are very low, especially on web surveys, which may introduce bias in the data. Our research might be subject to low consent rates and possible biases in the integrated data. For this reason, we also compare the data on a macroscopic level that does not require consent and use our small sample with linked data to further strengthen and verify our findings.

Integrating survey data with digital trace data is challenging in several aspects. Stier et al. (2020) describe three key issues that emerge when integrating survey data with digital trace data, that is, (i) consent when linking individual data, (ii) methodological and ethical issues of the analysis, and (iii) dealing with the multi-dimensionality of such data. All three issues apply to our research. Hence, to address issue (i), we collected individual data only from survey respondents who gave their informed consent. We informed the respondents about the nature of our research and explained that we will analyze their social media posts in case they share their handles with us. That procedure, plus the anonymization of all identifying personal information in any publication, reduces the ethical concerns (ii). The main emphasis of our paper is on the methodological challenges as expressed in (ii) and (iii). We tackle the problem of multi-modality of the data sources by comparing similar statistics across the different data types, that is, we compare agreement and sentiment using mean, variance, skewness, and kurtosis, as well as derived statistics such as the bimodality coefficient. Yet, we are aware of the different nature of our data and strive to avoid fallacies on inference between survey and social media data compared to just considering aggregate data. Regarding the linking types introduced by Stier et al. (2020), we use both aggregate-level and individual-level ex post linking, that is, both of which use historical data.

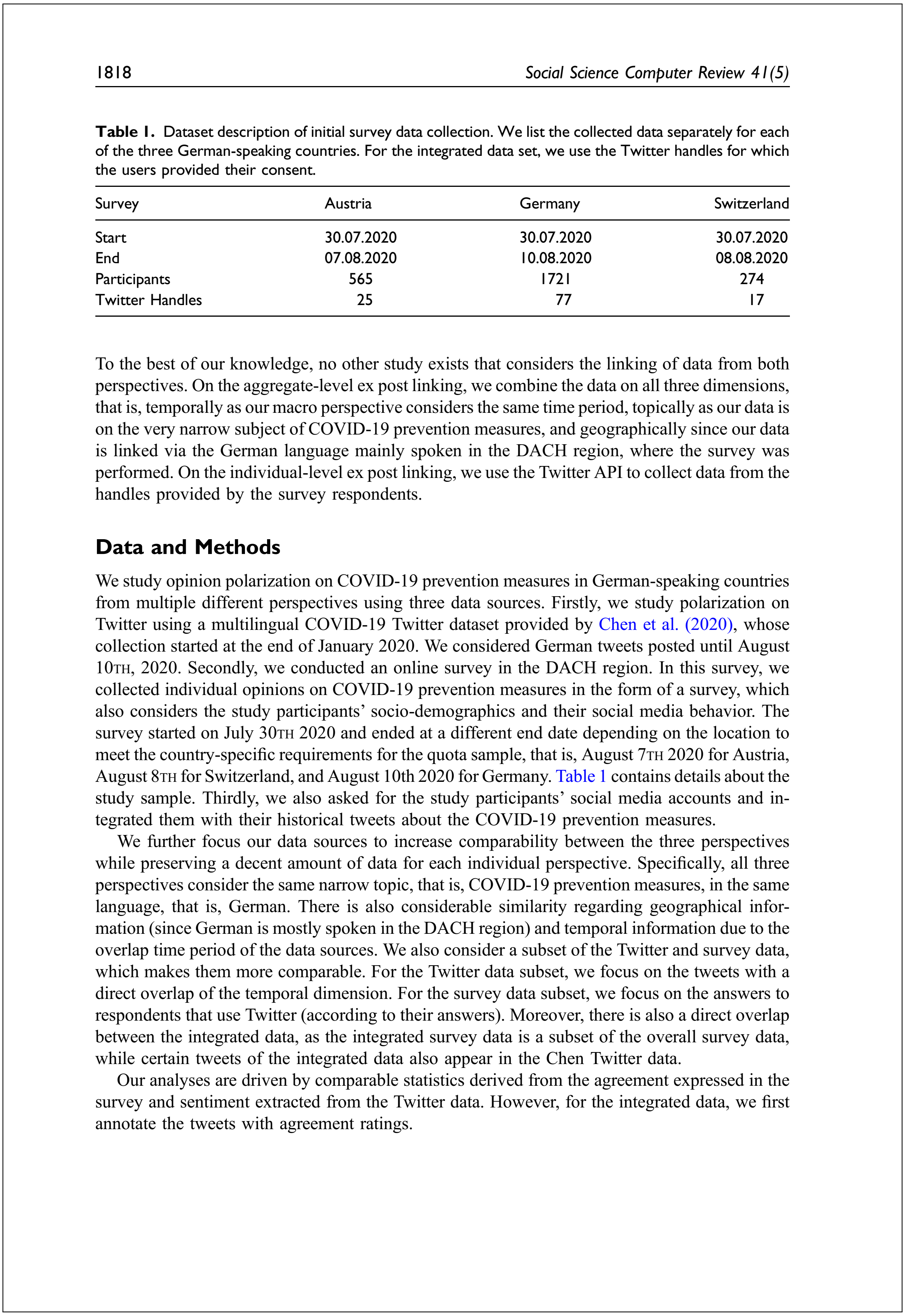


Table 1. Dataset description of initial survey data collection. We list the collected data separately for each of the three German-speaking countries. For the integrated data set, we use the Twitter handles for which the users provided their consent.

Survey	Austria	Germany	Switzerland
Start	30.07.2020	30.07.2020	30.07.2020
End	07.08.2020	10.08.2020	08.08.2020
Participants	565	1721	274
Twitter Handles	25	77	17

To the best of our knowledge, no other study exists that considers the linking of data from both perspectives. On the aggregate-level ex post linking, we combine the data on all three dimensions, that is, temporally as our macro perspective considers the same time period, topically as our data is on the very narrow subject of COVID-19 prevention measures, and geographically since our data is linked via the German language mainly spoken in the DACH region, where the survey was performed. On the individual-level ex post linking, we use the Twitter API to collect data from the handles provided by the survey respondents.

Data and Methods

We study opinion polarization on COVID-19 prevention measures in German-speaking countries from multiple different perspectives using three data sources. Firstly, we study polarization on Twitter using a multilingual COVID-19 Twitter dataset provided by Chen et al. (2020), whose collection started at the end of January 2020. We considered German tweets posted until August 10th, 2020. Secondly, we conducted an online survey in the DACH region. In this survey, we collected individual opinions on COVID-19 prevention measures in the form of a survey, which also considers the study participants' socio-demographics and their social media behavior. The survey started on July 30th 2020 and ended at a different end date depending on the location to meet the country-specific requirements for the quota sample, that is, August 7th 2020 for Austria, August 8th for Switzerland, and August 10th 2020 for Germany. Table 1 contains details about the study sample. Thirdly, we also asked for the study participants' social media accounts and integrated them with their historical tweets about the COVID-19 prevention measures.

We further focus our data sources to increase comparability between the three perspectives while preserving a decent amount of data for each individual perspective. Specifically, all three perspectives consider the same narrow topic, that is, COVID-19 prevention measures, in the same language, that is, German. There is also considerable similarity regarding geographical information (since German is mostly spoken in the DACH region) and temporal information due to the overlap time period of the data sources. We also consider a subset of the Twitter and survey data, which makes them more comparable. For the Twitter data subset, we focus on the tweets with a direct overlap of the temporal dimension. For the survey data subset, we focus on the answers to respondents that use Twitter (according to their answers). Moreover, there is also a direct overlap between the integrated data, as the integrated survey data is a subset of the overall survey data, while certain tweets of the integrated data also appear in the Chen Twitter data.

Our analyses are driven by comparable statistics derived from the agreement expressed in the survey and sentiment extracted from the Twitter data. However, for the integrated data, we first annotate the tweets with agreement ratings.

COVID-19 situation

Given that the responses and Tweets are influenced by the actual state of the pandemic, we now offer a brief overview of the situation during our data collection period. After the first peak in spring 2020, the pandemic situation in the German-speaking countries was rather calm during the summer. However, due to holiday traffic, infection rates started to rise again and containing measures were discussed anew. At the time of the survey, the stringency index (Ritchie et al., 2020), which records the strictness of active COVID-19 policies (from 0 to 100, 100 = strictest), was between 55.09 and 56.94 in Germany, between 39.35 and 43.06 in Switzerland, and stable at 37.96 in Austria. In all three countries, face masks were required in some public spaces over the whole period, comprehensive contact tracing of all cases was conducted and vaccination was not yet available. Furthermore, there were no stay-at-home requirements during this time span in all three countries, workplace closures and public event cancellations were required for some regions in Germany and Switzerland, in Austria it was recommended (Ritchie et al., 2020).

Twitter Data

We analyze the *Twitter data* using the publicly available dataset from Chen et al. (2020). This dataset contains the 1% sample of tweet IDs from the Twitter streaming API[1] on a predefined set of COVID-19-related accounts and keywords, for example, *COVID-19* and *Coronavirus*. Using these tweet IDs, we gathered tweets in the period of the survey, that is, we use the maximum length from January 29th 2020 to August 10th 2020, as shown in Table 1, retroactively, which is called hydration. As suggested by the authors, we hydrate the tweets using twarc,[2] which uses Twitter's lookup API. As a consequence of the hydration, some tweets might no longer be available, that is, the tweets were deleted.[3] We filtered the tweets to only include German tweets, resulting in 3, 336, 562 tweets for our analyses. Additionally, we also focused on the subset of tweets within the same time period of the survey data, that is, July 30th 2020 to August 10th 2020, which resulted in 547, 579 tweets. When referring to this subset, we explicitly state it, otherwise, we refer to the full Twitter Data.

We further filter the tweets according to three predefined word stems that resemble the three prevention measures to be considered. The rationale for using stems as identification of tweets is due to its simplicity, and thus interpretability while capturing virtually all target tweets.[4] Specifically, we use *impf* for vaccination, *mask* for mask wearing, and *trac* for contact tracing. These three stems capture virtually all tweets related to these measures. The stem *impf* captures the German noun *Impfung* and verb *impfen* for vaccination, as well as other words related to it such as the vaccine itself (i.e., *Impfstoff* in German). The stem *mask* captures the German noun *Maske* and related words such as mask mandate (i.e., *Maskenpflicht* in German). The stem *trac* captures both contact tracing and contact tracer, which have been Germanized and used by the public for COVID-19. For the full dataset, this results in 63, 676 for *impf*, 136, 198 for *mask*, and 13, 151 for *trac* tweets, respectively. Considering the subset number of tweets gets reduced to 12, 260 for *impf*, 31, 856 for *mask*, and 1, 385 for *trac*.

We conduct the sentiment analysis using the TextBlob library with the German language extension[5], which includes a sentiment polarity lexicon that we use for sentiment extraction. After extracting the sentiment, we remove tweets that express no sentiment to exclude purely objective statements, for example, in scientific discussions or very short statements in tweets, which would otherwise dominate the resulting distribution. The final full dataset for comparison consists of 25, 769 tweets expressing sentiment for vaccination, 60, 218 for mask wearing, and 4, 819 for contact tracing. For the subset, the numbers of tweets with sentiment are 5, 420 for vaccination, 15, 425 for mask wearing, and 634 for contact tweets. The extracted sentiments are on a numerical scale from −1 for negative to +1 for positive sentiment.

Please note that with this procedure, we exclude 57.37% of the tweets since they do not contain any words contained in the sentiment polarity lexicon. Here, potentially valuable tweets might be excluded, for example, from users, who choose to write their tweets using words not associated with sentiment. Furthermore, TextBlob only uses simple rules in combination with the lexicon, for example, to detect negations. Thus, more nuanced types of statements, such as sarcasm, are unlikely to be detected and might be associated with the wrong polarity. Finally, the method only considers the text itself, but not contextual features such as conversation threads and user attributes. As a result, the sentiment analysis, while being well-established, can only act as a proxy since true opinions are unavailable in Twitter data.

Survey Data

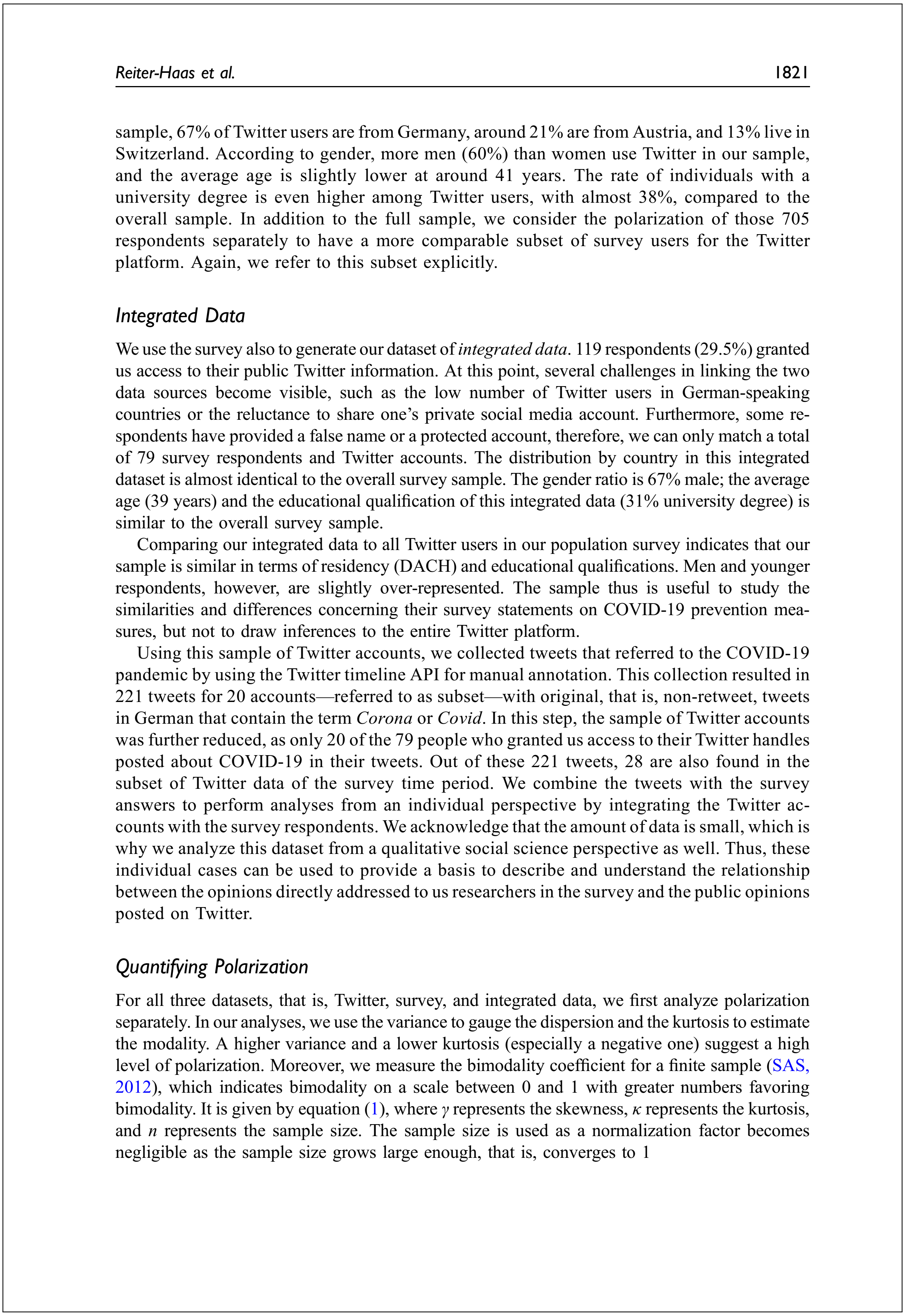
The *survey data* was collected through an online survey in July and August 2020 in Germany, Austria, and Switzerland as a representative quota sample and comprises 2560 respondents. The targeted quotas were based on the official distribution of gender, age, and federal state/canton in the population of the three DACH countries. As these quotas were met, it can be assumed that the sample is representative of the overall population in these countries regarding those aspects. However, only people with Internet access were considered for the sample, as the survey was conducted online. The questionnaire consists of opinions on polarizing topics (including the COVID-19 prevention measures of vaccination, mask wearing, and contact tracing), social media use (including private Twitter handles), and socio-demographics.[6]

The items concerning attitudes toward polarizing topics were taken from the questionnaire of the project ”The measurement of CO2 relevant environmental behaviors and other environmental attitudes through surveys,” funded by the Austrian National Bank (OeNB) and carried out by the Institute of Sociology (University of Graz) in 2019, and adapted to the desired requirements, that is, COVID-19. The items regarding social media use were taken from the questionnaire of the project “Future of Life,” conducted by the Institute of Sociology (University of Graz) in 2018/2019.

In our sample, 67% of the respondents are from Germany, another 22% live in Austria, and the remaining 11% are from Switzerland, whereas Austrians are oversampled and the Swiss sample is limited to the German-speaking area. Gender is equally distributed, and the average age is 44 years. The sample shows a high level of education as it contains an above-average number of respondents with a university degree, with almost 30% (the rate of people with university degrees varies in the DACH region between 16% and 21% (Bundesamt für Statistik Schweiz, 2021; Statistik Austria, 2018; Statistisches Bundesamt Deutschland Destatis, 2020).

We analyze polarization in terms of agreement with the COVID-19 prevention measure on an ordinal scale from 1 for strong disagreement to 5 for strong agreement.

In the survey, we also asked participants about their Twitter use and consent to use their private data for our analyses. First, we assured them of the confidential treatment of their Twitter data, and asked them for their consent to link this data to the survey data as well. Respondents first had to give their consent to provide us with their personal Twitter username and access to their data before being asked about their actual Twitter handle in a follow-up question (where we provided an example, i.e., @jane.doe). Of the 2560 respondents in our population survey, 705 people (27.5%) use Twitter between “several times a day” and “less than once a week.” In this respect, our data reflect the findings of social media use statistics that Twitter is far less widespread in German-speaking countries compared to other social media platforms, such as Facebook or Instagram Newman et al., 2021). As in the overall



sample, 67% of Twitter users are from Germany, around 21% are from Austria, and 13% live in Switzerland. According to gender, more men (60%) than women use Twitter in our sample, and the average age is slightly lower at around 41 years. The rate of individuals with a university degree is even higher among Twitter users, with almost 38%, compared to the overall sample. In addition to the full sample, we consider the polarization of those 705 respondents separately to have a more comparable subset of survey users for the Twitter platform. Again, we refer to this subset explicitly.

Integrated Data

We use the survey also to generate our dataset of *integrated data*. 119 respondents (29.5%) granted us access to their public Twitter information. At this point, several challenges in linking the two data sources become visible, such as the low number of Twitter users in German-speaking countries or the reluctance to share one's private social media account. Furthermore, some respondents have provided a false name or a protected account, therefore, we can only match a total of 79 survey respondents and Twitter accounts. The distribution by country in this integrated dataset is almost identical to the overall survey sample. The gender ratio is 67% male; the average age (39 years) and the educational qualification of this integrated data (31% university degree) is similar to the overall survey sample.

Comparing our integrated data to all Twitter users in our population survey indicates that our sample is similar in terms of residency (DACH) and educational qualifications. Men and younger respondents, however, are slightly over-represented. The sample thus is useful to study the similarities and differences concerning their survey statements on COVID-19 prevention measures, but not to draw inferences to the entire Twitter platform.

Using this sample of Twitter accounts, we collected tweets that referred to the COVID-19 pandemic by using the Twitter timeline API for manual annotation. This collection resulted in 221 tweets for 20 accounts—referred to as subset—with original, that is, non-retweet, tweets in German that contain the term *Corona* or *Covid*. In this step, the sample of Twitter accounts was further reduced, as only 20 of the 79 people who granted us access to their Twitter handles posted about COVID-19 in their tweets. Out of these 221 tweets, 28 are also found in the subset of Twitter data of the survey time period. We combine the tweets with the survey answers to perform analyses from an individual perspective by integrating the Twitter accounts with the survey respondents. We acknowledge that the amount of data is small, which is why we analyze this dataset from a qualitative social science perspective as well. Thus, these individual cases can be used to provide a basis to describe and understand the relationship between the opinions directly addressed to us researchers in the survey and the public opinions posted on Twitter.

Quantifying Polarization

For all three datasets, that is, Twitter, survey, and integrated data, we first analyze polarization separately. In our analyses, we use the variance to gauge the dispersion and the kurtosis to estimate the modality. A higher variance and a lower kurtosis (especially a negative one) suggest a high level of polarization. Moreover, we measure the bimodality coefficient for a finite sample (SAS, 2012), which indicates bimodality on a scale between 0 and 1 with greater numbers favoring bimodality. It is given by equation (1), where γ represents the skewness, κ represents the kurtosis, and n represents the sample size. The sample size is used as a normalization factor becomes negligible as the sample size grows large enough, that is, converges to 1

$$\beta = \frac{\gamma^2 + 1}{\kappa + 3\frac{(n-1)^2}{(n-2)(n-3)}} \qquad (1)$$

The bimodality coefficient has some caveats regarding its use for identifying true bimodal distributions (Pfister et al., 2013). However, it captures the basic intuition for quantifying polarization, that is, both high skewness and low kurtosis are associated with a higher amount of polarization. Consequently, and in line with intuition, it also assigns a high value in the case of an unimodal but highly skewed distribution.

Evaluating Congruence

In the integrated data, we first investigate the tweet content to get a better understanding of the specific topics that Twitter users are discussing. Following qualitative content analysis according to Mayring (2015), we inductively categorize a subset of COVID-19-related tweets of our integrated data survey users. This way, we want to identify the specific topics that survey users talk about on Twitter when using keywords regarding COVID-19 or hashtags such as *#COVID-19* and *#Corona*.

To evaluate the congruence of the survey and Twitter data, we manually annotated the subset of 20 users with a total of 221 tweets on the COVID-19 prevention measures by two annotators and compared these with the survey data. For the annotations, we chose the same labels as in the survey, that is, an ordinal rating scale of agreement. We calculate the binary inter-annotator agreement, which only considers perfect matches, between the survey answers and the Twitter annotations. Evaluating the congruence is an important aspect for ensuring the comparability between the survey data and the Twitter data.

Inductive Category Formation. To analyze whether the provided content fits the case for the congruence evaluation, we perform a qualitative content analysis to inductively categorize the content. This approach provides insights into the topics discussed by the integrated users.

The content analysis includes 221 tweets from 20 survey users and discovers a huge variety of categories. Political topics (both local and global politics, over 70 times in total) were most frequently addressed in connection with COVID-19. Here, politicians' handling of the pandemic was frequently discussed and criticized. There was also frequent debate about how dangerous COVID-19 was (almost 50 times). Comparisons were often made with influenza, or personal experiences with COVID-19 were reported. Furthermore, different prevention measures were mentioned about 25 times, and individual problems, as well as societal challenges due to the pandemic, were reported (about 20 times). In addition, private and professional changes in everyday life were reported a few times (more than 10 times). Financial support from the government and how relief funds should be distributed was mentioned in similar frequency. Tweets about scientific research results and data were shared around 10 times, and tweets about fake news and conspiracy theories similarly often. Topics that were less frequently mentioned were polarizing role attributions (e.g., COVID-19 deniers), Corona apps, demonstrations, maintaining occupations, future scenarios, or toilet paper.

Alongside such content-related topics, jokes (sarcasm, irony) about the current situation as well as emotions were frequently found in the COVID-19-related tweets (around 30 times). Here, mainly negative emotions such as annoyance, disappointment, or nervous breakdowns were reported. However, there were also positive emotions mentioned, such as good wishes or hope.

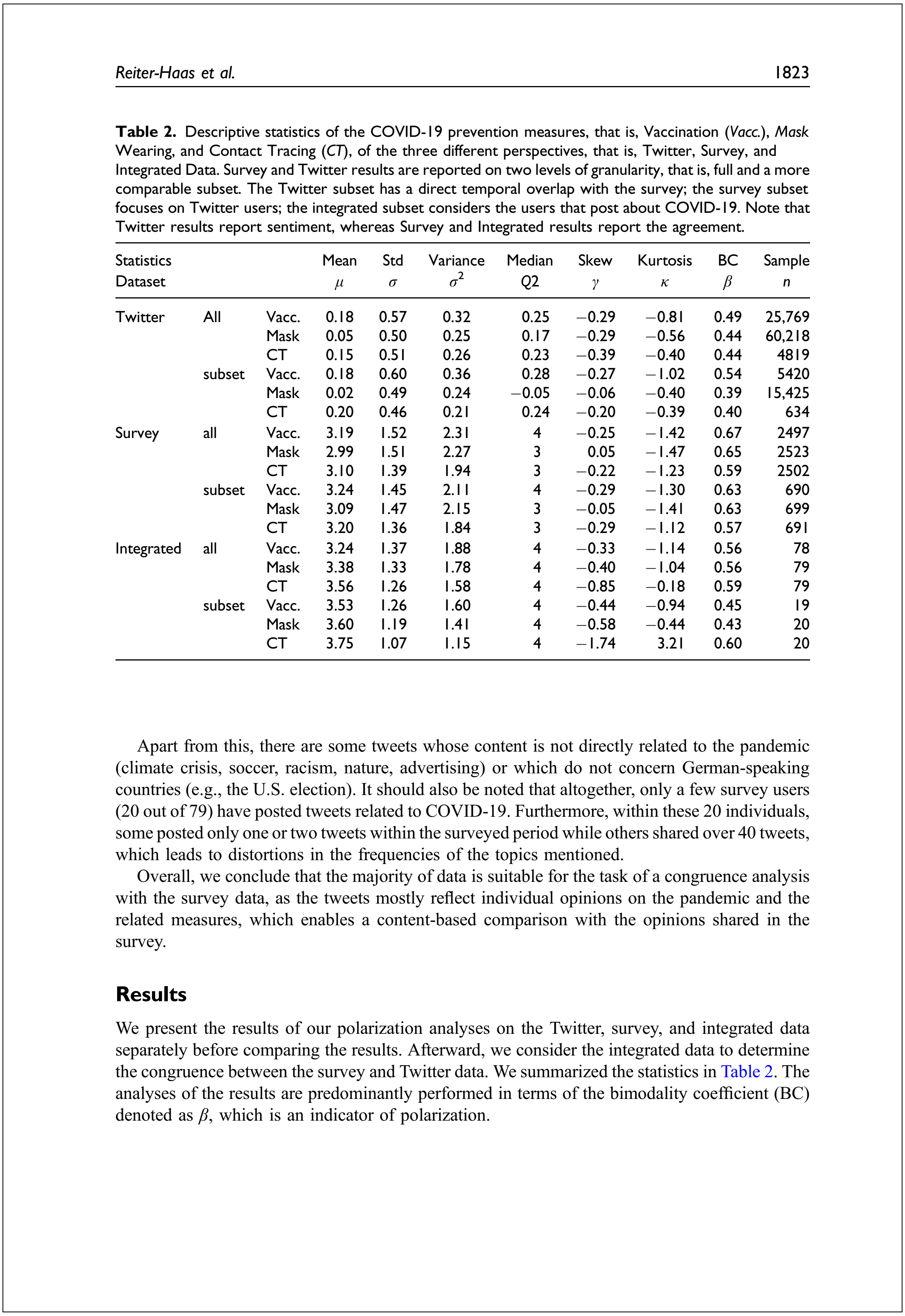


Table 2. Descriptive statistics of the COVID-19 prevention measures, that is, Vaccination (*Vacc.*), *Mask* Wearing, and Contact Tracing (*CT*), of the three different perspectives, that is, Twitter, Survey, and Integrated Data. Survey and Twitter results are reported on two levels of granularity, that is, full and a more comparable subset. The Twitter subset has a direct temporal overlap with the survey; the survey subset focuses on Twitter users; the integrated subset considers the users that post about COVID-19. Note that Twitter results report sentiment, whereas Survey and Integrated results report the agreement.

Statistics			Mean	Std	Variance	Median	Skew	Kurtosis	BC	Sample
Dataset			μ	σ	σ^2	Q2	γ	κ	β	*n*
Twitter	All	Vacc.	0.18	0.57	0.32	0.25	−0.29	−0.81	0.49	25,769
		Mask	0.05	0.50	0.25	0.17	−0.29	−0.56	0.44	60,218
		CT	0.15	0.51	0.26	0.23	−0.39	−0.40	0.44	4819
	subset	Vacc.	0.18	0.60	0.36	0.28	−0.27	−1.02	0.54	5420
		Mask	0.02	0.49	0.24	−0.05	−0.06	−0.40	0.39	15,425
		CT	0.20	0.46	0.21	0.24	−0.20	−0.39	0.40	634
Survey	all	Vacc.	3.19	1.52	2.31	4	−0.25	−1.42	0.67	2497
		Mask	2.99	1.51	2.27	3	0.05	−1.47	0.65	2523
		CT	3.10	1.39	1.94	3	−0.22	−1.23	0.59	2502
	subset	Vacc.	3.24	1.45	2.11	4	−0.29	−1.30	0.63	690
		Mask	3.09	1.47	2.15	3	−0.05	−1.41	0.63	699
		CT	3.20	1.36	1.84	3	−0.29	−1.12	0.57	691
Integrated	all	Vacc.	3.24	1.37	1.88	4	−0.33	−1.14	0.56	78
		Mask	3.38	1.33	1.78	4	−0.40	−1.04	0.56	79
		CT	3.56	1.26	1.58	4	−0.85	−0.18	0.59	79
	subset	Vacc.	3.53	1.26	1.60	4	−0.44	−0.94	0.45	19
		Mask	3.60	1.19	1.41	4	−0.58	−0.44	0.43	20
		CT	3.75	1.07	1.15	4	−1.74	3.21	0.60	20

Apart from this, there are some tweets whose content is not directly related to the pandemic (climate crisis, soccer, racism, nature, advertising) or which do not concern German-speaking countries (e.g., the U.S. election). It should also be noted that altogether, only a few survey users (20 out of 79) have posted tweets related to COVID-19. Furthermore, within these 20 individuals, some posted only one or two tweets within the surveyed period while others shared over 40 tweets, which leads to distortions in the frequencies of the topics mentioned.

Overall, we conclude that the majority of data is suitable for the task of a congruence analysis with the survey data, as the tweets mostly reflect individual opinions on the pandemic and the related measures, which enables a content-based comparison with the opinions shared in the survey.

Results

We present the results of our polarization analyses on the Twitter, survey, and integrated data separately before comparing the results. Afterward, we consider the integrated data to determine the congruence between the survey and Twitter data. We summarized the statistics in Table 2. The analyses of the results are predominantly performed in terms of the bimodality coefficient (BC) denoted as β, which is an indicator of polarization.

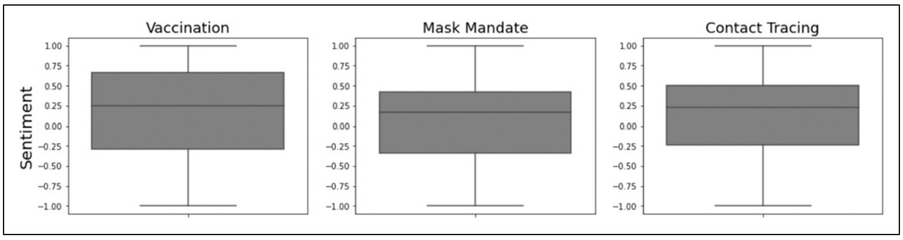


Figure 1. Polarization in Twitter data (all: n = 90, 806 tweets) in terms of sentiment in the three prevention measures, that is, vaccination, mask wearing, and contact tracing. The sentiments are measured per tweet on a range from −1 for the maximum negative sentiment to +1 for the maximum positive sentiment. Tweets with neutral sentiment are excluded. Vaccination shows high variance which indicates a high level of polarization, but also the highest median suggesting a more positive leaning toward the measure.

Polarization in Twitter Data

We analyzed polarization regarding the prevention measures in the COVID-19 German dataset and found that all three measures are polarized as shown in Figure 1. We observe that vaccination has the highest dispersion, that is, a variance of 0.32 compared to 0.25 and 0.26, which is already an indicator for polarization. Investigating the kurtosis further strengthens this observation, which is far lower than the other two measures, that is, −0.81 compared to −0.56 and −0.4. Considering the skewness, we observe similar results, but vaccination has the highest mean (0.18) and median (0.25). This shows that the approval of the measures is higher than the rejection. Moreover, all three prevention measures are leaning more toward the positive, that is, approval side with the mean and median being positive. Computing the bimodality coefficient reaffirms the observation that vaccination is the most polarizing with β = 0.49.

The results are very similar for the temporal subset of Twitter data as it has similar medians and dispersion of the data (due to its marginal differences to Figure 1 we omitted showing the boxplot). However, we observe a noticeable change in the bimodality coefficient. This results in an increase for the bimodality coefficient in vaccination with β = 0.54 and a decrease for the other two prevention measures with a β of 0.39 and 0.4. Overall, we conclude that the prevention measures are polarizing in terms of sentiment, and find that there are differences in the opinions depending on the prevention measures, as vaccination is substantially more polarizing compared to the other two prevention measures.

Polarization in Survey Data

The polarization of public opinion is particularly evident with regard to socio-political measures addressing the COVID-19 pandemic. We investigated the agreement to the introduction of compulsory vaccination, voluntary wearing of face masks, and contact tracing. In the entire sample, both supporters and opponents of all three prevention measures can be found to a similar extent. The highest level of support can be found for the introduction of compulsory vaccination (51% agree absolutely or rather agree), the strongest opposition can be observed against the voluntary wearing of face masks (45% disagree or rather disagree).

The distribution of the variables regarding the different COVID-19 prevention measures for the overall sample can be seen in Figure 2. Compulsory vaccination receives the highest level of agreement, with a mean of 4. The variables considered here are ordinal, but we nevertheless consider certain statistical indicators of dispersion for the sake of comparability with the Twitter

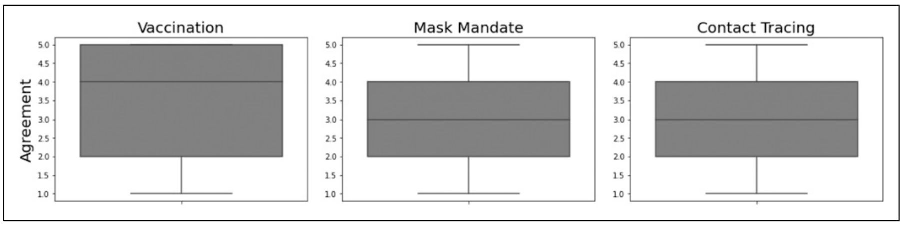


Figure 2. Polarization in Survey data (all: n = 2560 respondents) in terms of agreement to the three prevention measures, that is, vaccination, mask wearing, and contact tracing. The agreement is measured per respondent on a range from 1 for strong disagreement to 5 for strong agreement. Vaccination shows high variance which indicates a high level of polarization, but also the highest median suggesting a more positive leaning toward the measure.

analysis. The first quartile for all three prevention measures lies at 2, which means that 25% of respondents are below this level and do not agree with the prevention measures. The 75% quartile is highest for compulsory vaccination, which again indicates the highest level of agreement with this prevention measure. An additional comparison by country shows that respondents from Germany express the strongest support for all three prevention measures. Meanwhile, respondents from Austria show the highest level of rejection of the prevention measures, especially of contact tracing and compulsory vaccination. It can be noted that in the overall DACH region there tends to be a higher level of support for those three COVID-19 prevention measures than the rejection of the same.

Considering the bimodality coefficient, we observe that all three prevention measures are polarizing with vaccination being the most polarizing by having a bimodality coefficient of 0.67 compared to 0.65 for mask wearing, and 0.59 for contact tracing. Considering the subset of Twitter users shows similar results (the boxplot is almost identical to Figure 2 and thus omitted), but leads to a noticeable drop in the bimodality coefficient. This observation suggests that Twitter users in our sample are less polarized compared to the overall population.

Polarization in Integrated Data

Here, we analyzed the 79 respondents, whose Twitter handles could be successfully matched between the opinions expressed in the survey and the tweets posted online. This group turned out to be more likely in favor of the prevention measures compared to all respondents who use Twitter—especially contact tracing (63% vs. 48%) and wearing of face masks (55% vs. 44%), whereas compulsory vaccination is seen similar (51% vs. 51%).

Figure 3 shows the distributions of agreement on the three COVID-19 prevention measures of the respondents analyzed in this section. The median of all three prevention measures is located at the upper end of the boxes and, in case of contact tracing and face masks, higher than in the overall sample as well as in the subsample of Twitter users.

A decrease of polarization is also reflected in the bimodality coefficient of 0.56 for vaccination and mask wearing, and 0.59 for contact tracing. Interestingly, in this dataset vaccination has a lower bimodality coefficient than contact tracing, whereas vaccination was consistently the highest in terms of the bimodality coefficient for all other datasets.

Considering the subset, we observe an additional drop for vaccination to 0.45 and mask wearing to 0.43. Whereas contact tracing rises to 0.6 as a result of a very negative skewness, that is only partially counteracted by a high kurtosis. This is mainly due to the size of the subsample (n =

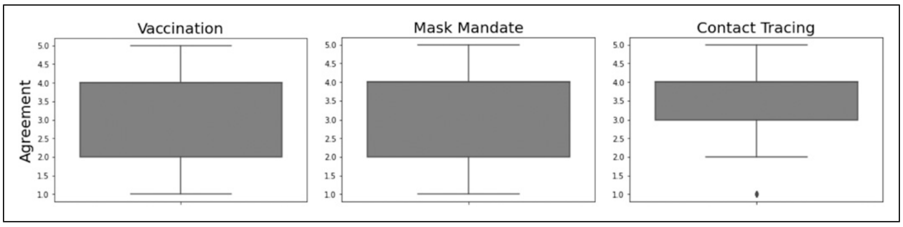


Figure 3. Polarization in the Integrated data (all: n = 79 respondents) in terms of agreement among the three prevention measures, that is, vaccination, mask wearing, and contact tracing. The agreement is measured per respondent on a range from 1 for strong disagreement to 5 for strong agreement. Both, vaccination and mask wearing, show a high variance which indicates a high amount of polarization. All three measures have a median of 4, suggesting a leaning toward approval of the measures.

20), since the smaller the data set, the greater the impact of outliers. Also, agreement on contact tracing is far more unevenly distributed than among the other two preventive measures (65% of respondents expressed an agreement value of 4 on a scale from 1 to 5).

Comparison between Polarization Results

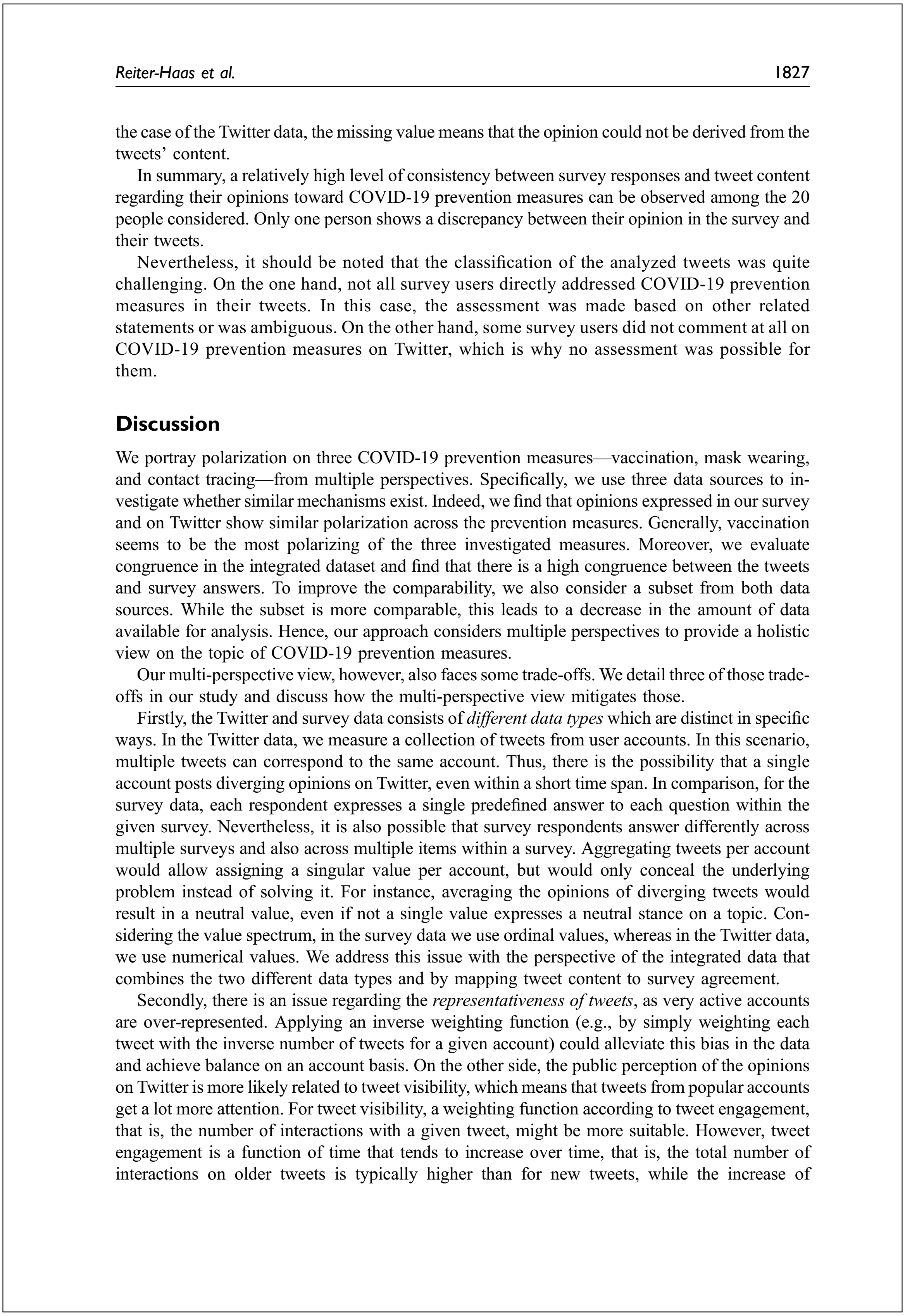
To discuss the COVID-19 measures holistically, we compare the distributions of the provided boxplots, that is, Figure 1 for Twitter, Figure 2 for the survey, and Figure 3 for the integrated data.

We observe that the sentiments in the Twitter data are less dispersed compared to the survey data and also have a lower bimodality coefficient. Note that Twitter data is collected at the level of tweets and measured in terms of sentiment, whereas survey data is based on a single response per item and respondent and measured in terms of agreement. Nevertheless, we find that the overall characteristics are rather similar in all distributions. The prevention measures of contact tracing and mask wearing are less polarized and do not display a clear tendency toward either side, whereas prevention measures on vaccinations are highly polarized and skewed toward agreement/positive sentiment.

Our observation indicates that the opinions of survey participants directly relate to the opinions of Twitter users. To test this assumption, we compare the tweets in the integrated data, that is, which were provided by the survey participants, with their respective survey answers.

Congruence of Opinions in Integrated Data. Multiple tweets in our Twitter data can belong to one account, whereas for the survey data, we have a single answer per respondent. This fact limits the comparability between the two data sources. To mitigate this issue, we also consider the association between the opinions expressed in the survey and through their Twitter accounts within the integrated data. In this regard, we enable a direct comparison of the two different data types, that is, sentiment and agreement, by manually annotating the tweets.

The binary inter-annotator agreement for the assessment of tweets is α = 0.7 and includes missing values, that is, where the stance toward the prevention measure could not be derived. In comparison, random annotations would only agree 1/6 of the time (scale 1–5 and missing). The rating scale is similar to the survey scale, based on agreement on a prevention measure of 1–5 (1 for strong disagreement and 5 for strong agreement). Both rating scales also allow for missing values, but the meaning differs slightly. In the case of the survey data, missing values mean that participants either have no opinion or that the participants do not want to specify their opinion. In



the case of the Twitter data, the missing value means that the opinion could not be derived from the tweets' content.

In summary, a relatively high level of consistency between survey responses and tweet content regarding their opinions toward COVID-19 prevention measures can be observed among the 20 people considered. Only one person shows a discrepancy between their opinion in the survey and their tweets.

Nevertheless, it should be noted that the classification of the analyzed tweets was quite challenging. On the one hand, not all survey users directly addressed COVID-19 prevention measures in their tweets. In this case, the assessment was made based on other related statements or was ambiguous. On the other hand, some survey users did not comment at all on COVID-19 prevention measures on Twitter, which is why no assessment was possible for them.

Discussion

We portray polarization on three COVID-19 prevention measures—vaccination, mask wearing, and contact tracing—from multiple perspectives. Specifically, we use three data sources to investigate whether similar mechanisms exist. Indeed, we find that opinions expressed in our survey and on Twitter show similar polarization across the prevention measures. Generally, vaccination seems to be the most polarizing of the three investigated measures. Moreover, we evaluate congruence in the integrated dataset and find that there is a high congruence between the tweets and survey answers. To improve the comparability, we also consider a subset from both data sources. While the subset is more comparable, this leads to a decrease in the amount of data available for analysis. Hence, our approach considers multiple perspectives to provide a holistic view on the topic of COVID-19 prevention measures.

Our multi-perspective view, however, also faces some trade-offs. We detail three of those trade-offs in our study and discuss how the multi-perspective view mitigates those.

Firstly, the Twitter and survey data consists of *different data types* which are distinct in specific ways. In the Twitter data, we measure a collection of tweets from user accounts. In this scenario, multiple tweets can correspond to the same account. Thus, there is the possibility that a single account posts diverging opinions on Twitter, even within a short time span. In comparison, for the survey data, each respondent expresses a single predefined answer to each question within the given survey. Nevertheless, it is also possible that survey respondents answer differently across multiple surveys and also across multiple items within a survey. Aggregating tweets per account would allow assigning a singular value per account, but would only conceal the underlying problem instead of solving it. For instance, averaging the opinions of diverging tweets would result in a neutral value, even if not a single value expresses a neutral stance on a topic. Considering the value spectrum, in the survey data we use ordinal values, whereas in the Twitter data, we use numerical values. We address this issue with the perspective of the integrated data that combines the two different data types and by mapping tweet content to survey agreement.

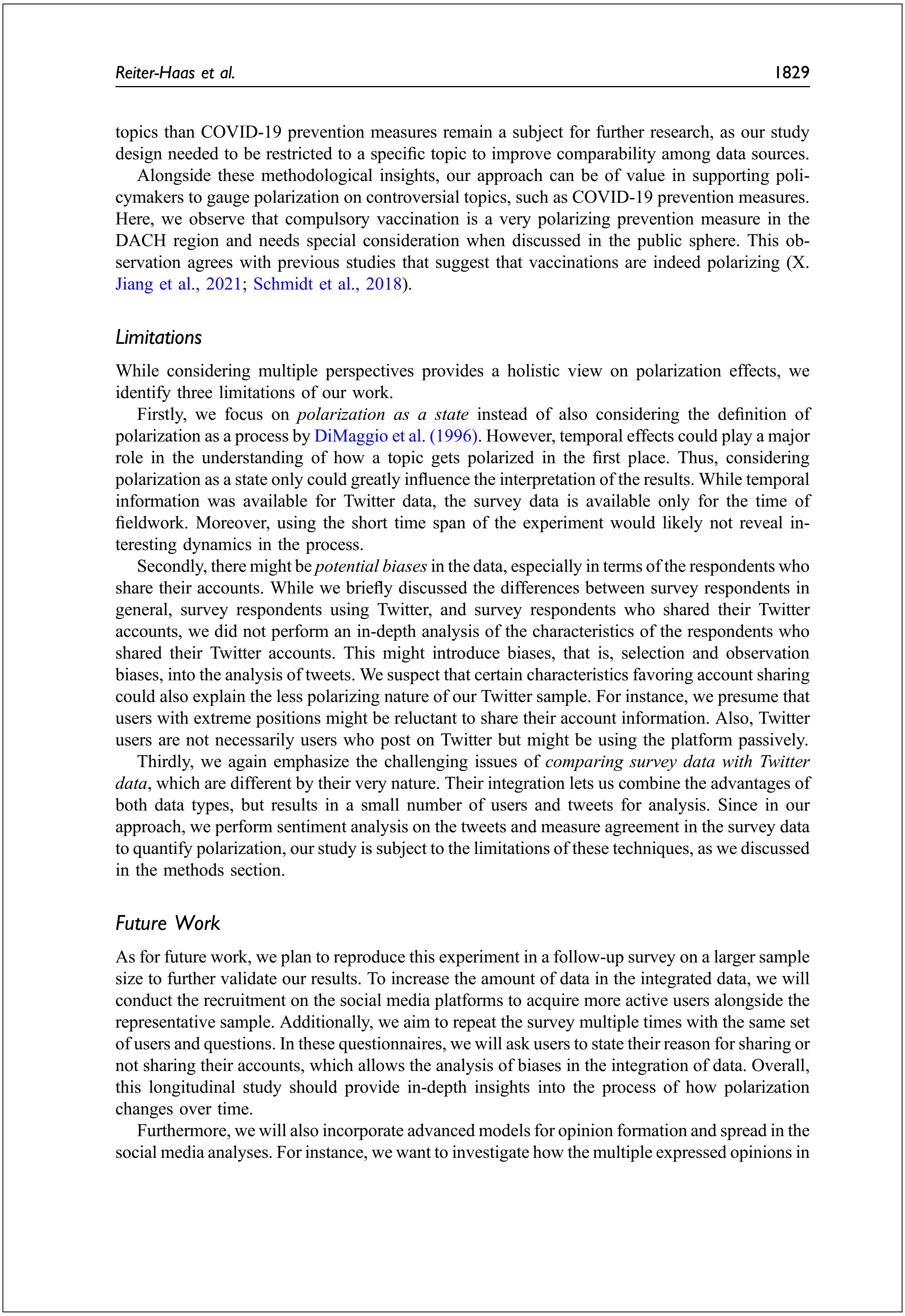
Secondly, there is an issue regarding the *representativeness of tweets*, as very active accounts are over-represented. Applying an inverse weighting function (e.g., by simply weighting each tweet with the inverse number of tweets for a given account) could alleviate this bias in the data and achieve balance on an account basis. On the other side, the public perception of the opinions on Twitter is more likely related to tweet visibility, which means that tweets from popular accounts get a lot more attention. For tweet visibility, a weighting function according to tweet engagement, that is, the number of interactions with a given tweet, might be more suitable. However, tweet engagement is a function of time that tends to increase over time, that is, the total number of interactions on older tweets is typically higher than for new tweets, while the increase of interactions is higher for newer tweets as they get more attention. We opted for a naive approach and omitted weighting tweets due to a lack of knowledge on which weighting function best captures the relevance of each tweet to its corresponding account. Moreover, treating tweets uniformly lies between the account-level weighting, that is, treating each account as equally important, and visibility-level weighting, that is, according to the public perception. As such, it provides a balance between those extreme weightings, while providing a natural way of representing the importance of social media content. Our approach mitigates some of the issues of representativeness, as we consider the polarization at different levels of granularity, including the very fine-grained level of our integrated data. In the integrated data, individual tweets are aggregated, and an overall assessment is derived, thus, alleviating the issue.

Thirdly, we analyze the polarization in the Twitter data *using sentiment exclusively*, but not in terms of positions or emotions. Considering positions would be non-trivial due to a lack of well-defined dimensions such as political ideology. Regarding the measuring of affective polarization using emotions, we performed a prestudy in terms of emotions, which did not lead to noteworthy results. In particular, the results were comparable to sentiment in terms of emotional valence but less distinct. Thus, we focus our analyses on sentiment for conciseness reasons. This single view on the Twitter data becomes less prevalent as we also report the perspective of the agreement in the survey data.

Although we find that polarization is similar between the perspectives, there are still differences between each of the data sources. Comparing the Twitter data with the survey data, we observe that the Twitter data is less polarized considering the bimodality coefficient. However, we cannot conclude that Twitter as a platform acts in a depolarizing manner. Although we observe that the subsample of survey respondents that use Twitter are less polarized, two other observations indicate that other effects could be the cause for this phenomenon. Firstly, a temporal focus on the Twitter data within the survey time period results in a change of the bimodality coefficient. The subset of Twitter data shows higher polarization for vaccination but decreases for mask wearing and contact tracing. This outlines the importance of considering temporal factors in the analysis. Secondly, we also observe that there is a lower level of polarization in agreement in the integrated data compared to the complete and Twitter subset of the survey data. Still, the polarization is substantially higher compared to the level of polarization in the Twitter data. We attribute this difference to the different kinds of data that are measured, respectively. In the Twitter data, the sentiments of tweets are measured, and multiple tweets can belong to the same account, whereas agreement of individuals in the survey data is measured at the particular time of the fieldwork. These observations show that a direct comparison would be infeasible and is the reason we also evaluated the congruence in the integrated data.

Overall, we show that both survey data and social media data have their merits when studying opinion polarization; however, both provide an incomplete picture. Twitter data is more abundant, whereas survey data provides representativeness. Additionally, considering the integrated data combines the advantage of both perspectives, but comes at the cost of difficulties in obtaining the data. As we found in our experiments, only a limited amount of data can be collected with such an approach. A possible remedy to increase the sample size could be to move the recruitment of survey participants to social media platforms, for example, similar to the approach described in Pötzschke and Weiß (2021), or to target specific user interests and user demographics.

Considering the perspectives for our topic, that is, COVID-19 prevention measures in the German-speaking DACH region, we find that there is a congruence of the different perspectives, but with variations in how pronounced the observed polarization is. Thus, each individual perspective would result in a similar conclusion, but the polarization is more noticeable in the survey data. Still, our research illustrates the importance of considering multiple perspectives, as there are noticeable differences between the perspectives. Whether our findings also apply to other



topics than COVID-19 prevention measures remain a subject for further research, as our study design needed to be restricted to a specific topic to improve comparability among data sources.

Alongside these methodological insights, our approach can be of value in supporting policymakers to gauge polarization on controversial topics, such as COVID-19 prevention measures. Here, we observe that compulsory vaccination is a very polarizing prevention measure in the DACH region and needs special consideration when discussed in the public sphere. This observation agrees with previous studies that suggest that vaccinations are indeed polarizing (X. Jiang et al., 2021; Schmidt et al., 2018).

Limitations

While considering multiple perspectives provides a holistic view on polarization effects, we identify three limitations of our work.

Firstly, we focus on *polarization as a state* instead of also considering the definition of polarization as a process by DiMaggio et al. (1996). However, temporal effects could play a major role in the understanding of how a topic gets polarized in the first place. Thus, considering polarization as a state only could greatly influence the interpretation of the results. While temporal information was available for Twitter data, the survey data is available only for the time of fieldwork. Moreover, using the short time span of the experiment would likely not reveal interesting dynamics in the process.

Secondly, there might be *potential biases* in the data, especially in terms of the respondents who share their accounts. While we briefly discussed the differences between survey respondents in general, survey respondents using Twitter, and survey respondents who shared their Twitter accounts, we did not perform an in-depth analysis of the characteristics of the respondents who shared their Twitter accounts. This might introduce biases, that is, selection and observation biases, into the analysis of tweets. We suspect that certain characteristics favoring account sharing could also explain the less polarizing nature of our Twitter sample. For instance, we presume that users with extreme positions might be reluctant to share their account information. Also, Twitter users are not necessarily users who post on Twitter but might be using the platform passively.

Thirdly, we again emphasize the challenging issues of *comparing survey data with Twitter data*, which are different by their very nature. Their integration lets us combine the advantages of both data types, but results in a small number of users and tweets for analysis. Since in our approach, we perform sentiment analysis on the tweets and measure agreement in the survey data to quantify polarization, our study is subject to the limitations of these techniques, as we discussed in the methods section.

Future Work

As for future work, we plan to reproduce this experiment in a follow-up survey on a larger sample size to further validate our results. To increase the amount of data in the integrated data, we will conduct the recruitment on the social media platforms to acquire more active users alongside the representative sample. Additionally, we aim to repeat the survey multiple times with the same set of users and questions. In these questionnaires, we will ask users to state their reason for sharing or not sharing their accounts, which allows the analysis of biases in the integration of data. Overall, this longitudinal study should provide in-depth insights into the process of how polarization changes over time.

Furthermore, we will also incorporate advanced models for opinion formation and spread in the social media analyses. For instance, we want to investigate how the multiple expressed opinions in tweets relate to the single innate opinion of a social media account user. Using these models, we will try to further improve the understanding of how online content relates to the survey answers.

Acknowledgements

We thank the rectorates of Graz University of Technology and University of Graz for their support of this research. We thank the anonymous reviewers for their helpful comments on this paper.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Markus Reiter-Haas https://orcid.org/0000-0001-9852-8206
Beate Klösch https://orcid.org/0000-0002-8061-6088
Markus Hadler https://orcid.org/0000-0002-0359-5789
Elisabeth Lex https://orcid.org/0000-0001-5293-2967

Supplemental Material

Supplemental material for this article is available online. The archive of the survey information is available at: https://data.aussda.at/dataset.xhtml?persistentId=doi:10.11587/OVHKTR. The repository with the code is available at: https://github.com/socialcomplab/sscr-opinion-polarization.

Notes

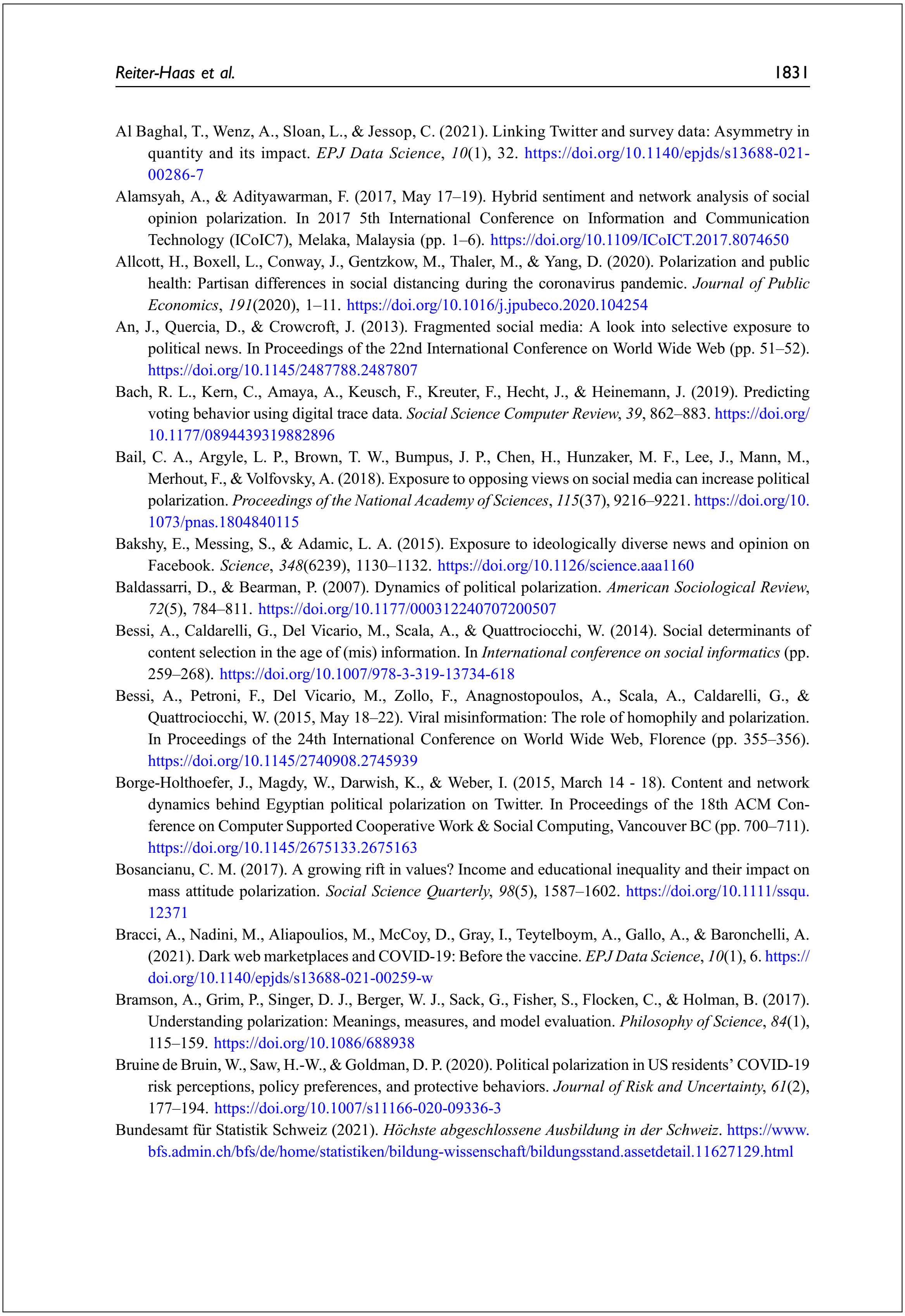
1. https://developer.twitter.com/en/docs/tutorials/consuming-streaming-data Note that, while the sample is supposedly random, Pfeffer et al. (2018) showed that it should not be regard as such due limitations in the sampling algorithm.
2. https://github.com/DocNow/twarc
3. This is due to the Twitter policy that researchers are only allowed to share tweet IDs instead of the complete tweets: https://developer.twitter.com/en/developer-terms/agreement-and-policy
4. We also experimented with topic models such as Latent Dirichlet allocation (LDA), but perceived the interpretation of the resulting topics and their unsatisfactory quality as a hindrance in our analysis.
5. https://textblob-de.readthedocs.io
6. The questionnaire items are provided in the Supplemental Materials

References

Abril, E. P. (2018). Subduing attitude polarization?: How partisan news may not affect attitude polarization for online publics. *Politics and the Life Sciences*, *37*(1), 68–77. https://doi.org/10.1017/pls.2017.1

Adamic, L. A., & Glance, N. (2005). The political blogosphere and the 2004 US election: Divided they blog. In Proceedings of the 3rd international workshop on Link discovery, Chicago, IL (pp. 36–43). https://doi.org/10.1145/1134271.1134277

Al Baghal, T., Sloan, L., Jessop, C., Williams, M. L., & Burnap, P. (2019). Linking Twitter and survey data: The impact of survey mode and demographics on consent rates across three uk studies. *Social Science Computer Review*, *38*(5), 517–532. https://doi.org/10.1177/0894439319828011



Al Baghal, T., Wenz, A., Sloan, L., & Jessop, C. (2021). Linking Twitter and survey data: Asymmetry in quantity and its impact. *EPJ Data Science*, *10*(1), 32. https://doi.org/10.1140/epjds/s13688-021-00286-7

Alamsyah, A., & Adityawarman, F. (2017, May 17–19). Hybrid sentiment and network analysis of social opinion polarization. In 2017 5th International Conference on Information and Communication Technology (ICoIC7), Melaka, Malaysia (pp. 1–6). https://doi.org/10.1109/ICoICT.2017.8074650

Allcott, H., Boxell, L., Conway, J., Gentzkow, M., Thaler, M., & Yang, D. (2020). Polarization and public health: Partisan differences in social distancing during the coronavirus pandemic. *Journal of Public Economics*, *191*(2020), 1–11. https://doi.org/10.1016/j.jpubeco.2020.104254

An, J., Quercia, D., & Crowcroft, J. (2013). Fragmented social media: A look into selective exposure to political news. In Proceedings of the 22nd International Conference on World Wide Web (pp. 51–52). https://doi.org/10.1145/2487788.2487807

Bach, R. L., Kern, C., Amaya, A., Keusch, F., Kreuter, F., Hecht, J., & Heinemann, J. (2019). Predicting voting behavior using digital trace data. *Social Science Computer Review*, *39*, 862–883. https://doi.org/10.1177/0894439319882896

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. F., Lee, J., Mann, M., Merhout, F., & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, *115*(37), 9216–9221. https://doi.org/10.1073/pnas.1804840115

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*(6239), 1130–1132. https://doi.org/10.1126/science.aaa1160

Baldassarri, D., & Bearman, P. (2007). Dynamics of political polarization. *American Sociological Review*, *72*(5), 784–811. https://doi.org/10.1177/000312240707200507

Bessi, A., Caldarelli, G., Del Vicario, M., Scala, A., & Quattrociocchi, W. (2014). Social determinants of content selection in the age of (mis) information. In *International conference on social informatics* (pp. 259–268). https://doi.org/10.1007/978-3-319-13734-618

Bessi, A., Petroni, F., Del Vicario, M., Zollo, F., Anagnostopoulos, A., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2015, May 18–22). Viral misinformation: The role of homophily and polarization. In Proceedings of the 24th International Conference on World Wide Web, Florence (pp. 355–356). https://doi.org/10.1145/2740908.2745939

Borge-Holthoefer, J., Magdy, W., Darwish, K., & Weber, I. (2015, March 14 - 18). Content and network dynamics behind Egyptian political polarization on Twitter. In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, Vancouver BC (pp. 700–711). https://doi.org/10.1145/2675133.2675163

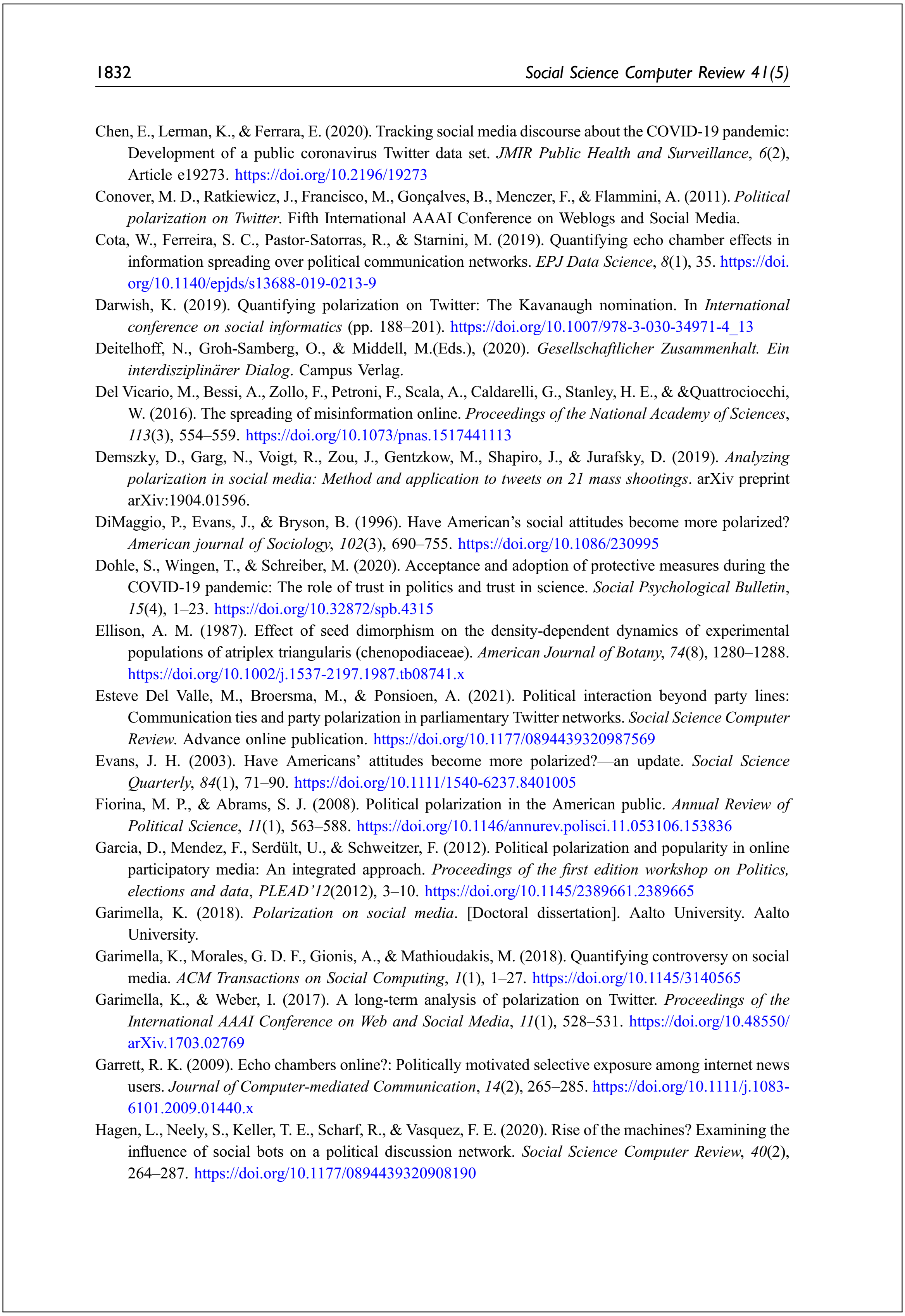
Bosancianu, C. M. (2017). A growing rift in values? Income and educational inequality and their impact on mass attitude polarization. *Social Science Quarterly*, *98*(5), 1587–1602. https://doi.org/10.1111/ssqu.12371

Bracci, A., Nadini, M., Aliapoulios, M., McCoy, D., Gray, I., Teytelboym, A., Gallo, A., & Baronchelli, A. (2021). Dark web marketplaces and COVID-19: Before the vaccine. *EPJ Data Science*, *10*(1), 6. https://doi.org/10.1140/epjds/s13688-021-00259-w

Bramson, A., Grim, P., Singer, D. J., Berger, W. J., Sack, G., Fisher, S., Flocken, C., & Holman, B. (2017). Understanding polarization: Meanings, measures, and model evaluation. *Philosophy of Science*, *84*(1), 115–159. https://doi.org/10.1086/688938

Bruine de Bruin, W., Saw, H.-W., & Goldman, D. P. (2020). Political polarization in US residents' COVID-19 risk perceptions, policy preferences, and protective behaviors. *Journal of Risk and Uncertainty*, *61*(2), 177–194. https://doi.org/10.1007/s11166-020-09336-3

Bundesamt für Statistik Schweiz (2021). *Höchste abgeschlossene Ausbildung in der Schweiz*. https://www.bfs.admin.ch/bfs/de/home/statistiken/bildung-wissenschaft/bildungsstand.assetdetail.11627129.html



Chen, E., Lerman, K., & Ferrara, E. (2020). Tracking social media discourse about the COVID-19 pandemic: Development of a public coronavirus Twitter data set. *JMIR Public Health and Surveillance*, *6*(2), Article e19273. https://doi.org/10.2196/19273

Conover, M. D., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., & Flammini, A. (2011). *Political polarization on Twitter*. Fifth International AAAI Conference on Weblogs and Social Media.

Cota, W., Ferreira, S. C., Pastor-Satorras, R., & Starnini, M. (2019). Quantifying echo chamber effects in information spreading over political communication networks. *EPJ Data Science*, *8*(1), 35. https://doi.org/10.1140/epjds/s13688-019-0213-9

Darwish, K. (2019). Quantifying polarization on Twitter: The Kavanaugh nomination. In *International conference on social informatics* (pp. 188–201). https://doi.org/10.1007/978-3-030-34971-4_13

Deitelhoff, N., Groh-Samberg, O., & Middell, M.(Eds.), (2020). *Gesellschaftlicher Zusammenhalt. Ein interdisziplinärer Dialog*. Campus Verlag.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & &Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, *113*(3), 554–559. https://doi.org/10.1073/pnas.1517441113

Demszky, D., Garg, N., Voigt, R., Zou, J., Gentzkow, M., Shapiro, J., & Jurafsky, D. (2019). *Analyzing polarization in social media: Method and application to tweets on 21 mass shootings*. arXiv preprint arXiv:1904.01596.

DiMaggio, P., Evans, J., & Bryson, B. (1996). Have American's social attitudes become more polarized? *American journal of Sociology*, *102*(3), 690–755. https://doi.org/10.1086/230995

Dohle, S., Wingen, T., & Schreiber, M. (2020). Acceptance and adoption of protective measures during the COVID-19 pandemic: The role of trust in politics and trust in science. *Social Psychological Bulletin*, *15*(4), 1–23. https://doi.org/10.32872/spb.4315

Ellison, A. M. (1987). Effect of seed dimorphism on the density-dependent dynamics of experimental populations of atriplex triangularis (chenopodiaceae). *American Journal of Botany*, *74*(8), 1280–1288. https://doi.org/10.1002/j.1537-2197.1987.tb08741.x

Esteve Del Valle, M., Broersma, M., & Ponsioen, A. (2021). Political interaction beyond party lines: Communication ties and party polarization in parliamentary Twitter networks. *Social Science Computer Review*. Advance online publication. https://doi.org/10.1177/0894439320987569

Evans, J. H. (2003). Have Americans' attitudes become more polarized?—an update. *Social Science Quarterly*, *84*(1), 71–90. https://doi.org/10.1111/1540-6237.8401005

Fiorina, M. P., & Abrams, S. J. (2008). Political polarization in the American public. *Annual Review of Political Science*, *11*(1), 563–588. https://doi.org/10.1146/annurev.polisci.11.053106.153836

Garcia, D., Mendez, F., Serdült, U., & Schweitzer, F. (2012). Political polarization and popularity in online participatory media: An integrated approach. *Proceedings of the first edition workshop on Politics, elections and data*, *PLEAD'12*(2012), 3–10. https://doi.org/10.1145/2389661.2389665

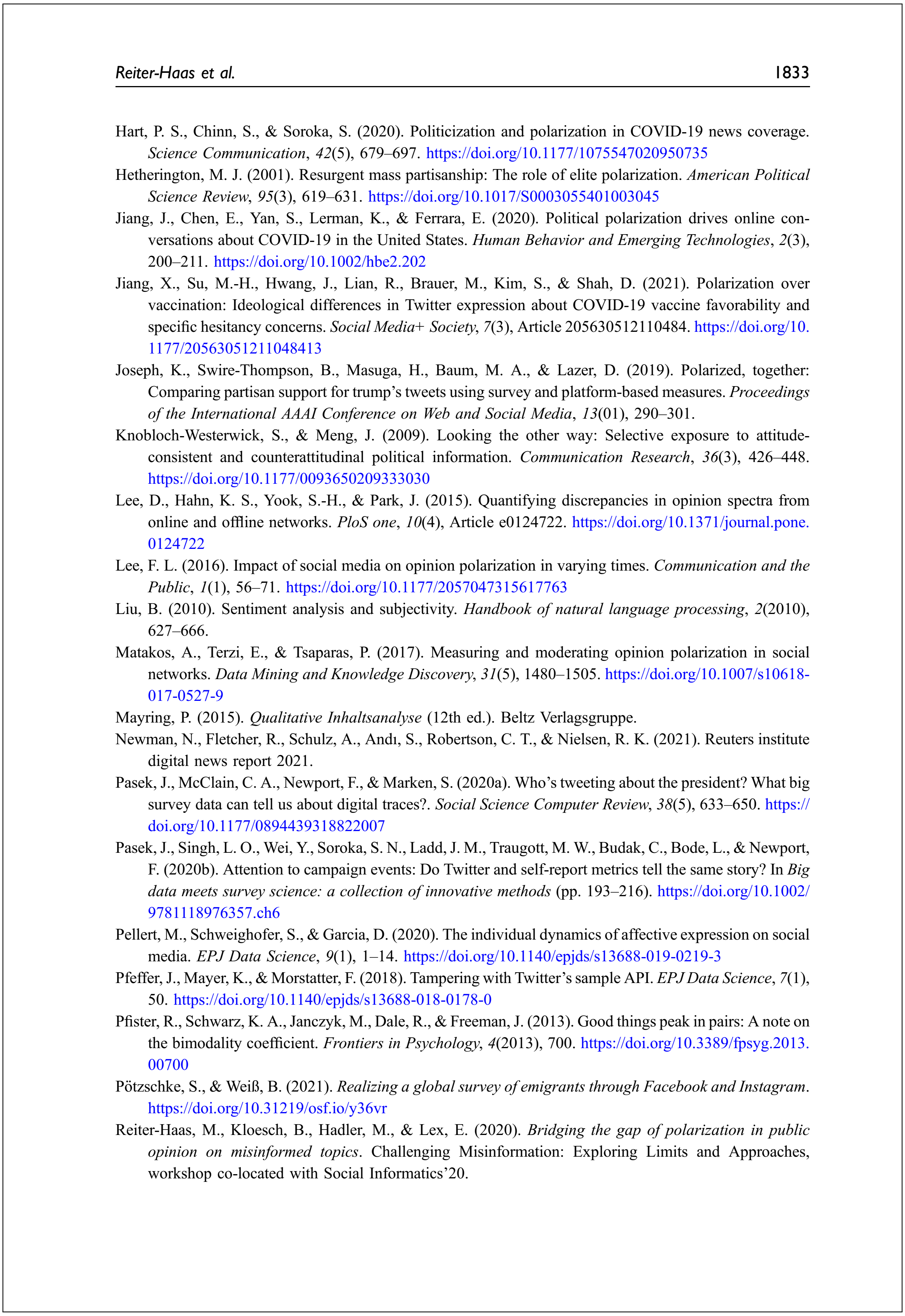
Garimella, K. (2018). *Polarization on social media*. [Doctoral dissertation]. Aalto University. Aalto University.

Garimella, K., Morales, G. D. F., Gionis, A., & Mathioudakis, M. (2018). Quantifying controversy on social media. *ACM Transactions on Social Computing*, *1*(1), 1–27. https://doi.org/10.1145/3140565

Garimella, K., & Weber, I. (2017). A long-term analysis of polarization on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1), 528–531. https://doi.org/10.48550/arXiv.1703.02769

Garrett, R. K. (2009). Echo chambers online?: Politically motivated selective exposure among internet news users. *Journal of Computer-mediated Communication*, *14*(2), 265–285. https://doi.org/10.1111/j.1083-6101.2009.01440.x

Hagen, L., Neely, S., Keller, T. E., Scharf, R., & Vasquez, F. E. (2020). Rise of the machines? Examining the influence of social bots on a political discussion network. *Social Science Computer Review*, *40*(2), 264–287. https://doi.org/10.1177/0894439320908190



Hart, P. S., Chinn, S., & Soroka, S. (2020). Politicization and polarization in COVID-19 news coverage. *Science Communication*, *42*(5), 679–697. https://doi.org/10.1177/1075547020950735

Hetherington, M. J. (2001). Resurgent mass partisanship: The role of elite polarization. *American Political Science Review*, *95*(3), 619–631. https://doi.org/10.1017/S0003055401003045

Jiang, J., Chen, E., Yan, S., Lerman, K., & Ferrara, E. (2020). Political polarization drives online conversations about COVID-19 in the United States. *Human Behavior and Emerging Technologies*, *2*(3), 200–211. https://doi.org/10.1002/hbe2.202

Jiang, X., Su, M.-H., Hwang, J., Lian, R., Brauer, M., Kim, S., & Shah, D. (2021). Polarization over vaccination: Ideological differences in Twitter expression about COVID-19 vaccine favorability and specific hesitancy concerns. *Social Media+ Society*, *7*(3), Article 205630512110484. https://doi.org/10.1177/20563051211048413

Joseph, K., Swire-Thompson, B., Masuga, H., Baum, M. A., & Lazer, D. (2019). Polarized, together: Comparing partisan support for trump's tweets using survey and platform-based measures. *Proceedings of the International AAAI Conference on Web and Social Media*, *13*(01), 290–301.

Knobloch-Westerwick, S., & Meng, J. (2009). Looking the other way: Selective exposure to attitude-consistent and counterattitudinal political information. *Communication Research*, *36*(3), 426–448. https://doi.org/10.1177/0093650209333030

Lee, D., Hahn, K. S., Yook, S.-H., & Park, J. (2015). Quantifying discrepancies in opinion spectra from online and offline networks. *PloS one*, *10*(4), Article e0124722. https://doi.org/10.1371/journal.pone.0124722

Lee, F. L. (2016). Impact of social media on opinion polarization in varying times. *Communication and the Public*, *1*(1), 56–71. https://doi.org/10.1177/2057047315617763

Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of natural language processing*, *2*(2010), 627–666.

Matakos, A., Terzi, E., & Tsaparas, P. (2017). Measuring and moderating opinion polarization in social networks. *Data Mining and Knowledge Discovery*, *31*(5), 1480–1505. https://doi.org/10.1007/s10618-017-0527-9

Mayring, P. (2015). *Qualitative Inhaltsanalyse* (12th ed.). Beltz Verlagsgruppe.

Newman, N., Fletcher, R., Schulz, A., Andı, S., Robertson, C. T., & Nielsen, R. K. (2021). Reuters institute digital news report 2021.

Pasek, J., McClain, C. A., Newport, F., & Marken, S. (2020a). Who's tweeting about the president? What big survey data can tell us about digital traces?. *Social Science Computer Review*, *38*(5), 633–650. https://doi.org/10.1177/0894439318822007

Pasek, J., Singh, L. O., Wei, Y., Soroka, S. N., Ladd, J. M., Traugott, M. W., Budak, C., Bode, L., & Newport, F. (2020b). Attention to campaign events: Do Twitter and self-report metrics tell the same story? In *Big data meets survey science: a collection of innovative methods* (pp. 193–216). https://doi.org/10.1002/9781118976357.ch6

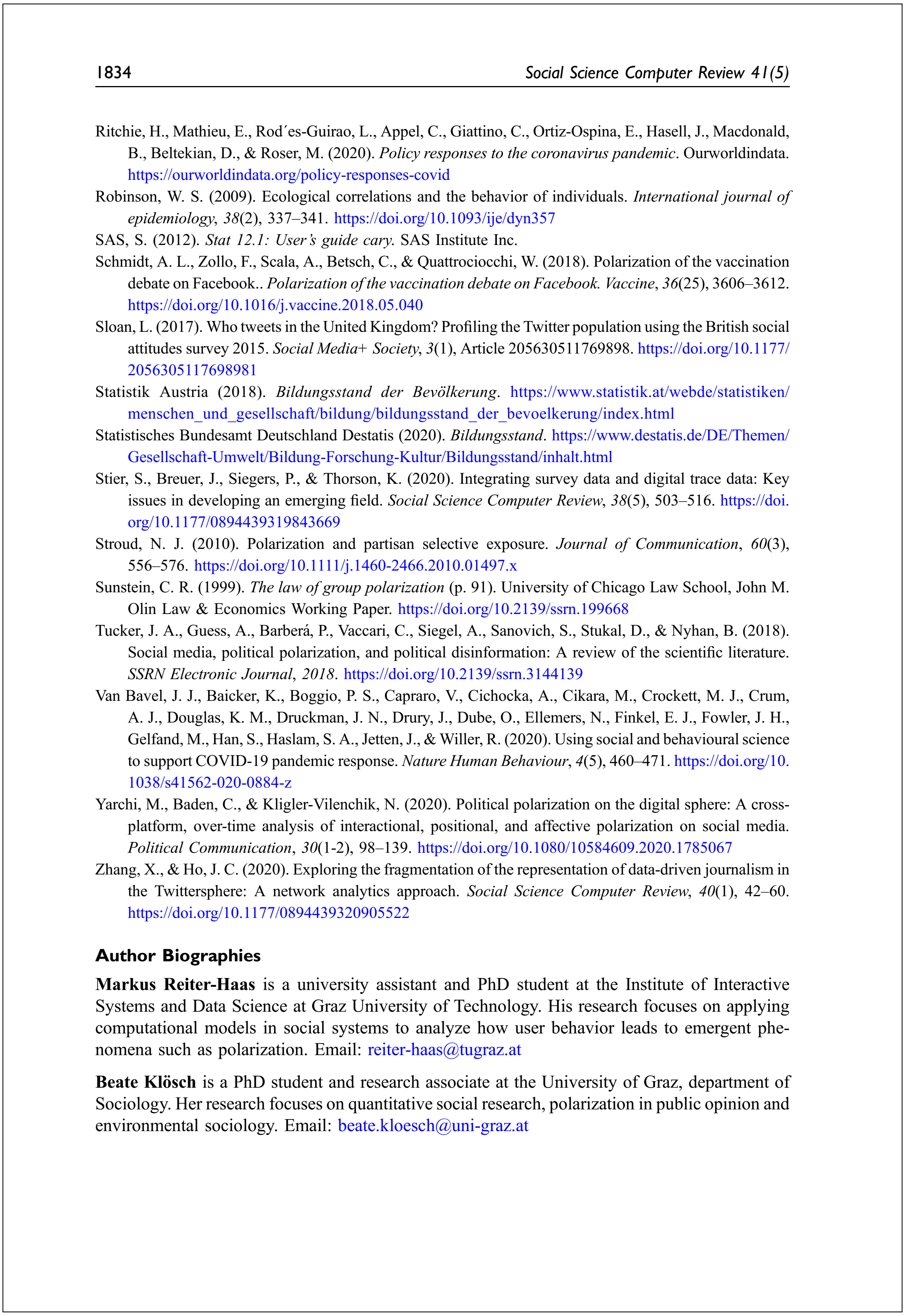
Pellert, M., Schweighofer, S., & Garcia, D. (2020). The individual dynamics of affective expression on social media. *EPJ Data Science*, *9*(1), 1–14. https://doi.org/10.1140/epjds/s13688-019-0219-3

Pfeffer, J., Mayer, K., & Morstatter, F. (2018). Tampering with Twitter's sample API. *EPJ Data Science*, *7*(1), 50. https://doi.org/10.1140/epjds/s13688-018-0178-0

Pfister, R., Schwarz, K. A., Janczyk, M., Dale, R., & Freeman, J. (2013). Good things peak in pairs: A note on the bimodality coefficient. *Frontiers in Psychology*, *4*(2013), 700. https://doi.org/10.3389/fpsyg.2013.00700

Pötzschke, S., & Weiß, B. (2021). *Realizing a global survey of emigrants through Facebook and Instagram*. https://doi.org/10.31219/osf.io/y36vr

Reiter-Haas, M., Kloesch, B., Hadler, M., & Lex, E. (2020). *Bridging the gap of polarization in public opinion on misinformed topics*. Challenging Misinformation: Exploring Limits and Approaches, workshop co-located with Social Informatics'20.



Ritchie, H., Mathieu, E., Rod´es-Guirao, L., Appel, C., Giattino, C., Ortiz-Ospina, E., Hasell, J., Macdonald, B., Beltekian, D., & Roser, M. (2020). *Policy responses to the coronavirus pandemic*. Ourworldindata. https://ourworldindata.org/policy-responses-covid

Robinson, W. S. (2009). Ecological correlations and the behavior of individuals. *International journal of epidemiology*, *38*(2), 337–341. https://doi.org/10.1093/ije/dyn357

SAS, S. (2012). *Stat 12.1: User's guide cary*. SAS Institute Inc.

Schmidt, A. L., Zollo, F., Scala, A., Betsch, C., & Quattrociocchi, W. (2018). Polarization of the vaccination debate on Facebook.. *Polarization of the vaccination debate on Facebook. Vaccine*, *36*(25), 3606–3612. https://doi.org/10.1016/j.vaccine.2018.05.040

Sloan, L. (2017). Who tweets in the United Kingdom? Profiling the Twitter population using the British social attitudes survey 2015. *Social Media+ Society*, *3*(1), Article 205630511769898. https://doi.org/10.1177/2056305117698981

Statistik Austria (2018). *Bildungsstand der Bevölkerung*. https://www.statistik.at/webde/statistiken/menschen_und_gesellschaft/bildung/bildungsstand_der_bevoelkerung/index.html

Statistisches Bundesamt Deutschland Destatis (2020). *Bildungsstand*. https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bildung-Forschung-Kultur/Bildungsstand/inhalt.html

Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2020). Integrating survey data and digital trace data: Key issues in developing an emerging field. *Social Science Computer Review*, *38*(5), 503–516. https://doi.org/10.1177/0894439319843669

Stroud, N. J. (2010). Polarization and partisan selective exposure. *Journal of Communication*, *60*(3), 556–576. https://doi.org/10.1111/j.1460-2466.2010.01497.x

Sunstein, C. R. (1999). *The law of group polarization* (p. 91). University of Chicago Law School, John M. Olin Law & Economics Working Paper. https://doi.org/10.2139/ssrn.199668

Tucker, J. A., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., & Nyhan, B. (2018). Social media, political polarization, and political disinformation: A review of the scientific literature. *SSRN Electronic Journal*, *2018*. https://doi.org/10.2139/ssrn.3144139

Van Bavel, J. J., Baicker, K., Boggio, P. S., Capraro, V., Cichocka, A., Cikara, M., Crockett, M. J., Crum, A. J., Douglas, K. M., Druckman, J. N., Drury, J., Dube, O., Ellemers, N., Finkel, E. J., Fowler, J. H., Gelfand, M., Han, S., Haslam, S. A., Jetten, J., & Willer, R. (2020). Using social and behavioural science to support COVID-19 pandemic response. *Nature Human Behaviour*, *4*(5), 460–471. https://doi.org/10.1038/s41562-020-0884-z

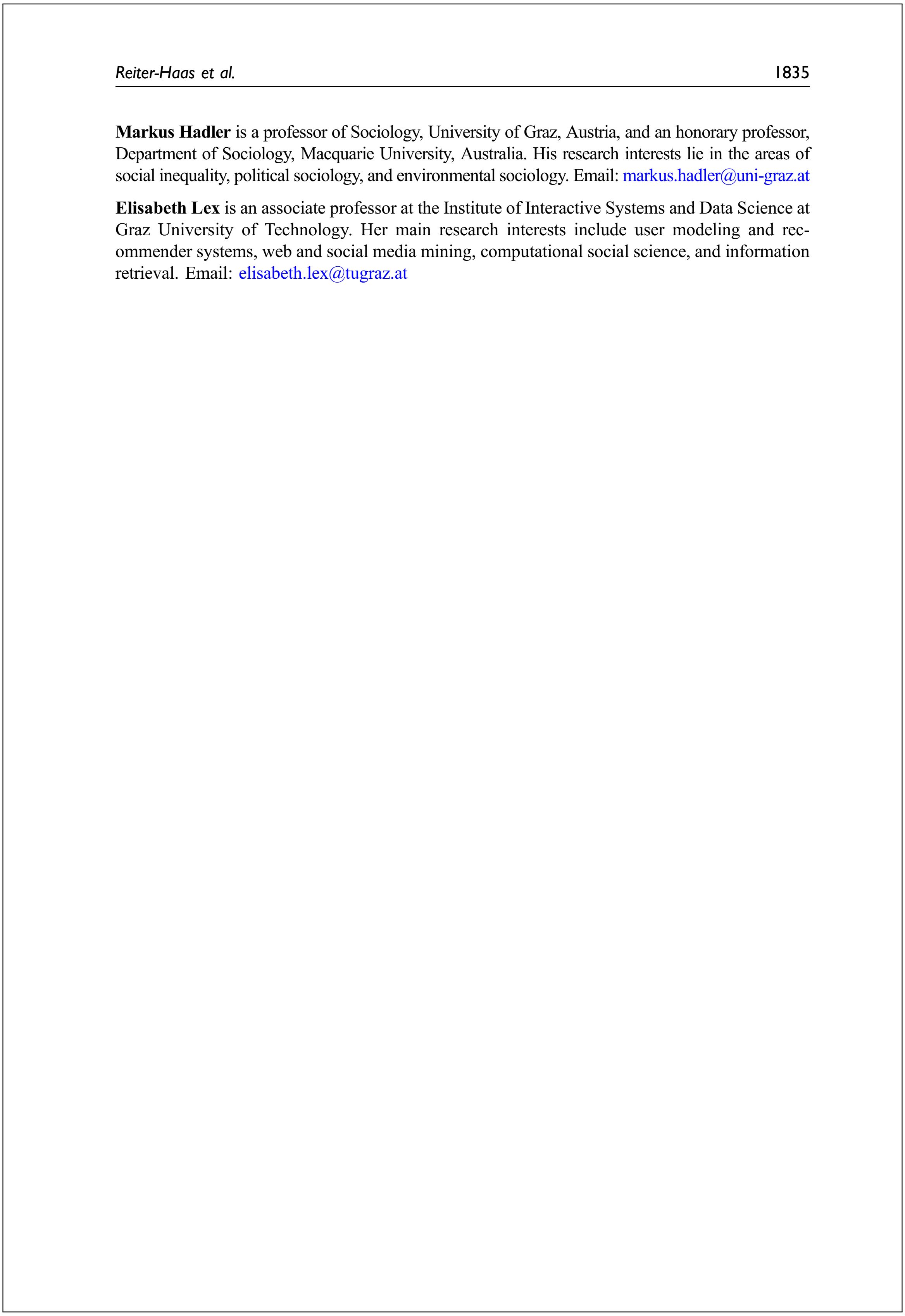
Yarchi, M., Baden, C., & Kligler-Vilenchik, N. (2020). Political polarization on the digital sphere: A cross-platform, over-time analysis of interactional, positional, and affective polarization on social media. *Political Communication*, *30*(1-2), 98–139. https://doi.org/10.1080/10584609.2020.1785067

Zhang, X., & Ho, J. C. (2020). Exploring the fragmentation of the representation of data-driven journalism in the Twittersphere: A network analytics approach. *Social Science Computer Review*, *40*(1), 42–60. https://doi.org/10.1177/0894439320905522

Author Biographies

Markus Reiter-Haas is a university assistant and PhD student at the Institute of Interactive Systems and Data Science at Graz University of Technology. His research focuses on applying computational models in social systems to analyze how user behavior leads to emergent phenomena such as polarization. Email: reiter-haas@tugraz.at

Beate Klösch is a PhD student and research associate at the University of Graz, department of Sociology. Her research focuses on quantitative social research, polarization in public opinion and environmental sociology. Email: beate.kloesch@uni-graz.at



Markus Hadler is a professor of Sociology, University of Graz, Austria, and an honorary professor, Department of Sociology, Macquarie University, Australia. His research interests lie in the areas of social inequality, political sociology, and environmental sociology. Email: markus.hadler@uni-graz.at

Elisabeth Lex is an associate professor at the Institute of Interactive Systems and Data Science at Graz University of Technology. Her main research interests include user modeling and recommender systems, web and social media mining, computational social science, and information retrieval. Email: elisabeth.lex@tugraz.at

Exploration of Framing Biases in Polarized Online Content Consumption

Markus Reiter-Haas
supervised by Elisabeth Lex
Graz University of Technology, Institute of Interactive Systems and Data Science
Graz, Austria
reiter-haas@tugraz.at

ABSTRACT

The study of framing bias on the Web is crucial in our digital age, as the framing of information can influence human behavior and decision on critical issues such as health or politics. Traditional frame analysis requires a curated set of frames derived from manual content analysis by domain experts. In this work, we introduce a frame analysis approach based on pretrained Transformer models that let us capture frames in an exploratory manner beyond predefined frames. In our experiments on two public online news and social media datasets, we show that our approach lets us identify underexplored conceptualizations, such as that health-related content is framed in terms of beliefs for conspiracy media, while mainstream media is instead concerned with science. We anticipate our work to be a starting point for further research on exploratory computational framing analysis using pretrained Transformers.

CCS CONCEPTS

• **Information systems** → *Content analysis and feature selection*; **World Wide Web**; Language models.

KEYWORDS

computational frame extraction, content bias, exploratory content analysis, text processing, semantic representations

ACM Reference Format:
Markus Reiter-Haas. 2023. Exploration of Framing Biases in Polarized Online Content Consumption. In *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3543873.3587534

1 INTRODUCTION

The Web affects society at large, but also reflects the inherent biases of people [4]. Biases have been studied extensively regarding online behavior patterns, e.g., in terms of popularity bias [1, 22, 25] and confirmation bias [20, 21, 43]. Beyond behavioral patterns, biases can also stem from Web content itself. Herein, Draws et al. [16] show that the viewpoint of biased content influences user attitudes, while Rekabsaz et al. [40] highlight the impact of societal biases (e.g., gender) in retrieved content on the representation of particular groups in retrieval results. Similarly, the way content is *framed* can lead to biases and has also been shown to affect human behavior, public opinion and decision-making [45]. Framing corresponds to the selection and saliency of certain aspects in communicating texts [17]. Although research on framing has been thoroughly conducted for media [e.g., 11, 24], framing remains largely unexplored in Web content and its users' consumption patterns, in particular, when content is polarized or negative and thus receiving increased attention [30, 34]. Also, traditional frame analysis techniques frequently require a set of known frames for a topic, which need to be identified manually by domain experts [24].

In this work, we aim to explore framing biases in polarized Web content and their effects on content consumption behavior. We introduce three complementary approaches for exploratory framing analysis based on pretrained Transformers [46]. We subsequently categorize the extracted frames and conduct behavior analysis in openly available online news and social media corpora [28, 48]. We find that polarized health-related news is largely framed in terms of science vs. beliefs. Given that this frame is not part of established prior conceptualizations [e.g., 11], this finding underpins the merits of our approach. We believe that our research will considerably improve the understanding of framing bias and users' consumption patterns on the Web. Besides, we hope that our approach inspires novel debiasing methods to mitigate polarized Web-based retrieval.

This work is licensed under a Creative Commons Attribution International 4.0 License.

WWW '23 Companion, April 30–May 04, 2023, Austin, TX, USA
© 2023 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9419-2/23/04.
https://doi.org/10.1145/3543873.3587534

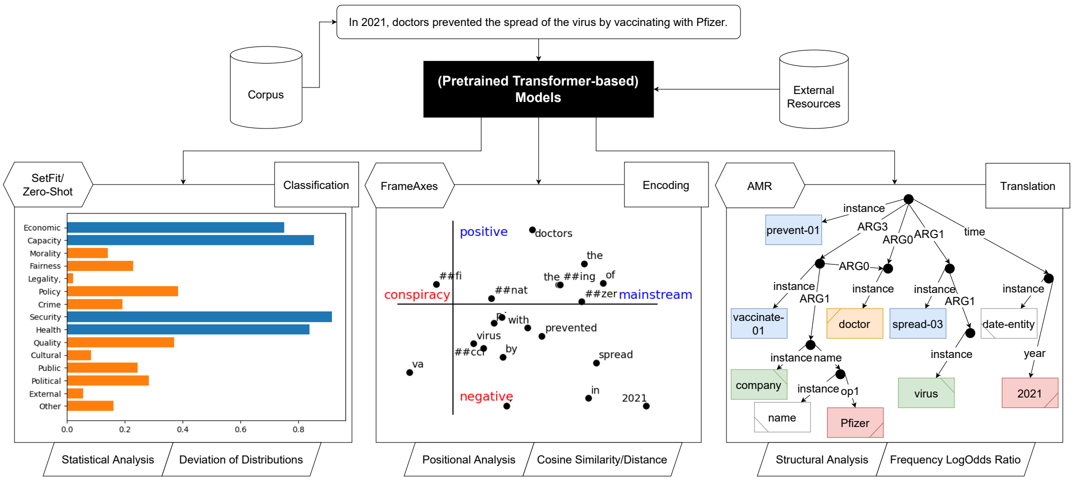
2 PROBLEM

Framing of digital media relates to societal effects such as polarization. However, framing biases are difficult to detect and characterize. Moreover, acquiring labeled data on framing is challenging and labor-intensive. This issue becomes even more apparent in non-English settings, where the lack of data is more severe.

In our work, we investigate text representation, such as embeddings, to capture the semantic differences related to framing in text. As our approach is exploratory in nature, it applies to a setting with a low amount of labeled or entirely unlabeled data. Similarly, our approach is not restricted to a single language, as it is based on language models, and hence, can use a multilingual base encoder.

We aim to answer three research questions on framing analysis:

RQ1: How can we extract frames from polarized Web content without prior conceptualization?
RQ2: How can the extracted frames be categorized for specific contexts, e.g., health-related topics?
RQ3: What is the relationship between frames and viewpoint diversity in users' Web content consumption?



(a) Framing Labels via Label Probabilities

(b) Framing Dimensions via Embedding Space

(c) Framing Structure via Semantic Graph

Figure 1: Overview of the three complementary approaches. Subfigures show the result of a transformation with each approach. In (a), a predefined set of labels is predicted, here in a zero-shot setting, and the label probabilities are plotted (blue for predicted labels with high probability). In (b), the tokens of the sentence are projected onto framing axes in 2-dimensional embedding space, where the axis poles are opposing each other (e.g., positive vs negative). In (c), the text is transformed into a semantic (rooted, directed, and acyclic) graph.

3 STATE OF THE ART

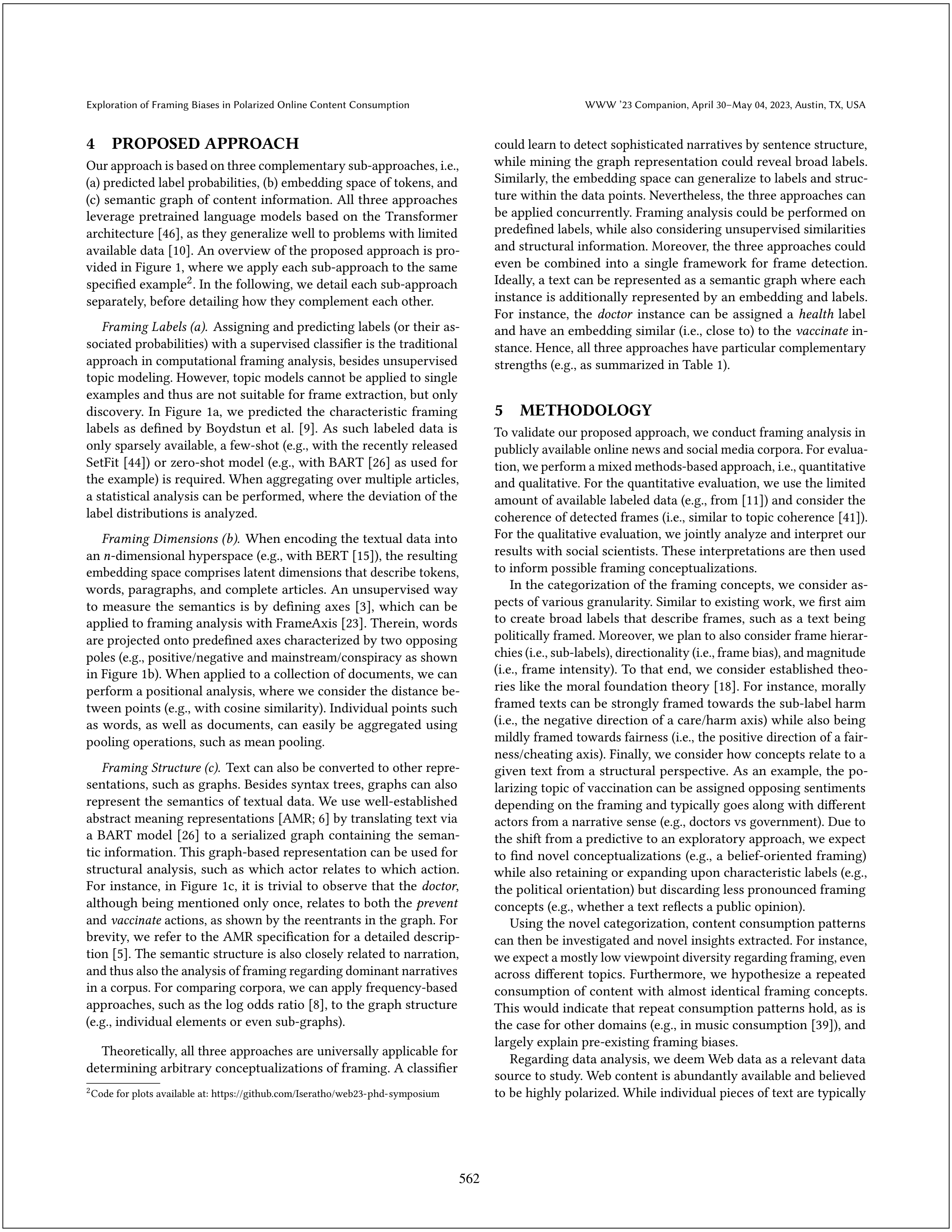
Framing is a fractured paradigm in literature, but essentially deals with the *selection* and *salience* of some aspects of a communicating text [17]. For example, measures against the COVID-19 pandemic can be framed in terms of the *prevention of the spread* or *fight against the virus*, thereby highlighting distinct features of the problem and suggesting opposing solutions. As framing promotes the alteration of the perceived reality and its interpretation [17], it also affects human judgments and choices [45]. Consequently, social scientists have researched the framing of important topics, such as the responses and social movements towards the COVID-19 pandemic [19, 27, 33]. Traditionally, framing analysis involves careful manual analysis of data. More recently, computational methods [e.g., 42, 47] have been suggested to automatically determine the framing of textual content. For example, the *SemEval 2023 Task 3 (Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup)* [32] [1]. aims to predict the framing of text based on a predefined taxonomy.

Computational frame analysis studies comprise various types of frames, such as war [47], terrorists [14], morality [29], or blame [42]. These studies focus on different conceptualizations of framing and are not necessarily comparable. Moreover, their methodological approaches differ drastically and depend on the preselected frames for the study. Hence, other conceptualizations of framing that would be more characteristic might not be detected.

Ali and Hassan [2] provide a comprehensive survey of computational framing extraction methods. The main approaches include various kinds of topic modeling and cluster analysis as unsupervised approaches. In comparison, neural networks, also including pretrained Transformer-based language models, are mainly used in a supervised manner. Other methods include parsing semantic relations, frequency-based models, and semantic axes, i.e., FrameAxis [23]. The current state-of-the-art predominantly investigates a predefined set of frames. We strive to alleviate this limitation by utilizing an exploratory approach based on semantic information embedded within the textual content. We adapt several of the previously mentioned methods, i.e., neural networks with pretrained Transformer-based language models, semantic relations, frequency-based models, and semantic axes. Due to the exploratory manner, our approach enables novel and unexpected new conceptualizations predefined selection of frames or labels. This is unlike the existing *OpenFraming* tool [7], which, although exploratory in nature, still requires a preselection of frames and labeling of data.

Finally, framing theory [12] relates to various other concepts, such as public opinion and values. Especially in media frames [13], narratives are another important aspect to consider. Hence, other computational endeavors like computational narrative understanding [31] benefit from improved frame extraction methods. For brevity, we omit a detailed discussion here and refer to Reiter-Haas et al. [37], where we thoroughly discuss the relationship between narratives and framing.

[1] Challenge Website: https://propaganda.math.unipd.it/semeval2023task3/



4 PROPOSED APPROACH

Our approach is based on three complementary sub-approaches, i.e., (a) predicted label probabilities, (b) embedding space of tokens, and (c) semantic graph of content information. All three approaches leverage pretrained language models based on the Transformer architecture [46], as they generalize well to problems with limited available data [10]. An overview of the proposed approach is provided in Figure 1, where we apply each sub-approach to the same specified example[2]. In the following, we detail each sub-approach separately, before detailing how they complement each other.

Framing Labels (a). Assigning and predicting labels (or their associated probabilities) with a supervised classifier is the traditional approach in computational framing analysis, besides unsupervised topic modeling. However, topic models cannot be applied to single examples and thus are not suitable for frame extraction, but only discovery. In Figure 1a, we predicted the characteristic framing labels as defined by Boydstun et al. [9]. As such labeled data is only sparsely available, a few-shot (e.g., with the recently released SetFit [44]) or zero-shot model (e.g., with BART [26] as used for the example) is required. When aggregating over multiple articles, a statistical analysis can be performed, where the deviation of the label distributions is analyzed.

Framing Dimensions (b). When encoding the textual data into an *n*-dimensional hyperspace (e.g., with BERT [15]), the resulting embedding space comprises latent dimensions that describe tokens, words, paragraphs, and complete articles. An unsupervised way to measure the semantics is by defining axes [3], which can be applied to framing analysis with FrameAxis [23]. Therein, words are projected onto predefined axes characterized by two opposing poles (e.g., positive/negative and mainstream/conspiracy as shown in Figure 1b). When applied to a collection of documents, we can perform a positional analysis, where we consider the distance between points (e.g., with cosine similarity). Individual points such as words, as well as documents, can easily be aggregated using pooling operations, such as mean pooling.

Framing Structure (c). Text can also be converted to other representations, such as graphs. Besides syntax trees, graphs can also represent the semantics of textual data. We use well-established abstract meaning representations [AMR; 6] by translating text via a BART model [26] to a serialized graph containing the semantic information. This graph-based representation can be used for structural analysis, such as which actor relates to which action. For instance, in Figure 1c, it is trivial to observe that the *doctor*, although being mentioned only once, relates to both the *prevent* and *vaccinate* actions, as shown by the reentrants in the graph. For brevity, we refer to the AMR specification for a detailed description [5]. The semantic structure is also closely related to narration, and thus also the analysis of framing regarding dominant narratives in a corpus. For comparing corpora, we can apply frequency-based approaches, such as the log odds ratio [8], to the graph structure (e.g., individual elements or even sub-graphs).

Theoretically, all three approaches are universally applicable for determining arbitrary conceptualizations of framing. A classifier could learn to detect sophisticated narratives by sentence structure, while mining the graph representation could reveal broad labels. Similarly, the embedding space can generalize to labels and structure within the data points. Nevertheless, the three approaches can be applied concurrently. Framing analysis could be performed on predefined labels, while also considering unsupervised similarities and structural information. Moreover, the three approaches could even be combined into a single framework for frame detection. Ideally, a text can be represented as a semantic graph where each instance is additionally represented by an embedding and labels. For instance, the *doctor* instance can be assigned a *health* label and have an embedding similar (i.e., close to) to the *vaccinate* instance. Hence, all three approaches have particular complementary strengths (e.g., as summarized in Table 1).

5 METHODOLOGY

To validate our proposed approach, we conduct framing analysis in publicly available online news and social media corpora. For evaluation, we perform a mixed methods-based approach, i.e., quantitative and qualitative. For the quantitative evaluation, we use the limited amount of available labeled data (e.g., from [11]) and consider the coherence of detected frames (i.e., similar to topic coherence [41]). For the qualitative evaluation, we jointly analyze and interpret our results with social scientists. These interpretations are then used to inform possible framing conceptualizations.

In the categorization of the framing concepts, we consider aspects of various granularity. Similar to existing work, we first aim to create broad labels that describe frames, such as a text being politically framed. Moreover, we plan to also consider frame hierarchies (i.e., sub-labels), directionality (i.e., frame bias), and magnitude (i.e., frame intensity). To that end, we consider established theories like the moral foundation theory [18]. For instance, morally framed texts can be strongly framed towards the sub-label harm (i.e., the negative direction of a care/harm axis) while also being mildly framed towards fairness (i.e., the positive direction of a fairness/cheating axis). Finally, we consider how concepts relate to a given text from a structural perspective. As an example, the polarizing topic of vaccination can be assigned opposing sentiments depending on the framing and typically goes along with different actors from a narrative sense (e.g., doctors vs government). Due to the shift from a predictive to an exploratory approach, we expect to find novel conceptualizations (e.g., a belief-oriented framing) while also retaining or expanding upon characteristic labels (e.g., the political orientation) but discarding less pronounced framing concepts (e.g., whether a text reflects a public opinion).

Using the novel categorization, content consumption patterns can then be investigated and novel insights extracted. For instance, we expect a mostly low viewpoint diversity regarding framing, even across different topics. Furthermore, we hypothesize a repeated consumption of content with almost identical framing concepts. This would indicate that repeat consumption patterns hold, as is the case for other domains (e.g., in music consumption [39]), and largely explain pre-existing framing biases.

Regarding data analysis, we deem Web data as a relevant data source to study. Web content is abundantly available and believed to be highly polarized. While individual pieces of text are typically

[2] Code for plots available at: https://github.com/Iseratho/web23-phd-symposium

Criteria	Classifier (a)	Embeddings (b)	Graph (c)
Unsupervised	×[†]	✓	✓
Exploratory	×	~	✓
Narratives	×	×	✓
Challenge	✓	×	×
Dimensions	scalar	*n*-D	irregular
Data Type	int/float	float	int
Aggregation	trivial	intuitive	challenging

Table 1: Summary of the comparison between the three sub-approaches. The complexity increases from left to right, but similarly increases in exploratory potential.

lacking manual framing annotations, online sources are often associated with certain features, such as political preferences and credibility. Hence, our approach specifically focuses on Web content, such as news websites and social media.

We aim to use recent and large datasets containing textual and log data, such as LOCO [28] for testing the approaches and frame categorization, and MIND [48] for content consumption analysis. LOCO provides online articles regarding a range of topics (including health-related ones), while also containing labels on whether they belong to conspiracy or mainstream media. Hence, we expect a noteworthy difference in framing between the two types of sources. MIND, on the other hand, provides click and impression logs (i.e., consumption data) in addition to content data. Thus, both datasets are suitable for their respective tasks.

6 RESULTS

In our initial paper on framing [38, similar to Figure 1b], we investigate the morality framing of political tweets in the US and Austria with word vectors and FrameAxis [23]. In the study, we find that the framing is coherent with previous findings on US politicians regarding their party's dominant morality. However, followers of Austrian politicians frame their tweets regarding COVID-19 similarly to topic-specific political messages in the public, rather than the usual party-associated dimensions. Opposite to expectations, the left-leaning Social Democratic Party emphasizes authority, while the ruling conservative party focuses on care. For context, at the time of data collection, the conservative party aimed to slow the spread of the virus with public messaging regarding mutual care. Conversely, the leader of the Social Democratic Party, being an epidemiologist herself, repeatedly insisted on listening to doctors and scientists. Hence, just focusing on a predefined conceptualization might lose valuable information regarding the framing of messages that might even lead to counterintuitive pictures.

In our next and most substantial contribution so far [37, i.e., Figure 1c], we demonstrate that semantic graphs are a perfect fit to represent the framing of the narrative information embedded in the textual content. Specifically, we use abstract meaning representations [6] to extract health-related narratives from the LOCO dataset [28] containing articles from both mainstream and conspiracy media. Although the approach requires no predefined conceptualization of framing, we extract differences that support our intuition. Most notably, we find that conspiracy media revolved around belief narratives, whereas mainstream media focused on science instead. Such conceptualization of framing goes beyond the typical analysis of framing classes and dimensions. Moreover, we investigated more specific narratives concerning the interplay of actors and actions. For instance, we find that the *prevent* action is concerned with *the government preventing individuals* for conspiracy media, rather than *the vaccination preventing the virus*, as is the case in mainstream media. This shows that the approach is very suitable for an exploratory framing analysis.

In our recently concluded experiments [35, refer to Figure 1a for an example], we provide our contribution to the SemEval challenge 2023 Task 3 using a SetFit-inspired approach for few-shot predictions [44]. The challenge provides predefined frames on which the performance is measured, but only provides a low amount of labels. Unlike the more exploratory approaches, we deem a classical label prediction approach more applicable for a competitive scenario with predefined labels, i.e., conceptualizations. Hence, our approach adopts contrastive multi-label loss functions for fine-tuning a multilingual base encoder. We achieved the first position on the zero-shot Spanish framing detection subtask.

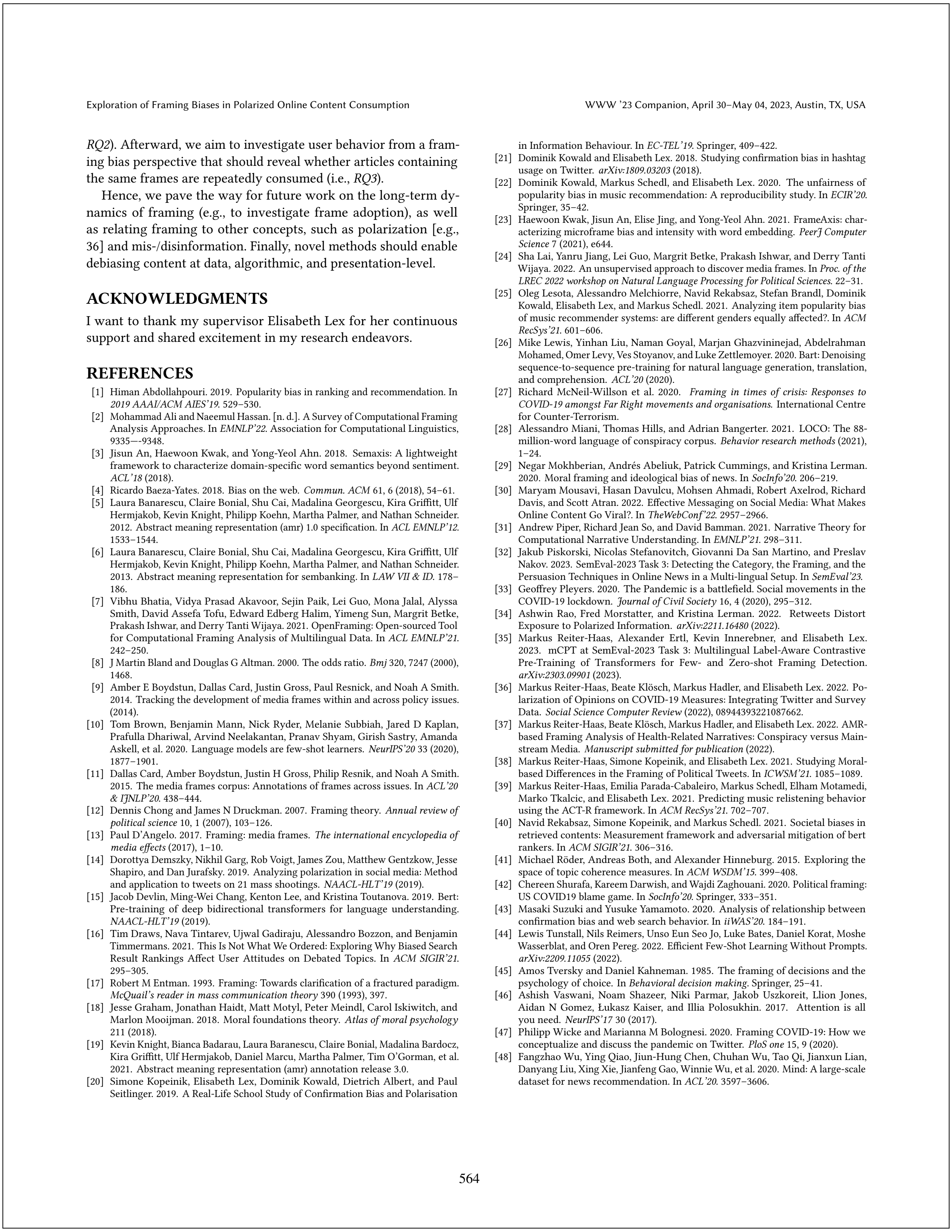
We summarize our findings regarding the three approaches in Table 1. All three approaches have their merits, but the graph-based approach is best in terms of exploratory potential. Also, all three approaches can be employed in an unsupervised manner, but the classifier can only do so in a zero-shot setting that harms its predictive performance. The graph-based approach is the only one that naively allows the extraction of narratives rather than simpler conceptions. This can be attributed to the irregularity of graphs in comparison to the simpler hyper-dimensional structure of embedding spaces and scalar-valued label probabilities. Regarding the data types, the classifier can be used both for discrete label and continuous label probability predictions. In the embedding space, continuous values are the norm to specify the positions, whereas graph elements and sub-graphs are frequency-based. This also affects the complexity of the aggregation in a corpus, where labels (or their probabilities) are trivial to combine. Embedding spaces, while more complex, are still intuitive to aggregate (e.g., mean embedding). However, the aggregation of sub-graphs or their elements is challenging. Altogether, we highlight that the three approaches are complementary in nature, and considering the information from all three approaches is beneficial.

7 CONCLUSION

In this work, we introduced our efforts towards an exploratory approach for framing analysis, which is a multi-faceted problem. We demonstrated in previous works that the semantic information extracted from pretrained Transformers provides richer representations for comparison between different corpora. There, we also showed that such approaches tie in neatly with the current state-of-the-art, and hence, allow for a more comprehensive analysis.

As currently ongoing research, we aim to consolidate the previously distinct directions of research into a holistic approach and openly available framework for framing analysis (i.e., to conclude *RQ1*). Furthermore, we started working on the categorization of health-related frames using the knowledge of the prior research (i.e.,

[†] A classifier can be used in an unsupervised fashion with zero-shot learning.



RQ2). Afterward, we aim to investigate user behavior from a framing bias perspective that should reveal whether articles containing the same frames are repeatedly consumed (i.e., *RQ3*).

Hence, we pave the way for future work on the long-term dynamics of framing (e.g., to investigate frame adoption), as well as relating framing to other concepts, such as polarization [e.g., 36] and mis-/disinformation. Finally, novel methods should enable debiasing content at data, algorithmic, and presentation-level.

ACKNOWLEDGMENTS

I want to thank my supervisor Elisabeth Lex for her continuous support and shared excitement in my research endeavors.

REFERENCES

[1] Himan Abdollahpouri. 2019. Popularity bias in ranking and recommendation. In *2019 AAAI/ACM AIES'19*. 529–530.

[2] Mohammad Ali and Naeemul Hassan. [n. d.]. A Survey of Computational Framing Analysis Approaches. In *EMNLP'22*. Association for Computational Linguistics, 9335—-9348.

[3] Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. 2018. Semaxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment. *ACL'18* (2018).

[4] Ricardo Baeza-Yates. 2018. Bias on the web. *Commun. ACM* 61, 6 (2018), 54–61.

[5] Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2012. Abstract meaning representation (amr) 1.0 specification. In *ACL EMNLP'12*. 1533–1544.

[6] Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *LAW VII & ID*. 178–186.

[7] Vibhu Bhatia, Vidya Prasad Akavoor, Sejin Paik, Lei Guo, Mona Jalal, Alyssa Smith, David Assefa Tofu, Edward Edberg Halim, Yimeng Sun, Margrit Betke, Prakash Ishwar, and Derry Tanti Wijaya. 2021. OpenFraming: Open-sourced Tool for Computational Framing Analysis of Multilingual Data. In *ACL EMNLP'21*. 242–250.

[8] J Martin Bland and Douglas G Altman. 2000. The odds ratio. *Bmj* 320, 7247 (2000), 1468.

[9] Amber E Boydstun, Dallas Card, Justin Gross, Paul Resnick, and Noah A Smith. 2014. Tracking the development of media frames within and across policy issues. (2014).

[10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *NeurIPS'20* 33 (2020), 1877–1901.

[11] Dallas Card, Amber Boydstun, Justin H Gross, Philip Resnik, and Noah A Smith. 2015. The media frames corpus: Annotations of frames across issues. In *ACL'20 & IJNLP'20*. 438–444.

[12] Dennis Chong and James N Druckman. 2007. Framing theory. *Annual review of political science* 10, 1 (2007), 103–126.

[13] Paul D'Angelo. 2017. Framing: media frames. *The international encyclopedia of media effects* (2017), 1–10.

[14] Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Matthew Gentzkow, Jesse Shapiro, and Dan Jurafsky. 2019. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. *NAACL-HLT'19* (2019).

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT'19* (2019).

[16] Tim Draws, Nava Tintarev, Ujwal Gadiraju, Alessandro Bozzon, and Benjamin Timmermans. 2021. This Is Not What We Ordered: Exploring Why Biased Search Result Rankings Affect User Attitudes on Debated Topics. In *ACM SIGIR'21*. 295–305.

[17] Robert M Entman. 1993. Framing: Towards clarification of a fractured paradigm. *McQuail's reader in mass communication theory* 390 (1993), 397.

[18] Jesse Graham, Jonathan Haidt, Matt Motyl, Peter Meindl, Carol Iskiwitch, and Marlon Mooijman. 2018. Moral foundations theory. *Atlas of moral psychology* 211 (2018).

[19] Kevin Knight, Bianca Badarau, Laura Baranescu, Claire Bonial, Madalina Bardocz, Kira Griffitt, Ulf Hermjakob, Daniel Marcu, Martha Palmer, Tim O'Gorman, et al. 2021. Abstract meaning representation (amr) annotation release 3.0.

[20] Simone Kopeinik, Elisabeth Lex, Dominik Kowald, Dietrich Albert, and Paul Seitlinger. 2019. A Real-Life School Study of Confirmation Bias and Polarisation in Information Behaviour. In *EC-TEL'19*. Springer, 409–422.

[21] Dominik Kowald and Elisabeth Lex. 2018. Studying confirmation bias in hashtag usage on Twitter. *arXiv:1809.03203* (2018).

[22] Dominik Kowald, Markus Schedl, and Elisabeth Lex. 2020. The unfairness of popularity bias in music recommendation: A reproducibility study. In *ECIR'20*. Springer, 35–42.

[23] Haewoon Kwak, Jisun An, Elise Jing, and Yong-Yeol Ahn. 2021. FrameAxis: characterizing microframe bias and intensity with word embedding. *PeerJ Computer Science* 7 (2021), e644.

[24] Sha Lai, Yanru Jiang, Lei Guo, Margrit Betke, Prakash Ishwar, and Derry Tanti Wijaya. 2022. An unsupervised approach to discover media frames. In *Proc. of the LREC 2022 workshop on Natural Language Processing for Political Sciences*. 22–31.

[25] Oleg Lesota, Alessandro Melchiorre, Navid Rekabsaz, Stefan Brandl, Dominik Kowald, Elisabeth Lex, and Markus Schedl. 2021. Analyzing item popularity bias of music recommender systems: are different genders equally affected?. In *ACM RecSys'21*. 601–606.

[26] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *ACL'20* (2020).

[27] Richard McNeil-Willson et al. 2020. *Framing in times of crisis: Responses to COVID-19 amongst Far Right movements and organisations*. International Centre for Counter-Terrorism.

[28] Alessandro Miani, Thomas Hills, and Adrian Bangerter. 2021. LOCO: The 88-million-word language of conspiracy corpus. *Behavior research methods* (2021), 1–24.

[29] Negar Mokhberian, Andrés Abeliuk, Patrick Cummings, and Kristina Lerman. 2020. Moral framing and ideological bias of news. In *SocInfo'20*. 206–219.

[30] Maryam Mousavi, Hasan Davulcu, Mohsen Ahmadi, Robert Axelrod, Richard Davis, and Scott Atran. 2022. Effective Messaging on Social Media: What Makes Online Content Go Viral?. In *TheWebConf'22*. 2957–2966.

[31] Andrew Piper, Richard Jean So, and David Bamman. 2021. Narrative Theory for Computational Narrative Understanding. In *EMNLP'21*. 298–311.

[32] Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. SemEval-2023 Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup. In *SemEval'23*.

[33] Geoffrey Pleyers. 2020. The Pandemic is a battlefield. Social movements in the COVID-19 lockdown. *Journal of Civil Society* 16, 4 (2020), 295–312.

[34] Ashwin Rao, Fred Morstatter, and Kristina Lerman. 2022. Retweets Distort Exposure to Polarized Information. *arXiv:2211.16480* (2022).

[35] Markus Reiter-Haas, Alexander Ertl, Kevin Innerebner, and Elisabeth Lex. 2023. mCPT at SemEval-2023 Task 3: Multilingual Label-Aware Contrastive Pre-Training of Transformers for Few- and Zero-shot Framing Detection. *arXiv:2303.09901* (2023).

[36] Markus Reiter-Haas, Beate Klösch, Markus Hadler, and Elisabeth Lex. 2022. Polarization of Opinions on COVID-19 Measures: Integrating Twitter and Survey Data. *Social Science Computer Review* (2022), 08944393221087662.

[37] Markus Reiter-Haas, Beate Klösch, Markus Hadler, and Elisabeth Lex. 2022. AMR-based Framing Analysis of Health-Related Narratives: Conspiracy versus Mainstream Media. *Manuscript submitted for publication* (2022).

[38] Markus Reiter-Haas, Simone Kopeinik, and Elisabeth Lex. 2021. Studying Moral-based Differences in the Framing of Political Tweets. In *ICWSM'21*. 1085–1089.

[39] Markus Reiter-Haas, Emilia Parada-Cabaleiro, Markus Schedl, Elham Motamedi, Marko Tkalcic, and Elisabeth Lex. 2021. Predicting music relistening behavior using the ACT-R framework. In *ACM RecSys'21*. 702–707.

[40] Navid Rekabsaz, Simone Kopeinik, and Markus Schedl. 2021. Societal biases in retrieved contents: Measurement framework and adversarial mitigation of bert rankers. In *ACM SIGIR'21*. 306–316.

[41] Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *ACM WSDM'15*. 399–408.

[42] Chereen Shurafa, Kareem Darwish, and Wajdi Zaghouani. 2020. Political framing: US COVID19 blame game. In *SocInfo'20*. Springer, 333–351.

[43] Masaki Suzuki and Yusuke Yamamoto. 2020. Analysis of relationship between confirmation bias and web search behavior. In *iiWAS'20*. 184–191.

[44] Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient Few-Shot Learning Without Prompts. *arXiv:2209.11055* (2022).

[45] Amos Tversky and Daniel Kahneman. 1985. The framing of decisions and the psychology of choice. In *Behavioral decision making*. Springer, 25–41.

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *NeurIPS'17* 30 (2017).

[47] Philipp Wicke and Marianna M Bolognesi. 2020. Framing COVID-19: How we conceptualize and discuss the pandemic on Twitter. *PloS one* 15, 9 (2020).

[48] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, et al. 2020. Mind: A large-scale dataset for news recommendation. In *ACL'20*. 3597–3606.

mCPT at SemEval-2023 Task 3: Multilingual Label-Aware Contrastive Pre-Training of Transformers for Few- and Zero-shot Framing Detection

Markus Reiter-Haas[*] [†], Alexander Ertl[*], Kevin Innerhofer, Elisabeth Lex
Graz University of Technology, Institute of Interactive Systems and Data Science
Sandgasse 36/III, 8010, Graz, Austria
`reiter-haas@tugraz.at, ertl@student.tugraz.at`
`innerebner@student.tugraz.at, elisabeth.lex@tugraz.at`

Abstract

This paper presents the winning system for the zero-shot Spanish framing detection task, which also achieves competitive places in eight additional languages. The challenge of the framing detection task lies in identifying a set of 14 frames when only a few or zero samples are available, i.e., a multilingual multi-label few- or zero-shot setting. Our developed solution employs a pre-training procedure based on multilingual Transformers using a label-aware contrastive loss function. In addition to describing the system, we perform an embedding space analysis and ablation study to demonstrate how our pre-training procedure supports framing detection to advance computational framing analysis.

1 Introduction

Approaches for computational framing detection are diverse (Ali and Hassan, 2022), as the framing concept itself is often just casually defined (Entman, 1993). Consequently, framing detection is challenging on its own, but also suffers from a lack of sufficient data (Kwak et al., 2020), especially in multilingual settings. The SemEval 2023 Task 3 Subtask 2 (Piskorski et al., 2023) aims at predicting 14 distinct media frames (Boydstun et al., 2013) present within news articles in 9 languages. Due to label imbalances, as a result of the high dimension of the label space compared to the number of samples, traditional paradigms, e.g., per-label binary classification, do not apply well to the given setting without adaptions (Tarekegn et al., 2021).

We introduce *mCPT*, the label-aware Contrastive Pre-training of Transformers based on a multilingual encoder model (original team name on the leaderboard[1]: *PolarIce*). We exploit two features of the task: (i) multi-label information and (ii) multilingual data for pre-training.

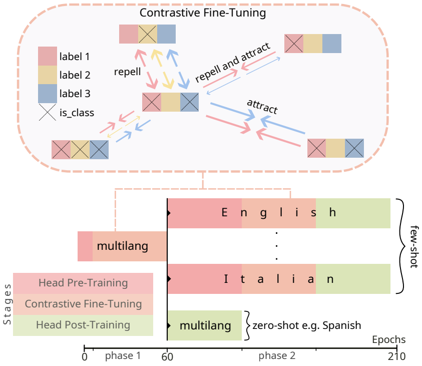


Figure 1: Our system performs label-aware contrastive fine-tuning (top). Embeddings of samples with similar labels are attracted, while they are repelled for dissimilar labels. The two-phase procedure (bottom) interleaves contrastive fine-tuning in both the multilingual and target language training.

First, we leverage the label information by adopting a contrastive loss function, i.e., HeroConLoss (Zheng et al., 2022), for natural language processing that optimizes the embedding space with respect to the similarities of the label space. Therefore, samples with more similar labels occupy similar regions in the embedding space, whereas mostly dissimilar samples regarding their shared labels are pushed apart (refer to Figures 1 top and 2a).

Second, we design a custom two-phase procedure with multiple stages for multilingual training to maximize the available data (see Figure 1). In phase one, we train on all languages, while in phase two, we further fine-tune the model on the target language if such data exist i.e. few-shot setting, or continue training on all languages if not i.e. zero-shot setting.

Our system performs competitively (top 10) on

[*] equal contribution
[†] corresponding author
[1] `https://propaganda.math.unipd.it/semeval2023task3/SemEval2023testleaderboard.html`

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 941–949
July 13-14, 2023 ©2023 Association for Computational Linguistics

all six few-shot (i.e., English, German, French, Italian, Polish, and Russian) and three zero-shot (i.e., Spanish, Greek, and Georgian) settings, beating the baselines on all languages. On Spanish, which is the only zero-shot language with a common language family and alphabet as the training languages, our system is the winning contribution with a Micro-F1 of 0.571 (compared to 0.120 of the baseline). Therefore, we argue that our system generalizes well to unseen data, even when no training data is available in similar target languages.

In sum, our contribution is three-fold[2]:

C1 We adopt a multi-label contrastive loss function for natural language processing to optimize the embeddings of textual data.

C2 We describe a two-phase multi-stage training procedure for multilingual scenarios with limited data, i.e., few- and zero-shot predictions.

C3 We demonstrate the effectiveness of our winning system for framing detection supported by embedding and ablation studies.

2 Related Work

Framing Detection. According to Entman (1993), to frame is to select and emphasize some aspects of reality to encourage particular interpretations. That is, messages centered around a common topic may draw the receiver's attention to distinct features, thus suggesting different courses of action, causal interpretations, etc. As such, computational framing detection requires natural language processing (NLP) methods that capture nuances of *how* content is presented rather than just *what* topic is present. Therein, studies focus on detecting vastly different conceptualizations of framing, such as blame frames (Shurafa et al., 2020), war frames (Wicke and Bolognesi, 2020), moral frames (Reiter-Haas et al., 2021), or media frames (Boydstun et al., 2014; Kwak et al., 2020).

Regarding media frames, Boydstun et al. (2013) identified a set of relevant frames that formed the basis for the media frame corpus (Card et al., 2015). Within this supervised frame detection scenario, Liu et al. (2019) indicate that Transformer-based approaches vastly outperform approaches using less powerful architectures such as LSTMs. As such, we also employ Transformer models with label-aware contrastive pre-training.

Supervised Contrastive Learning. Contrastive learning, originally mainly used in computer vision settings (e.g., Chopra et al., 2005), has recently found increased attention in the NLP research community due to its efficacy on tasks with limited amounts of data and its applicability to Transformer embeddings (e.g., Tunstall et al., 2022). The general concept of supervised contrastive learning (Khosla et al., 2020a) is that latent representations (or embeddings in NLP) of samples with the same labels should be close in embedding space, while samples with different labels should be further apart.

Su et al. (2022) and Zheng et al. (2022) have independently proposed contrastive learning methods for multi-label settings that weight similarities of samples by the similarity of their label vectors, i.e. hidden representations of samples with similar label vectors should be more similar than hidden representations of samples with less similar label vectors. Su et al. (2022) weight a Euclidean distance-based measure of embeddings by a normalized dot product of the label vectors. HeroCon (Zheng et al., 2022) generalizes supervised contrastive loss (Khosla et al., 2020b) and beats previous state-of-the-art contrastive learning paradigms in multi-label settings on multiple image data sets. Tunstall et al. (2022) introduce SETFIT, an algorithm for the data-efficient fine-tuning of sentence embeddings, primarily on binary labels. SETFIT first fine-tunes sentence embeddings in a contrastive manner before training a classification head.

We combine the idea of the contrastive pre-training stage from SETFIT and adopt HeroCon for NLP loss to improve performance on multi-label datasets.

3 Methods

At the core of our system lies a multilingual Transformer model with dense neural layers comprising the head. Contrastive fine-tuning is performed as part of a multi-stage training procedure.

3.1 Contrastive Fine-Tuning

Our contrastive fine-tuning objective (*C1*, Figure 1 top) is centered around the idea that embeddings of samples with similar labels should be close while embeddings of samples with very distinct label vectors should be distant. Following Zheng et al. (2022) for every batch and every class, we com-

[2] Our code and model are publicly available at: `https://github.com/socialcomplab/semeval23-mcpt`

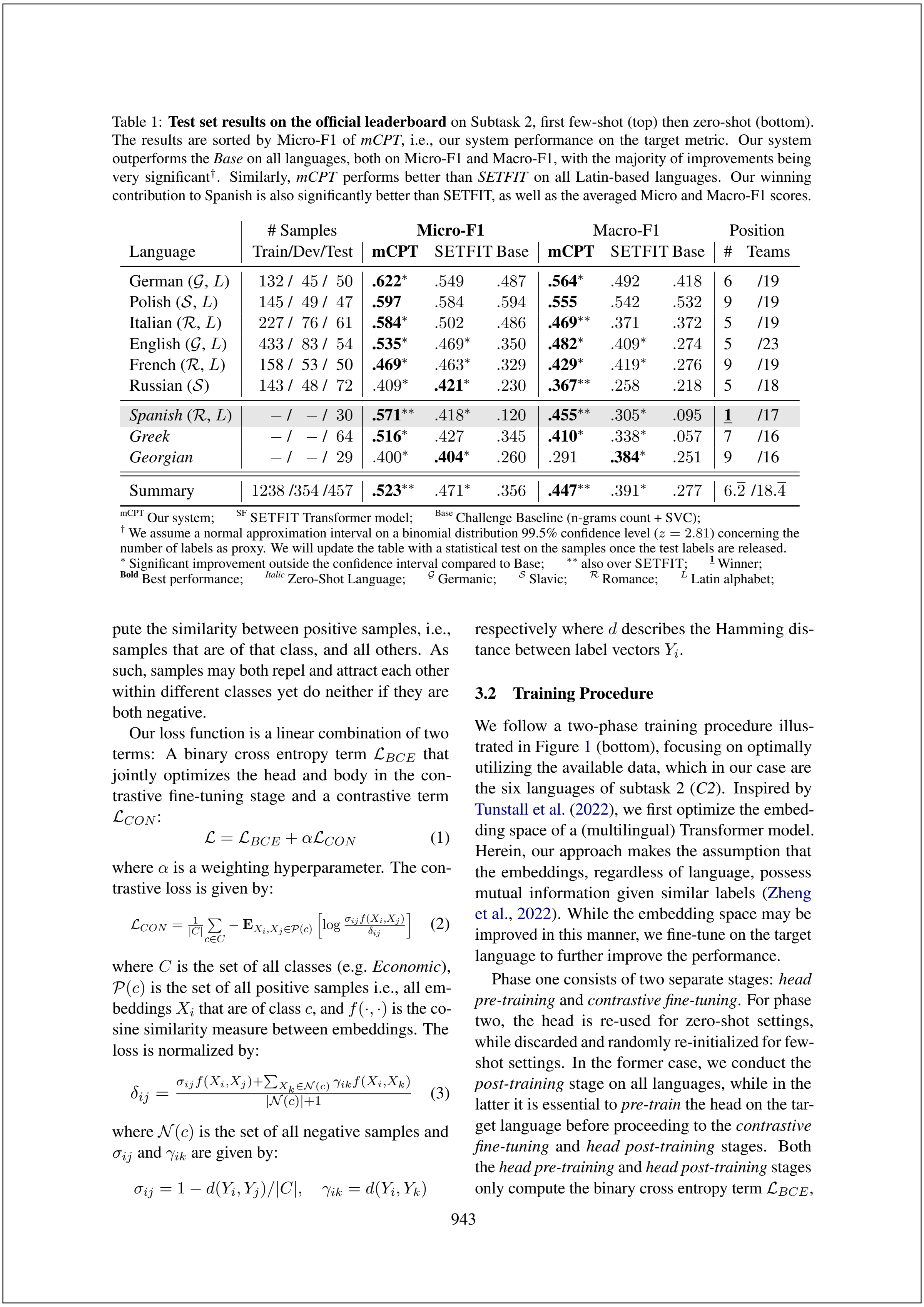


Table 1: **Test set results on the official leaderboard** on Subtask 2, first few-shot (top) then zero-shot (bottom). The results are sorted by Micro-F1 of *mCPT*, i.e., our system performance on the target metric. Our system outperforms the *Base* on all languages, both on Micro-F1 and Macro-F1, with the majority of improvements being very significant†. Similarly, *mCPT* performs better than *SETFIT* on all Latin-based languages. Our winning contribution to Spanish is also significantly better than SETFIT, as well as the averaged Micro and Macro-F1 scores.

	# Samples	**Micro-F1**			Macro-F1			Position	
Language	Train/Dev/Test	**mCPT**	SETFIT	Base	**mCPT**	SETFIT	Base	#	Teams
German ($\mathcal{G}$, L)	132 / 45 / 50	**.622***	.549	.487	**.564***	.492	.418	6	/19
Polish ($\mathcal{S}$, L)	145 / 49 / 47	**.597**	.584	.594	**.555**	.542	.532	9	/19
Italian ($\mathcal{R}$, L)	227 / 76 / 61	**.584***	.502	.486	**.469****	.371	.372	5	/19
English ($\mathcal{G}$, L)	433 / 83 / 54	**.535***	.469*	.350	**.482***	.409*	.274	5	/23
French ($\mathcal{R}$, L)	158 / 53 / 50	**.469***	.463*	.329	**.429***	.419*	.276	9	/19
Russian ($\mathcal{S}$)	143 / 48 / 72	.409*	**.421***	.230	**.367****	.258	.218	5	/18
Spanish ($\mathcal{R}$, L)	– / – / 30	**.571****	.418*	.120	**.455****	.305*	.095	<u>**1**</u>	/17
Greek	– / – / 64	**.516***	.427	.345	**.410***	.338*	.057	7	/16
Georgian	– / – / 29	.400*	**.404***	.260	.291	**.384***	.251	9	/16
Summary	1238 /354 /457	**.523****	.471*	.356	**.447****	.391*	.277	$6.\overline{2}$	$/18.\overline{4}$

mCPT Our system; SF SETFIT Transformer model; Base Challenge Baseline (n-grams count + SVC);
† We assume a normal approximation interval on a binomial distribution 99.5% confidence level ($z = 2.81$) concerning the number of labels as proxy. We will update the table with a statistical test on the samples once the test labels are released.
* Significant improvement outside the confidence interval compared to Base; ** also over SETFIT; <u>**1**</u> Winner;
Bold Best performance; *Italic* Zero-Shot Language; $^{\mathcal{G}}$ Germanic; $^{\mathcal{S}}$ Slavic; $^{\mathcal{R}}$ Romance; ^{L} Latin alphabet;

pute the similarity between positive samples, i.e., samples that are of that class, and all others. As such, samples may both repel and attract each other within different classes yet do neither if they are both negative.

Our loss function is a linear combination of two terms: A binary cross entropy term $\mathcal{L}_{BCE}$ that jointly optimizes the head and body in the contrastive fine-tuning stage and a contrastive term $\mathcal{L}_{CON}$:

$$\mathcal{L} = \mathcal{L}_{BCE} + \alpha\mathcal{L}_{CON} \qquad (1)$$

where α is a weighting hyperparameter. The contrastive loss is given by:

$$\mathcal{L}_{CON} = \frac{1}{|C|}\sum_{c\in C} -\mathbf{E}_{X_i,X_j\in\mathcal{P}(c)}\left[\log\frac{\sigma_{ij}f(X_i,X_j)}{\delta_{ij}}\right] \qquad (2)$$

where C is the set of all classes (e.g. *Economic*), $\mathcal{P}(c)$ is the set of all positive samples i.e., all embeddings X_i that are of class c, and $f(\cdot,\cdot)$ is the cosine similarity measure between embeddings. The loss is normalized by:

$$\delta_{ij} = \frac{\sigma_{ij}f(X_i,X_j)+\sum_{X_k\in\mathcal{N}(c)}\gamma_{ik}f(X_i,X_k)}{|\mathcal{N}(c)|+1} \qquad (3)$$

where $\mathcal{N}(c)$ is the set of all negative samples and σ_{ij} and γ_{ik} are given by:

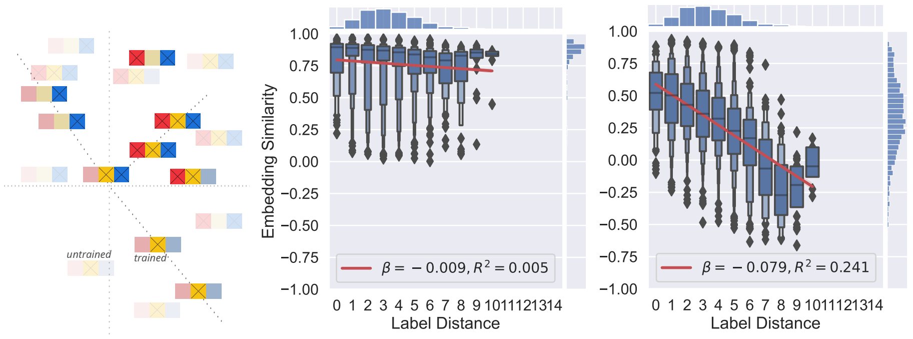
$$\sigma_{ij} = 1 - d(Y_i,Y_j)/|C|, \quad \gamma_{ik} = d(Y_i,Y_k)$$

respectively where d describes the Hamming distance between label vectors Y_i.

3.2 Training Procedure

We follow a two-phase training procedure illustrated in Figure 1 (bottom), focusing on optimally utilizing the available data, which in our case are the six languages of subtask 2 (*C2*). Inspired by Tunstall et al. (2022), we first optimize the embedding space of a (multilingual) Transformer model. Herein, our approach makes the assumption that the embeddings, regardless of language, possess mutual information given similar labels (Zheng et al., 2022). While the embedding space may be improved in this manner, we fine-tune on the target language to further improve the performance.

Phase one consists of two separate stages: *head pre-training* and *contrastive fine-tuning*. For phase two, the head is re-used for zero-shot settings, while discarded and randomly re-initialized for few-shot settings. In the former case, we conduct the *post-training* stage on all languages, while in the latter it is essential to *pre-train* the head on the target language before proceeding to the *contrastive fine-tuning* and *head post-training* stages. Both the *head pre-training* and *head post-training* stages only compute the binary cross entropy term $\mathcal{L}_{BCE}$,



(a) Repositioning due to loss (b) Before contrastive training (c) After contrastive training

Figure 2: *Effect of the loss function on the embedding space.* (a) Shows the repositioning of randomly generated samples (both embeddings and labels) in two-dimensional space. The contrast loss function on its own increases the cosine similarity of latent representations with similar labels, while decreasing the similarity of representations with different labels. Note the positioning of trained embeddings with identical labels along lines drawn from the origin. (b) Without contrastive pre-training, the pairs of embeddings in the English dev set are similar regardless of their label distance. (c) After 50 epochs, the embedding cosine similarity reflects the Hamming distance of the labels.

while simultaneously leaving the body unchanged, i.e., frozen. While the stages are identical, the rationale for each of them is very distinct: As the head is randomly initialized, we first *pre-train* it to avoid high gradients in the subsequent *contrastive fine-tuning* stage. In contrast, *post-training* allows the head to better fit the fine-tuned embeddings.

4 Experiments

We now present the results of our *mCPT* system, supported by embedding space and ablation studies (*C3*). We base *mCPT* on the multilingual[3] sentence Transformer model *paraphrase-multilingual-MiniLM-L12-v2* (Reimers and Gurevych, 2019) and demonstrate that competitive results can be achieved with a relatively small amount of parameters, i.e., $117M$ parameters (Wang et al., 2020), given a training method tailored to the task. The model was chosen for its sentence embedding performance on multiple languages and its small size relative to similar state-of-the-art multilingual Transformer models.

Our model architecture comprises mean-pooling, no normalization of embeddings, a dense head with one hidden layer of size 256, and a dropout of 0.5. We train the model with separate learning rates for the classification head ($1e{-}3$) and the body ($2e{-}5$), a weighting parameter α of 0.01, a batch size of 26 for 10 and 50 epochs for *head pre-training* and *contrastive fine-tuning* respectively in phase one (more details in Appendix A).

Baseline Models. We compare the performance of our system against two baselines. First, we consider the results of the official baseline *Base* (i.e., n-grams and support vector classification; Piskorski et al., 2023). Second, we compare against SETFIT (Tunstall et al., 2022) with the same base encoder as ours, on the post-challenge test set (details in Appendix B).

4.1 Main Results

mCPT performs better on Latin alphabets (marked by L) in both few- and zero-shot settings, and improves upon the two baselines (as presented in Table 1). In Slavic languages ($\mathcal{S}$), Polish is second-best in terms of Micro-F1 (0.597), but only slightly outperforms the baseline, whereas the improvement on Russian is very significant but only achieves the second-lowest score of 0.409. In comparison, both Germanic languages ($\mathcal{G}$), German and English perform well, where we also have our highest overall Micro-F1 of 0.622 for German, but also a high baseline of 0.487. Although we find significant improvement on Greek and Georgian (which are zero-shot languages that do not share a major branch with any other language) over the baseline, both perform poorly in terms of Micro-F1 (i.e., Georgian having the lowest Micro-F1 of

[3] The base model was trained on 50+ languages including all nine of the shared task, thus being suitable for the problem.

0.400). Hence, we suspect that not enough information from the other languages could be transferred. Finally, in the Romance languages ($\mathcal{R}$), Italian and Spanish perform well, while French with a Micro-F1 of 0.516 performs lower than Greek.

The performance of Spanish is especially noteworthy, as it is the only zero-shot language that shares an alphabet as well as a language family with the training data languages. Therefore, we argue that our winning performance on Spanish (Micro-F1 of 0.571 compared to 0.12 of the baseline) stems from the fact that the knowledge was successfully transferred from the other languages to the zero-shot setting. This is further supported by *SETFIT*, which improves upon the baseline but shows lower performance on Latin-based languages.

4.2 Embedding Space Analysis

Figure 2 demonstrates how our contrastive training procedure optimizes the embeddings of the Transformer body. Figure 2a exemplarily shows the repositioning of samples due to the loss function in two-dimensional space. Observe how the trained labels (opaque) align, with the blue-yellow label between the two blue-only labels and two yellow-only labels, while simultaneously pushing the red labels to the side (more detailed analysis in Appendix C). For the analysis of the high-dimensional embeddings (384) and label spaces (14) on the dev set, we use boxen plots concerning embedding cosine similarity for all pairwise samples within a given Hamming distance.

Regarding the pre-trained base model on English (Figure 2b), we find a suboptimal correlation with $R^2 = 0.005$ and $\beta = -0.009$. In comparison, after contrastively training the model, the correlation becomes much more pronounced, i.e., $R^2 = 0.241$ and $\beta = -0.079$ for English as shown in Figure 2c. Furthermore, the spread of pairwise embedding similarity distribution increases, as a greater amount of samples become dissimilar to each other, especially for higher label distance. Thus, we conclude that our system leads to higher utilization of the available embedding space, which in turn boosts performance. Appendix D contains the remaining languages.

Finally, we want to emphasize that our data set has no perfectly dissimilar label vector pairs, which would make hard negative mining approaches (Gao et al., 2021) infeasible, e.g., for using plain contrastive loss (Chopra et al., 2005).

Table 2: Ablation study (top) and the proposed contrastive sampling extension (bottom) on the dev set. In general, we observe that *mCPT* performs best with all components, i.e., pre-training (PT), contrastive loss ($\mathcal{L}_{CON}$), and end-to-end training (E2E), enabled. Contrast sampling (CS) suggest further improvements.

Model	en	it	ru	fr	ge	po
mCPT	**.682**	**.585**	**.520**	**.570**	.561	.636
- PT	.681	.545	.475	.563	.583	.616
- $\mathcal{L}_{CON}$	.657	.521	.436	.524	.570	**.645**
- E2E	.629	.519	.500	.535	**.586**	.633
mCPT+CS	**.688**	**.590**	.519	**.575**	**.591**	.638

4.3 Ablation Study and Extension

Table 2 indicates the effectiveness of our combined training approach *mCPT*. From *mCPT*, we remove components iteratively, first removing the multilingual pre-training phase, then the contrastive term (see Equation 2), and finally, end-to-end training leaving only a trained classification head with no embedding fine-tuning. Comparing the results of the ablation study with those of Table 1 it is interesting to note that our approach works best on languages with lower scores. A hypothesis is that the out-of-the-box Transformer embeddings already fit the data well and that *mCPT* is not able to improve upon the already strong baseline. Finally, we find that adding a contrast sampling extension could further improve the results (see Appendix E).

5 Conclusion

In this paper, we describe our system (*mCPT*) for the framing detection shared task (Piskorski et al., 2023). We introduce an approach based on a label-aware contrastive loss and training procedure for Transformers to deal with the challenges of multilingual multi-label prediction with few or even zero samples. The generalization ability of our system is demonstrated by providing the winning contribution for the Spanish framing detection subtask where no training samples were available[4]. Hence, we believe that our system is a notable advancement for computational framing research.

[4] We refer to Appendices F and G for discussions on limitations and ethical considerations, respectively.

Acknowledgements

We would like to thank the anonymous reviewers for taking the necessary time and effort to review the manuscript and appreciate their universal positive feedback.

References

Mohammad Ali and Naeemul Hassan. 2022. A survey of computational framing analysis approaches. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9335–9348, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Amber E Boydstun, Dallas Card, Justin Gross, Paul Resnick, and Noah A Smith. 2014. Tracking the development of media frames within and across policy issues.

Amber E Boydstun, Justin H Gross, Philip Resnik, and Noah A Smith. 2013. Identifying media frames and frame dynamics within and across policy issues. In *New Directions in Analyzing Text as Data Workshop, London*.

Dallas Card, Amber Boydstun, Justin H Gross, Philip Resnik, and Noah A Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444.

Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE.

Robert M Entman. 1993. Framing: Towards clarification of a fractured paradigm. *McQuail's reader in mass communication theory*, 390:397.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020a. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020b. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc.

Haewoon Kwak, Jisun An, and Yong-Yeol Ahn. 2020. A systematic media frame analysis of 1.5 million new york times articles from 2000 to 2017. In *12th ACM Conference on Web Science*, pages 305–314.

Qisheng Liao, Meiting Lai, and Preslav Nakov. 2023. Marseclipse at semeval-2023 task 3: Multi-lingual and multi-label framing detection with contrastive learning. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, SemEval 2023, Toronto, Canada.

Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019. Detecting frames in news headlines and its application to analyzing news framing trends surrounding us gun violence. In *Proceedings of the 23rd conference on computational natural language learning (CoNLL)*, pages 504–514.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. Semeval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, SemEval 2023, Toronto, Canada.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Markus Reiter-Haas, Simone Kopeinik, and Elisabeth Lex. 2021. Studying moral-based differences in the framing of political tweets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 1085–1089.

Chereen Shurafa, Kareem Darwish, and Wajdi Zaghouani. 2020. Political framing: Us covid19 blame game. In *International Conference on Social Informatics*, pages 333–351. Springer.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867.

Xi'ao Su, Ran Wang, and Xinyu Dai. 2022. Contrastive learning-enhanced nearest neighbor mechanism for multi-label text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 672–679, Dublin, Ireland. Association for Computational Linguistics.

Adane Nega Tarekegn, Mario Giacobini, and Krzysztof Michalak. 2021. A review of methods for imbalanced multi-label classification. *Pattern Recognition*, 118:107965.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts. *arXiv preprint arXiv:2209.11055*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.

Philipp Wicke and Marianna M Bolognesi. 2020. Framing covid-19: How we conceptualize and discuss the pandemic on twitter. *PloS one*, 15(9):e0240010.

Ben Wu, Olesya Razuvayevskaya, Freddy Heppell, João A. Leite, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2023. Sheffieldveraai at semeval-2023 task 3: Mono and multilingual approaches for news genre, topic and persuasion technique classification. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, SemEval 2023, Toronto, Canada.

Lecheng Zheng, Jinjun Xiong, Yada Zhu, and Jingrui He. 2022. Contrastive learning with complex heterogeneity.

A Training Environment

We performed the main experiments on the Kaggle platform (`www.kaggle.com`) with the P100 graphics card. We chose a free platform for the computation to demonstrate that our system is tailored towards the task at hand and is accessible for everybody, rather than relying on large amounts of computational resources. We empirically selected the hyperparameters to fit the platform. Herein, we chose a batch size of 26 which optimally utilizes the available GPU memory. The multilingual pre-training takes approximately 1.5 hours, while the language-specific fine-tuning takes 1 hour each.

B SETFIT Parameters

We choose *SETFIT* as it is similar in concept to our system, i.e., contrastive learning for Transformers, but not aligned with the shared task, i.e., does not explicitly consider multi-label problems. Hence, the comparison demonstrates how our system is an improvement over established approaches in this setting and emphasizes that the adaptions of the contrastive loss and training procedure are indeed beneficial. We report the results of *SETFIT* on the post-challenge leaderboard without further adaption after the initial submissions for fair comparisons on the test set.

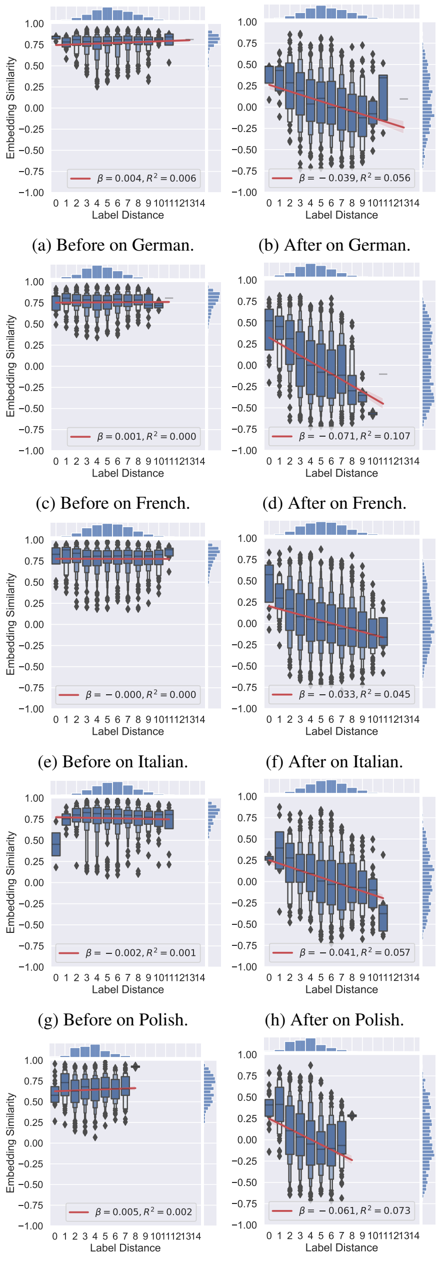
We mimic the parameters setting where applicable while preserving the standard training procedure to maximize comparability. We first contrastively train the body for 10 epochs before training the full model end-to-end for 50 epochs with a batch size of 26 with learning rates of $1e-3$ and $2e-5$ for the head and the body respectively. The body-then-end-to-end procedure was suggested by the usage guide. The training runtime is approximately 10 hours on the Kaggle platform (which again was chosen for a fair comparison). Initially, we experimented with SETFIT in the challenge period but decided to submit our presented system instead.

C Repositioning of Samples

In Figure 2a, we show the effect of the loss function and how the label and embedding space are intertwined. Specifically, we demonstrate how randomly generated embeddings in two-dimensional space with three-dimensional label vectors shift towards more optimal positions after applying the contrastive loss function. Accordingly, the initially random positions of embeddings with equivalent labels end up on straight lines drawn from the origin. The observed effect is a direct consequence of similar label vectors attracting and opposite labels repelling each other. Moreover, a partial label similarity with two distinct groups ends between those groups, as a result of both forces being active. For instance, consider the line from top left to bottom right: blue-only labels become attracted, and repel yellow-only labels, while the sample with blue and yellow lies in between. Due to the resulting positioning, the embedding space becomes disentangled leading to an increase in linear separability, which benefits classifiers such as our differentiable head.

D Embeddings and Labels Correlation

Here, we present the remaining languages for the embedding space analysis in Figure 3. When considering the plots before contrastive training is performed, the correlations between embedding spaces and label spaces are very weak, as well as most embeddings being very similar regardless of their label distance. Hence, virtually no pair of samples are opposite to each other, i.e., has a noteworthy negative cosine similarity. Conversely, all pairs are similar to a certain extent. We argue that in these given embedding spaces, it is challenging



(a) Before on German.

(b) After on German.

(c) Before on French.

(d) After on French.

(e) Before on Italian.

(f) After on Italian.

(g) Before on Polish.

(h) After on Polish.

(i) Before on Russian.

(j) After on Russian.

Figure 3: *Effect of the loss function on the embedding space.* Evaluated on various language dev sets.

for a classifier to learn a discriminative function. Moreover, due to the suboptimal positioning of embeddings, a substantial amount of the embedding space does not contribute to the prediction performance, thus wasting the model's potential expressiveness. Hence, the interpretation from Section 4.2 can be directly applied to the five other languages, as the effect is the same (although differently pronounced). For instance, models of German and Russian even have a slight upward slope when applied without contrastive training. Hence, the negative slope and regression fit increases with contrastive training, as intended and expected.

E Contrast Sampling Extension.

Adding a custom *contrast sampler* which ensures that at least one sample from every class is present per batch further improves consistency as well as performance. Due to the nature of the contrastive objective, it is imperative that every batch contain negative as well as positive pairs of samples for every class. This is not guaranteed by sampling randomly, especially if the label distribution is imbalanced. As illustrated by Table 2 it outperforms pure *mCPT* in five out of six languages while coming in second by a small margin in Russian. We attribute this largely to the variance introduced by using a relatively small batch size of 26 compared to the number of labels (14). This variance may lead to undesirable gradient updates in some iterations when batches contain label distributions that are not representative.

F Limitations

We recognize three main limitations of our work, which are distinct in their aspect.

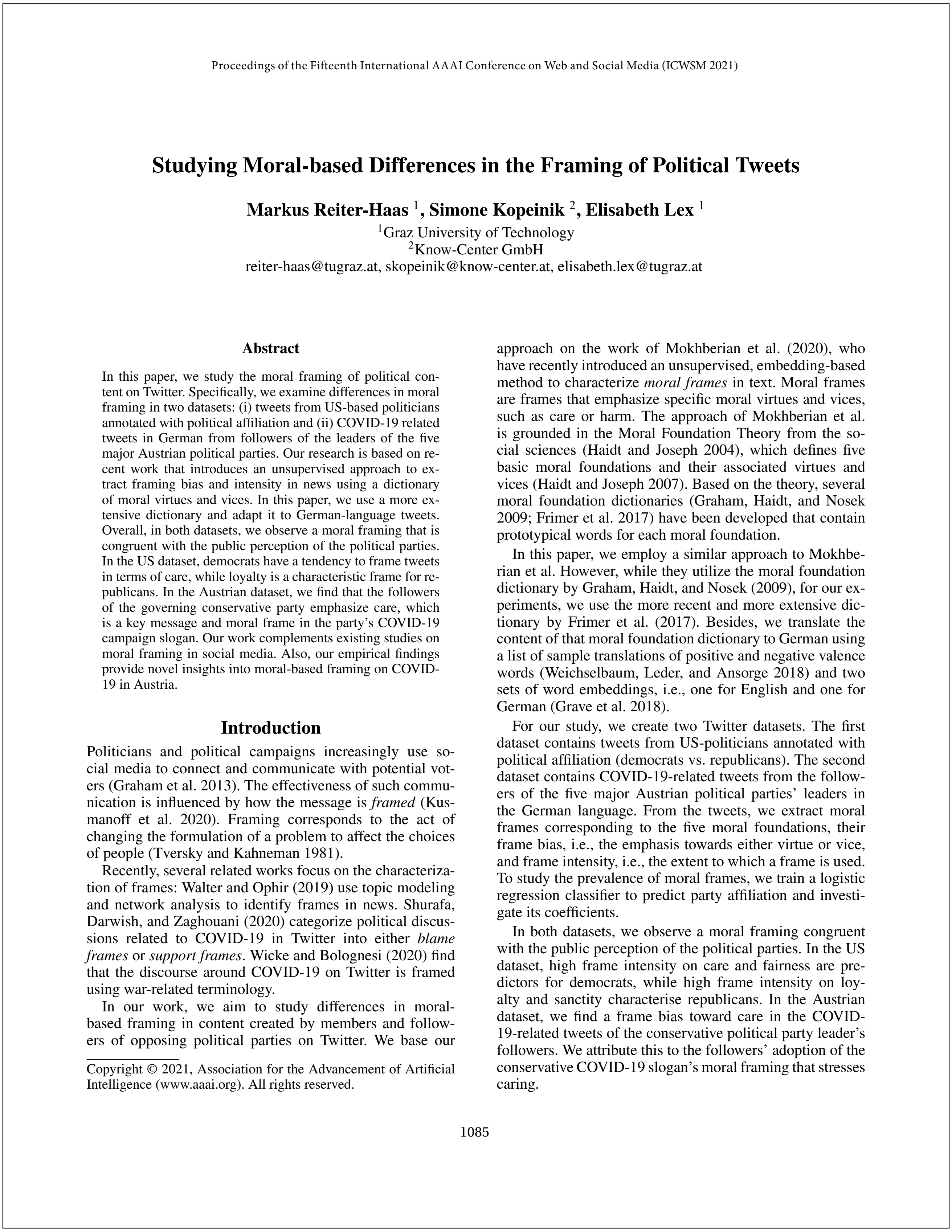
First, the *performance limitation*; while our system has competitive results across the board, it only performs best in one of the nine languages on the leaderboard. In comparison, team *MarsEclipse* (Liao et al., 2023), which also focused on the framing detection subtask, wins all six few-shot languages and performs second on two of three (i.e., Greek and Georgian) zero-shot languages. They only perform worse at Spanish (6th), which is opposite to our placement. Team *SheffieldVeraAI* (Wu et al., 2023), who also participated in the other two subtasks regarding news genre and persuasion technique detection, perform well across the board and wins the Greek and Georgian framing detection tasks. Hence, our system occupies the niche of zero-shot prediction when trained with similar languages (i.e, in our case Latin-based).

Second the *technical limitation*, our system was trained using a small multilingual model as we aimed towards adapting Transformer pre-training for the multi-label challenge in particular rather than achieving the highest performance with computationally expensive training. However, as a consequence, we do not know how well our system scales to bigger models, such as MPNet (Song et al., 2020), and plan to address this limitation in future work.

Third the *task setting limitation*, we want to emphasize a potential limitation resulting from the shared task setting. Ali and Hassan (2022) argue that the specified labels in the media frame corpus (Card et al., 2015) revolve around topics (i.e., the *what*) rather frames (i.e., the *how*). As the same labels were adopted for the shared task, the conceptualizations of frames are expected to be similar to a certain extent. They thus would also affect the resulting models and predictions.

G Ethics Statement

We want to discuss three ethical considerations of our system. First, our system is based on pretrained Transformers, which inherit *biases* from their training data. For the shared task, these biases are negligible, but are a concern for real-world applications. The second consideration relates to *fairness* concerns. The performance varies strongly between languages, with more researched languages typically resulting in better performance. We, thus, embrace the multilingual setting of the shared task with one-third zero-shot languages, but similarly achieved better performance in Latin-based languages. Third, our system leads to better detection of media frames, which is an important research direction. However, the system could in theory also be used in a disputed or even malicious manner, e.g., for *reframing* political statements. Hence, we do not advise specific applications of our system besides better framing detection.



Proceedings of the Fifteenth International AAAI Conference on Web and Social Media (ICWSM 2021)

Studying Moral-based Differences in the Framing of Political Tweets

Markus Reiter-Haas [1], Simone Kopeinik [2], Elisabeth Lex [1]
[1]Graz University of Technology
[2]Know-Center GmbH
reiter-haas@tugraz.at, skopeinik@know-center.at, elisabeth.lex@tugraz.at

Abstract

In this paper, we study the moral framing of political content on Twitter. Specifically, we examine differences in moral framing in two datasets: (i) tweets from US-based politicians annotated with political affiliation and (ii) COVID-19 related tweets in German from followers of the leaders of the five major Austrian political parties. Our research is based on recent work that introduces an unsupervised approach to extract framing bias and intensity in news using a dictionary of moral virtues and vices. In this paper, we use a more extensive dictionary and adapt it to German-language tweets. Overall, in both datasets, we observe a moral framing that is congruent with the public perception of the political parties. In the US dataset, democrats have a tendency to frame tweets in terms of care, while loyalty is a characteristic frame for republicans. In the Austrian dataset, we find that the followers of the governing conservative party emphasize care, which is a key message and moral frame in the party's COVID-19 campaign slogan. Our work complements existing studies on moral framing in social media. Also, our empirical findings provide novel insights into moral-based framing on COVID-19 in Austria.

Introduction

Politicians and political campaigns increasingly use social media to connect and communicate with potential voters (Graham et al. 2013). The effectiveness of such communication is influenced by how the message is *framed* (Kusmanoff et al. 2020). Framing corresponds to the act of changing the formulation of a problem to affect the choices of people (Tversky and Kahneman 1981).

Recently, several related works focus on the characterization of frames: Walter and Ophir (2019) use topic modeling and network analysis to identify frames in news. Shurafa, Darwish, and Zaghouani (2020) categorize political discussions related to COVID-19 in Twitter into either *blame frames* or *support frames*. Wicke and Bolognesi (2020) find that the discourse around COVID-19 on Twitter is framed using war-related terminology.

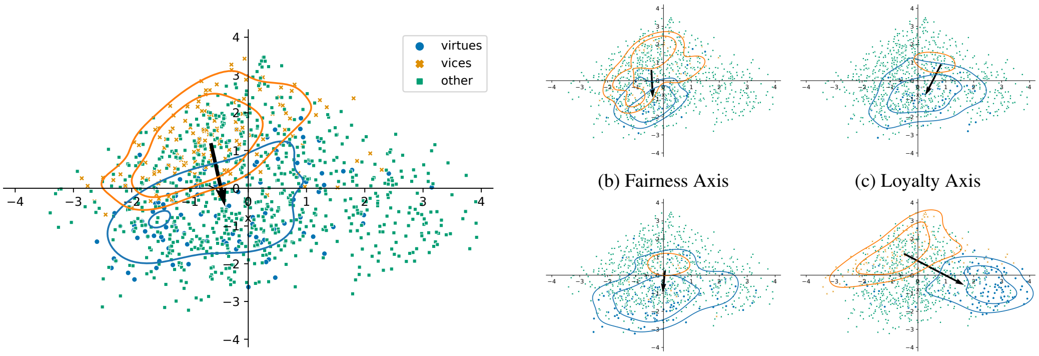
In our work, we aim to study differences in moral-based framing in content created by members and followers of opposing political parties on Twitter. We base our approach on the work of Mokhberian et al. (2020), who have recently introduced an unsupervised, embedding-based method to characterize *moral frames* in text. Moral frames are frames that emphasize specific moral virtues and vices, such as care or harm. The approach of Mokhberian et al. is grounded in the Moral Foundation Theory from the social sciences (Haidt and Joseph 2004), which defines five basic moral foundations and their associated virtues and vices (Haidt and Joseph 2007). Based on the theory, several moral foundation dictionaries (Graham, Haidt, and Nosek 2009; Frimer et al. 2017) have been developed that contain prototypical words for each moral foundation.

In this paper, we employ a similar approach to Mokhberian et al. However, while they utilize the moral foundation dictionary by Graham, Haidt, and Nosek (2009), for our experiments, we use the more recent and more extensive dictionary by Frimer et al. (2017). Besides, we translate the content of that moral foundation dictionary to German using a list of sample translations of positive and negative valence words (Weichselbaum, Leder, and Ansorge 2018) and two sets of word embeddings, i.e., one for English and one for German (Grave et al. 2018).

For our study, we create two Twitter datasets. The first dataset contains tweets from US-politicians annotated with political affiliation (democrats vs. republicans). The second dataset contains COVID-19-related tweets from the followers of the five major Austrian political parties' leaders in the German language. From the tweets, we extract moral frames corresponding to the five moral foundations, their frame bias, i.e., the emphasis towards either virtue or vice, and frame intensity, i.e., the extent to which a frame is used. To study the prevalence of moral frames, we train a logistic regression classifier to predict party affiliation and investigate its coefficients.

In both datasets, we observe a moral framing congruent with the public perception of the political parties. In the US dataset, high frame intensity on care and fairness are predictors for democrats, while high frame intensity on loyalty and sanctity characterise republicans. In the Austrian dataset, we find a frame bias toward care in the COVID-19-related tweets of the conservative political party leader's followers. We attribute this to the followers' adoption of the conservative COVID-19 slogan's moral framing that stresses caring.

Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.



(a) Care Axis

(b) Fairness Axis

(c) Loyalty Axis

(d) Authority Axis

(e) Sanctity Axis

Figure 1: Axis of the five moral foundations. Each axis is created by the centroid of words assigned to virtues and the centroid of words for vices and surrounded by moral words associated with the other axes. The black arrow goes from the vices' centroid to the virtues' centroid and describes the axes. The high-dimensional space is reduced with Principal Component Analysis (PCA). All the axis point approximately in the same direction, which indicates that virtues are more similar to other virtues than to their corresponding vices, and vice versa. A kernel density estimation of the underlying point cloud is used for the colored contours.

Characterization of Moral-based Framing

In the following, we describe our approach to investigate moral-based differences in the framing of tweets.

Capturing Moral Frames

In our work, similar to Mokhberian et al. (2020), we capture moral frames by combining the FrameAxis approach introduced in Kwak, An, and Ahn (2020) with a dictionary of moral values. FrameAxis enables the quantification of framing of a particular text using *semantic axes* (Kwak, An, and Ahn 2020). It is built upon the SemAxis approach (An, Kwak, and Ahn 2018), which defines semantic axes by the difference of opposing word pairs using their word embeddings in the vector space. FrameAxis learns in an unsupervised way by estimating the contribution of each word towards the target axis. The contribution per word is defined as the cosine similarity between its word embedding and the target axis in the vector space. For all contributions of every word in a given document, we calculate the *frame bias* and *frame intensity* of a moral frame. The frame bias corresponds to the mean of the contributions and the frame intensity to the variance of the contributions in relation to the baseline frame bias of the corpus. The latter denotes the mean of frame biases over the whole corpus.

As a dictionary of moral values, we use the Moral Foundation Dictionary version 2 (MFD-2) (Frimer et al. 2017). It is an extension of a moral values dictionary developed by Graham, Haidt, and Nosek (2009) and consists of prototypical words to *moral foundations*. Moral foundations are described in the moral foundation theory (MFT) as factors that guide emotional and ethical reactions to various social situations. MFT describes five foundations in the form of virtues and vices: (i) care/harm, i.e., the dislike for others' suffering, (ii) fairness/cheating, i.e., dislike of cheating, (iii) loyalty/betrayal, i.e., loyalty, (iv) authority/subversion, i.e., respect for authority, and (v) sanctity/degradation, i.e., concerns with purity (Mokhberian et al. 2020).

The Moral Foundation Dictionary MFD-2 assigns words to virtues and vices. As virtues and vices are opposing moral values, we use them as poles to create *moral frame axes*. Then, for each pole, we associate its words with word embeddings, i.e., the 300-dimensional GloVe representation (Pennington, Socher, and Manning 2014) trained on 840 billion tokens and calculate their centroids for virtues and vices. Each pair of virtue and vice centroid forms a semantic axis, i.e., *moral frame axis*, that we use for FrameAxis instead of individual words. For each axis, we extract the frame biases and intensities per tweet by aggregating its words' contributions (i.e., the cosine similarity with the axis) towards the corresponding moral value. Please note that we name axes using the name of the morals' virtues in the remainder of this paper, e.g., the care axis.

Validation of Moral Frame Axes

We define four properties of the word embedding space to investigate the validity of the moral frame axes. *P1: All axes should be close to the zero point.* Note that each axis is dividing a moral space into a positive and a negative part. *P1* prohibits the dominance of one pole (i.e., the pole closer to the zero point) that could be caused by an association of an overwhelming majority of words. *P2: The words associated with a pole should be semantically closer to each other than to words of the opposite pole.* If words are added to or removed from an axis, then *P2* ensures its stability. *P3: The orientation of axes should not oppose.* Adherence to *P3* allows the axes to be combinable and form a meta-axis for virtues and vices, e.g., care virtues are closer to fairness virtues than fairness vices. *P4: The orientation of axes should differ in the*

	Precision	Recall	F1-score	Accuracy
Reproduced with MFD-2				
Care	0.828	0.827	**0.827**	**0.827**
Fairness	0.729	0.728	**0.728**	0.728
Authority	0.754	0.754	0.754	0.754
Loyalty	0.895	0.889	**0.891**	**0.889**
Sanctity	0.881	0.880	0.880	0.880
Original with MFD-1				
Care	0.746	0.768	0.734	0.768
Fairness	0.662	0.774	0.681	**0.774**
Authority	0.808	0.875	**0.817**	**0.875**
Loyalty	0.802	0.873	0.816	0.873
Sanctity	0.910	0.935	**0.908**	**0.935**

Table 1: Results of classification of moral frames on the annotated Twitter corpus. The performance of MFD-2 is compared with the results from Mokhberian et al. (2020).

hyperspace. We expect the axes to be orthogonal to a certain degree. A violation of *P4*, i.e., two axes are pointing directly in the same direction, suggests that these axes likely relate to the same concept and could be combined.

A visual analysis of the moral frame axes (see Figure 1) shows the first two principal components of word embeddings using probabilistic Principal Component Analysis (PCA), moral frame axes, and up to three density regions for virtues and vices using a kernel density estimation, which has a lowest level threshold of 33%. Results indicate all the four properties hold, e.g., all the axes point in the same direction. Due to some ambiguous words, there is some overlap in the projected point clouds (e.g., unharmed). Furthermore, some words (e.g., wounds) belong to both poles, i.e., virtue and vice in the dictionary. In addition to the visual depiction, we also perform the validation numerically[1].

Validation on Annotated Tweets

To validate our approach, we perform classification of moral frames similar to Mokhberian et al. (2020) on the Twitter dataset provided by Hoover et al. (2020), which is annotated with virtues and vices. We conduct our experiments using a logistic regression classifier with the MFD-2 dictionary. Table 1 contains the results of this experiment, and a comparison of our results with the results of Mokhberian et al. (2020). While we observe similar results as Mokhberian et al., we find that the use of MFD-2 improves the F1-score on care, fairness, and loyalty, but performs worse on authority and sanctity. In terms of accuracy, we achieve a higher performance on care and loyalty using MFD-2, but a lower performance on fairness, authority, and sanctity. We conclude that the classifier accurately captures moral frames in tweets.

Experiments and Results

We perform experiments on two datasets: firstly, in tweets in the English language created by US-based politicians, which we gathered based on the Twitter user list provided by Barberá et al. (2015), and secondly, in German-speaking tweets that contain COVID-19 related content created by followers of the leaders of the five major Austrian parties. Our selection of datasets is motivated by their differences in contextual attributes, concretely their language (i.e., English vs. German), topics (i.e., various topics vs. COVID-19-related topics), account type (i.e., politicians vs. followers of top politicians), and distribution of political parties (i.e., two-party system in the US vs. multi-party system in Austria).

Datasets

For the US Twitter dataset, we collect the most recent tweets of democrats and republicans using the party-associated Twitter handles (Barberá et al. 2015). The resulting dataset consists of $1,388,198$ tweets, i.e., $704,392$ tweets from 243 democratic (D) and $683,806$ from 252 republican (R) accounts. We label the tweets according to the account owner's party affiliation.

For the Austrian Twitter dataset, we manually extract the Twitter handles of the five major Austrian parties' lead politicians, i.e., @BMeinl for the liberal party (NEOS), @WKogler for the green party (Greens), @norbertghofer for the national-focused freedom party (FPÖ), @rendiwagner for the social-democratic party (SPÖ), and @sebastiankurz for the conservative people's party (ÖVP). Then, we collect the most recent tweets of followers and labeled each tweet of the follower with the politician they follow. To avoid mutual labels, we restrict our collection to users that follow only one of the five accounts. Besides, we only consider tweets that contain COVID-19 related hashtags (e.g., #Corona). This results in a collection of $22,205$ tweets, i.e., $17,230$ tweets labeled with @sebastiankurz, $2,090$ labeled with @WKogler, $1,164$ labeled with @rendiwagner, 901 labeled with @BMeinl, and 820 labeled with @norbertghofer.

We normalize the tweets in both datasets and (i.e., lowercase, removing URLs, punctuation), remove stopwords, and apply tokenization before extracting the frame biases and intensities for training a logistic regression classifier[1].

Moral-based Framing in US-based Tweets

We group the tweets by parties and report the coefficients of the logistic regression classifier in Table 2a. The frame biases do not deviate considerably and, in general, share the same direction on all moral frames but on authority, which is positive for republicans and negative for democrats. We observe that democrats score higher in fairness and lower in sanctity, whereas republicans score higher in the frame bias for care and exhibit a high negative score in the frame bias for loyalty. Concerning the frame intensities, we observe opposing and more distinct results. The frame intensity for care is much higher for democrats, and conversely, the frame intensity for loyalty is higher for republicans. The frame intensities on fairness and sanctity agree with their corresponding frame biases, i.e., fairness has a higher frame intensity for democrats, while sanctity has a higher frame intensity for republicans. We find that our observations are congruent with Graham, Haidt, and Nosek (2009), i.e., liberals are predominantly associated with care and fairness.

[1] We provide the code, plots and examples of this research at: https://github.com/socialcomplab/icwsm21-framing

	Moral Frames	D	R
Bias	Care	2.505	3.791
	Fairness	2.130	1.115
	Authority	-0.385	2.343
	Loyalty	-1.419	-5.269
	Sanctity	0.476	2.102
Intensity	Care	**9.701**	-0.634
	Fairness	**4.376**	-8.154
	Authority	-3.453	-6.329
	Loyalty	0.166	**9.956**
	Sanctity	-2.967	**3.261**

(a) Moral frames in US politics. Democrats (D) and republicans (R) differ most in terms of frame intensities (in bold).

	Moral Frames	@BMeinl	@WKogler	@norbertghofer	@rendiwagner	@sebastiankurz
Bias	Care	-0.788	-0.463	**-4.682**	2.141	**6.931**
	Fairness	-1.375	**12.494**	-9.165	2.408	**-13.408**
	Authority	-0.035	**-1.078**	-0.366	**2.561**	-0.291
	Loyalty	-0.078	-0.998	**2.627**	**-7.987**	0.649
	Sanctity	**-3.673**	-0.457	**15.645**	-0.145	0.458
Intensity	Care	0.077	-0.039	0.001	-0.010	-0.011
	Fairness	0.072	-0.046	0.038	-0.066	0.018
	Authority	0.008	0.002	-0.009	0.007	-0.003
	Loyalty	0.003	0.003	-0.003	0.001	-0.009
	Sanctity	0.104	-0.008	-0.023	-0.020	-0.034

(b) Moral frames in Austrian politics. Frame biases are distinct between the followers of the party leaders, whereas intensities are very small in comparison. Minimum and maximum of frame biases per moral are in bold. Frame bias in fairness exhibits the greatest difference.

Table 2: Reported results correspond to the coefficients of the logistic regression classifier.

Moral-based Framing in Austrian-based Tweets

To investigate differences in moral-based framing in the Austrian Twitter dataset, we first translate the content of the MFD-2 dictionary. To that end, we use a list of sample translations of positive and negative valence words (Weichselbaum, Leder, and Ansorge 2018) and two sets of word embeddings, i.e., one for English and one for German (Grave et al. 2018). Using a translation matrix estimated from the valance word translations, we translate the words of the MFD-2 to similar words in German in terms of their word embeddings. We see that the top words seem to be congruent with the moral values, e.g., top translation of authority being *Befehl – command*, but also observe words of opposite moral values in their vicinity, e.g., harm having *Schadenfreude – malicious joy* as the second, and *Freude – joy* and third nearest neighbor. Such inconsistencies are expected since we previously established that some words are neither clearly associated with virtues nor vices.

Then, we group the tweets by followers of Austrian party leaders and report the coefficients of the logistic regression classifier in Table 2b. We find substantial differences in frame biases between the tweets of the groups, but not in their frame intensities. The reported frame biases reaffirm the parties' public perception, with fairness having a stronger association with left parties (with @WKogler followers being the highest), while sanctity is predominantly associating with right parties (i.e., the highest for @norbertghofer followers). Noteworthy, the followers of @sebastiankurz have the lowest association with fairness, which might indicate a contention point between the viewpoints of the governing coalition, i.e., the ÖVP (@sebastiankurz) and Greens (@WKogler). Moreover, the results show that @sebastiankurz followers are mostly associated with care, a moral frame that is prevalent in the government's COVID-19 information campaign through the slogan *”Schau auf dich - schau auf mich”*, which translates to *”take care of you - take care of me”*. Followers of @rendiwagner, who is also a scientist and epidemiologist, are associated with authority. We suspect that is the result of her emphasizing to listen to doctors and experts. For followers of @BMeinl, all frame biases are negative, which we relate to the party being an opposition party arguing against government COVID-19 policies. In summary, we find differences in the moral framing of the tweets on COVID-19 of the followers of the party leaders that reflect the ideology and messages of the corresponding political parties.

Conclusion

In conclusion, our experimental results show that the moral framing in the tweets of US-based politicians and the tweets of the followers of Austrian politicians is congruent with the public perception of the political parties. In the tweets from US-based politicians, we find that democrats are associated with high frame intensity in care and fairness, whereas high frame intensity in loyalty and sanctity is associated with republicans. In the tweets from followers of the five major Austrian parties' leaders, we find that high frame bias in fairness is mostly associated with followers of the green party's leader, while high frame bias in sanctity predominantly indicates followers of the freedom party's leader. Besides, we find that followers of the ruling conservative party's leader have a notable frame bias towards care in the case of COVID-19-related tweets. We attribute this to the followers' adoption of the framing of the conservative COVID-19 slogan that stresses caring. From a methodological perspective, our experiments show that the use of the extended moral foundations dictionary MFD-2 increases the accuracy of moral frame characterization.

We recognize several limitations of our work: our analysis is restricted to two specific political Twitter datasets. We chose these datasets, as the interpretation of results requires the researchers' domain understanding and language skills. Through making a validity analysis of the approach, we aimed to mitigate the potential impact of constraints. Also, since we did not filter out retweets, 63 tweets in the Austrian dataset are from the political party leaders.

For future work, we aim to research the interplay of frame bias and intensity in more detail. We will also study how followers engage with moral frames shared by politicians and if they are more prevalent in retweets or comments.

References

An, J.; Kwak, H.; and Ahn, Y.-Y. 2018. SemAxis: A Lightweight Framework to Characterize Domain-Specific Word Semantics Beyond Sentiment. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2450–2461.

Barberá, P.; Jost, J. T.; Nagler, J.; Tucker, J. A.; and Bonneau, R. 2015. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science* 26(10): 1531–1542.

Frimer, J.; Haidt, J.; Graham, J.; Dehghani, M.; and Boghrati, R. 2017. Moral foundations dictionaries for linguistic analyses, 2.0. *Unpublished Manuscript. Retrieved from:* www.jeremyfrimer.com/uploads/2/1/2/7/21278832/summary.pdf .

Graham, J.; Haidt, J.; and Nosek, B. A. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology* 96(5): 1029.

Graham, T.; Broersma, M.; Hazelhoff, K.; and Van'T Haar, G. 2013. Between broadcasting political messages and interacting with voters: The use of Twitter during the 2010 UK general election campaign. *Information, communication & society* 16(5): 692–716.

Grave, E.; Bojanowski, P.; Gupta, P.; Joulin, A.; and Mikolov, T. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Haidt, J.; and Joseph, C. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus* 133(4): 55–66.

Haidt, J.; and Joseph, C. 2007. The moral mind: How five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules. *The innate mind* 3: 367–391.

Hoover, J.; Portillo-Wightman, G.; Yeh, L.; Havaldar, S.; Davani, A. M.; Lin, Y.; Kennedy, B.; Atari, M.; Kamel, Z.; Mendlen, M.; et al. 2020. Moral Foundations Twitter Corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science* 11(8): 1057–1071.

Kusmanoff, A. M.; Fidler, F.; Gordon, A.; Garrard, G. E.; and Bekessy, S. A. 2020. Five lessons to guide more effective biodiversity conservation message framing. *Conservation Biology* 34(5): 1131–1141.

Kwak, H.; An, J.; and Ahn, Y.-Y. 2020. FrameAxis: Characterizing Framing Bias and Intensity with Word Embedding. *arXiv preprint arXiv:2002.08608* .

Mokhberian, N.; Abeliuk, A.; Cummings, P.; and Lerman, K. 2020. Moral Framing and Ideological Bias of News. In *International Conference on Social Informatics*, 206–219. Springer.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

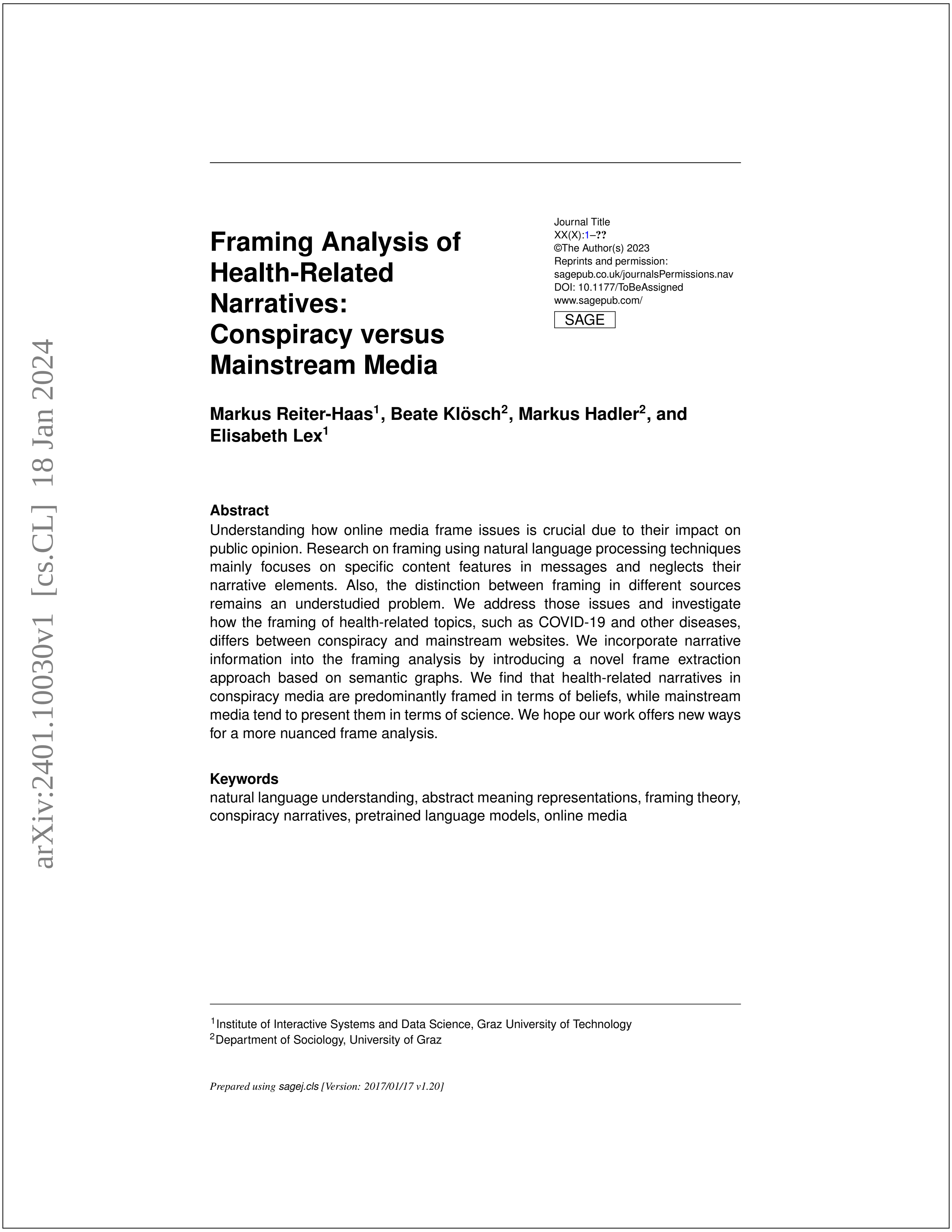
Shurafa, C.; Darwish, K.; and Zaghouani, W. 2020. Political Framing: US COVID19 Blame Game. In *International Conference on Social Informatics*, 333–351. Springer.

Tversky, A.; and Kahneman, D. 1981. The framing of decisions and the psychology of choice. *Science* 211(4481): 453–458.

Walter, D.; and Ophir, Y. 2019. News frame analysis: an inductive mixed-method computational approach. *Communication Methods and Measures* 13(4): 248–266.

Weichselbaum, H.; Leder, H.; and Ansorge, U. 2018. Implicit and explicit evaluation of visual symmetry as a function of art expertise. *i-Perception* 9(2): 2041669518761464.

Wicke, P.; and Bolognesi, M. M. 2020. Framing COVID-19: How we conceptualize and discuss the pandemic on Twitter. *PLoS ONE* 15(9): e0240010.



Journal Title
XX(X):1–??
©The Author(s) 2023
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Framing Analysis of Health-Related Narratives: Conspiracy versus Mainstream Media

Markus Reiter-Haas[1], Beate Klösch[2], Markus Hadler[2], and Elisabeth Lex[1]

arXiv:2401.10030v1 [cs.CL] 18 Jan 2024

Abstract

Understanding how online media frame issues is crucial due to their impact on public opinion. Research on framing using natural language processing techniques mainly focuses on specific content features in messages and neglects their narrative elements. Also, the distinction between framing in different sources remains an understudied problem. We address those issues and investigate how the framing of health-related topics, such as COVID-19 and other diseases, differs between conspiracy and mainstream websites. We incorporate narrative information into the framing analysis by introducing a novel frame extraction approach based on semantic graphs. We find that health-related narratives in conspiracy media are predominantly framed in terms of beliefs, while mainstream media tend to present them in terms of science. We hope our work offers new ways for a more nuanced frame analysis.

Keywords

natural language understanding, abstract meaning representations, framing theory, conspiracy narratives, pretrained language models, online media

[1] Institute of Interactive Systems and Data Science, Graz University of Technology
[2] Department of Sociology, University of Graz

Prepared using sagej.cls *[Version: 2017/01/17 v1.20]*

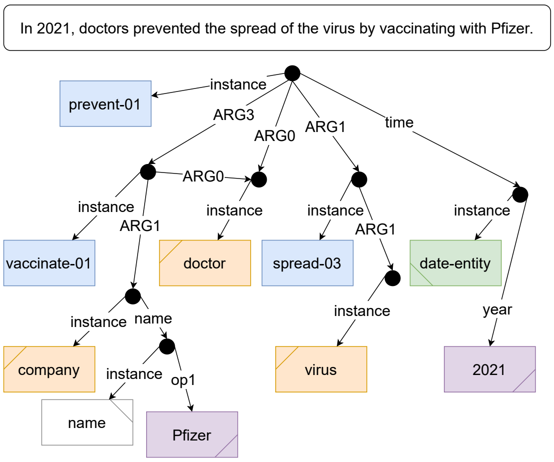


Figure 1. Example sentence (top) with its extracted AMR graph using a BART-based model. Given this representation, we can identify the narrative elements, while syntactical information such as tenses is omitted. Within the narrative, three characters are present, i.e., a *doctor* who acts twice (i.e., two *ARG0* relations) as character (orange), as well as a *company* and a *virus* (both with *ARG1* relations). The plot (blue; predicates with word senses) revolves around three frames, namely *prevent*, *vaccinate*, and *spread*. Additionally, the year 2021 (i.e., *date-entity* = setting; green) and the company *name* Pfizer are depicted as entities (purple).

Introduction

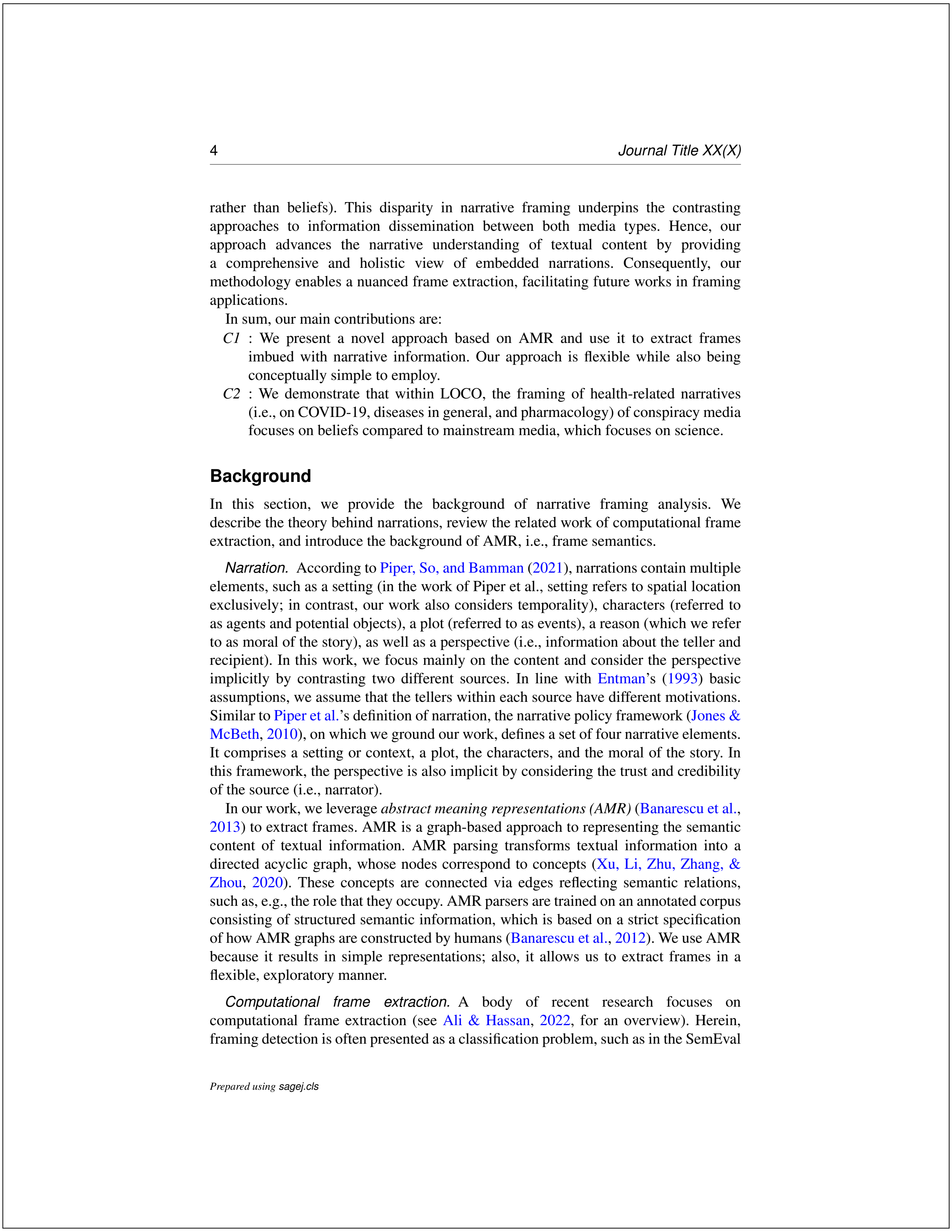
The perception of reality in human communication heavily relies on how messages are framed, leading to significant effects on human behavior. In their groundbreaking study, Tversky and Kahneman (1985) demonstrate that altering the formulation of a problem impacts people's decision making. Hence, understanding the role of framing in textual communication is a critical research direction. While many existing computational framing research studies have primarily focused on narrow topics, such as war (Wicke & Bolognesi, 2020), terrorists (Demszky et al., 2019), morality (Reiter-Haas, Kopeinik, & Lex, 2021), or blame (Shurafa, Darwish, & Zaghouani, 2020), these works often overlook the narrative content embedded within the frames. However, understanding the narrative content is essential as it plays an important role in transmitting the underlying information. Furthermore, framing in textual content is defined as promoting particular aspects of information through the selection and salience of content (Entman, 1993). Hence, a comprehensive framing analysis needs to extend beyond the identification of frames themselves and interpret why frames were used, such as supporting a particular narrative.

Narrative framing is especially critical in domains where the differences in the messages can be subtle, as exemplified in conspiracy theories. According to McLeod et al. (2022), narrative frames are abstract constructs that refer to entire messages rather than individual content features, which makes them difficult to identify. Fong, Roozenbeek, Goldwert, Rathje, and van der Linden (2021) find that despite conflicting narratives of conspiracy theories versus scientific narratives, the language employed within them has similarities in terms of word frequencies (using LIWC - Linguistic Inquiry and Word Count; Tausczik & Pennebaker, 2010). Contrary to Fong et al. hypotheses, these similarities also extend to causal (e.g., "how," "why," "because") and outgroup language (e.g., "they," "them"). Still, the authors highlight consistent differences in specific linguistic patterns, such as the prevalent use of anger-based wording to convey negative emotion within conspiratorial discourse. In our work, we aim to bridge the gap between word-frequency of messages and their framing. To that end, we propose to incorporate narrative information into the framing analysis of conspiracy theories.

We argue that most conventional text analysis methods employed in framing research, including various types of topic modeling (see Ali & Hassan, 2022, for an overview) do not accurately capture narrative information. Therefore, we propose the extraction of **semantic graphs** from textual data to conduct frame analysis. Our approach draws on the recent work of Jing and Ahn (2021), in which semantic relations are mined from textual data in the form of triples, i.e., subject, predicate, and object. In this representation, subjects and objects correspond to semantic roles such as agents and patients, while the predicate (i.e., the *verb*) connects these roles. In our work, however, we leverage semantic graphs as they allow for a more comprehensive representation of concepts, entities, and relations in textual data. This enables us to capture a broader range of semantic information beyond roles and predicates.

To that end, we utilize abstract meaning representations (AMR; Banarescu et al., 2013) as a means to transform textual content into semantic graphs. AMR gives us a structured representation of textual content in the form of AMR graphs, from which we extract their inherent semantic frames using edge information; thus considering the embedded narrative content contained within them. To validate the effectiveness of our approach, we apply it to health-related narratives mined from the language of a conspiracy corpus (LOCO; Miani, Hills, & Bangerter, 2021). LOCO contains text documents from mainstream and conspiracy websites from a time period of 2004 until 2020. The study of Miani et al. (2021) shows that the detection of conspiracy content compared to mainstream content can be challenging for humans, primarily due to ambiguity, highlighting the potential benefits of algorithmic support tools for humans. Our work aims to address this challenge by showing the difference in narrative elements, such as setting, characters, plot, and the moral of the story. Figure 1 exemplarily shows an extracted AMR graph on COVID-19 and its narrative elements. It depicts how narrative elements can be determined by edge traversal of the given semantic information. This approach allows for a more effective analysis of narratives, as compared to traditional methods relying on syntactical or word-based extraction techniques, which would prove considerably more challenging in this context.

Through our analysis, we uncover a distinctive pattern in the narrative framing employed by conspiracy media compared to mainstream media. Specifically, conspiracy media tend to employ belief-based rather than science-based arguments. Conversely, mainstream media shows the opposite tendency (i.e., towards science



rather than beliefs). This disparity in narrative framing underpins the contrasting approaches to information dissemination between both media types. Hence, our approach advances the narrative understanding of textual content by providing a comprehensive and holistic view of embedded narrations. Consequently, our methodology enables a nuanced frame extraction, facilitating future works in framing applications.

In sum, our main contributions are:

C1 : We present a novel approach based on AMR and use it to extract frames imbued with narrative information. Our approach is flexible while also being conceptually simple to employ.

C2 : We demonstrate that within LOCO, the framing of health-related narratives (i.e., on COVID-19, diseases in general, and pharmacology) of conspiracy media focuses on beliefs compared to mainstream media, which focuses on science.

Background

In this section, we provide the background of narrative framing analysis. We describe the theory behind narrations, review the related work of computational frame extraction, and introduce the background of AMR, i.e., frame semantics.

Narration. According to Piper, So, and Bamman (2021), narrations contain multiple elements, such as a setting (in the work of Piper et al., setting refers to spatial location exclusively; in contrast, our work also considers temporality), characters (referred to as agents and potential objects), a plot (referred to as events), a reason (which we refer to as moral of the story), as well as a perspective (i.e., information about the teller and recipient). In this work, we focus mainly on the content and consider the perspective implicitly by contrasting two different sources. In line with Entman's (1993) basic assumptions, we assume that the tellers within each source have different motivations. Similar to Piper et al.'s definition of narration, the narrative policy framework (Jones & McBeth, 2010), on which we ground our work, defines a set of four narrative elements. It comprises a setting or context, a plot, the characters, and the moral of the story. In this framework, the perspective is also implicit by considering the trust and credibility of the source (i.e., narrator).

In our work, we leverage *abstract meaning representations (AMR)* (Banarescu et al., 2013) to extract frames. AMR is a graph-based approach to representing the semantic content of textual information. AMR parsing transforms textual information into a directed acyclic graph, whose nodes correspond to concepts (Xu, Li, Zhu, Zhang, & Zhou, 2020). These concepts are connected via edges reflecting semantic relations, such as, e.g., the role that they occupy. AMR parsers are trained on an annotated corpus consisting of structured semantic information, which is based on a strict specification of how AMR graphs are constructed by humans (Banarescu et al., 2012). We use AMR because it results in simple representations; also, it allows us to extract frames in a flexible, exploratory manner.

Computational frame extraction. A body of recent research focuses on computational frame extraction (see Ali & Hassan, 2022, for an overview). Herein, framing detection is often presented as a classification problem, such as in the SemEval

2023 shared task (Piskorski, Stefanovitch, Da San Martino, & Nakov, 2023). In another example, Tourni et al. (2021) consider gun violence as portrayed in news headlines and lead images. In their work, the authors formalize the notion of frame concreteness derived from the tangibility of words within their headlines. They relate it to the relevance of images to the given headline. Their experiments show that news about politics has a high concreteness and relevance, whereas news about society/culture is low on both. Huguet Cabot, Dankers, Abadi, Fischer, and Shutova (2020) consider the frames security and defense, morality and fairness, and equality in the context of immigration, gun control, and death penalty. Their approach is based on RoBERTa (Liu et al., 2019), and they predict the framing of policy issues based on a joint model of emotion, metaphors, and political rhetoric. From a methodological perspective, many works also employ unsupervised learning to extract frames, such as clustering and sentiment analysis (e.g., Burscher, Vliegenthart, & Vreese, 2016, uses both). Herein, Jing and Ahn (2021) extract frames in partisan tweets related to COVID-19 by combining BERT with semantic role labeling (Shi & Lin, 2019).

We similarly position our paper as frame extraction but focus on narration instead. In our work, we extract frames related to COVID-19, albeit from conspiracy and mainstream media. Different from previous works (e.g., Jing & Ahn, 2021), we use AMR instead of topic modeling or semantic role labeling[1]. By structuring concepts within a text rather than tagging text spans, AMR allows for more flexible extraction of semantic information, which in turn benefits the interpretability of the data. For instance, extracted roles (i.e., agents and patients) are often long sequences of text in semantic role labeling, as modifiers are also included in the tagged spans of text. Hence, the full sequence of *the doctors who only recently graduated* is tagged as an agent instead of just extracting the *doctor* concept.

Frame Semantics. Frame semantics has a long history in the natural language processing community since its initial introduction by Fillmore (1976). The use of frame semantics gained momentum with the FrameNet project (Baker, Fillmore, & Lowe, 1998) and aided their adoption in natural language understanding (Fillmore & Baker, 2001). PropBank (short for proposition bank; Palmer, Gildea, & Kingsbury, 2005) provided an annotated corpus of frames and the relations to their arguments, and hence, paved the way for widely used computational methods in natural language understanding, such as semantic role labeling (Shi & Lin, 2019) and AMR parsing (Banarescu et al., 2013). Our work uses the latter due to its matching properties for the task.

As shown in Figure 1, AMR is a graph-based representation of the semantic content in the text without explicit syntax. To be more concise, AMR is a rooted, directed, acyclic graph with labeled edges and leaf nodes. AMR parsing converts texts to structured information beyond the capabilities of simple textual extraction methods. Firstly, it enriches the textual information with semantic information about data types (e.g., *date-entity*) and information (e.g., *name*). It also simplifies the semantic information by normalization (e.g., removing tenses – *prevented* to *prevent*, singularizing nouns – *doctors* to *doctor*, considering word senses – *spread-03* to indicate distribution instead of smearing, omitting the distinction between nouns and

Listing 1: Penman Notation of AMR graph

```
(p / prevent-01
   :ARG0 (d / doctor)
   :ARG1 (s / spread-03
         :ARG1 (v / virus))
   :ARG3 (v2 / vaccinate-01
         :ARG0 d
         :ARG1 (c / company
               :name (n / name
                     :op1 "Pfizer")))
   :time (d2 / date-entity
         :year 2021))
```

verbs – converting both *vaccinating* and *vaccination* to the common form *vaccinate-03*, and even substituting for named entities – *company* instead of using its name *Pfizer*).

Notation and Definitions

As the term "frame" is used in various ways in the literature (e.g., compare Entman (1993) and Fillmore (1976)), we briefly clarify at this point the specific meaning of the underlying representation and most important terms for the remainder of the paper. Our definitions are adapted to be specific for the *framing analysis imbued with narrative information*.

As an alternative to the graph-based representation, AMR graphs can also be represented as serialized text using the Penman notation (Kasper, 1989). The Penman notation applies to connected, rooted, directed, acyclic, and labeled graphs, such as AMR, which is often even used synonymously (Goodman, 2019). The notation has a recursive structure concerning its relations denoted by parenthesis, typically also indicated using newlines and indentation as a convention for human readability. As an example, Figure 1 is equivalent to the Penman notation in Listing 1.

In the following, we will clarify the definitions using the AMR annotation guideline (Banarescu et al., 2012) and the provided example. We highlight the **main building blocks** in bold, the references to the *Penman example* in italic, and ’lexical definitions’ with single quotation marks.

Semantic frames are defined in a language resource (here, PropBank[2]), which comprises a predefined set of **predicates** including their **sense** and associated **frame arguments**. For instance, consider the first two lines in Listing 1. The frame *(p / prevent-01 :ARG0 (d / doctor))* comprises a predicate (*prevent-01*), with *doctor* as frame argument. Frame arguments have a **semantic role** assigned to them (e.g., *ARG0* for *doctor*). In the given example, the *prevent* predicate only has a single sense (i.e., 01 with the meaning of ’stopping in advance’). However, when considering the next two lines (*:ARG1 (s / spread-03 :ARG1 (v / virus))*), *spread-03* refers to spread in the third sense - ’cause to be widely located or distributed’ rather than referring to ’smear’ (i.e., spread-01) or ’extend’ (i.e., spread-02). Also, frame arguments can themselves be frames, with *spread-03* being an argument of *prevent-01* as denoted by the *ARG1* relation. Therefore, frames can contain substructures, such as *virus* belonging to *spread-03* as argument. Alternatively, frame arguments can be **concepts** comprising words and phrases (e.g., *doctor* or *virus*). Furthermore, Penman uses variables (equivalent to nodes in the graph) to distinguish between their **instances** and denote instance relations with a slash. Hence, in the example, the *d* variable of the semantic frame *(v2 / vaccinate-01 :ARG0 d)* refers to the same *doctor* as in the *(p / prevent-01 :ARG0 (d / doctor))*. Finally, nodes can have associated **attributes** (e.g., *company* has the *name* attribute of *"Pfizer"*).

Here, we want to emphasize the subtle difference between semantic frames (also called "linguistic frames") and narrative framing (i.e., a form of communicative frames), which operate on different levels of language and communication, respectively (Sullivan, 2023)[3]. In the task at hand, language is essential for studying narrative framing, and therefore depend on semantic frames (which is not necessarily the case for other types of communicative frames, such as art (Sullivan, 2023)). Consequently, we use the precisely defined semantic frames as a basis to study more complex communicative frames (i.e., narrative frames).

When considering the narrative information in the framing analyses, i.e., **narrative framing**, we refer to AMR-subgraphs as potential narratives, such as *(p / prevent-01 :ARG0 (d / doctor) :ARG1 (s / spread-03 :ARG1 (v / virus)))*, and instances, as well as attributes, as **narrative elements**. In the remainder of the paper, we refer to semantic frames as *frames*, frame arguments as *arguments*, and narrative elements as *elements*. For brevity, we omit semantic roles and use subject-verb-object notation where applicable (e.g., *doctor prevent-01 spread-03*).

Method

We present our approach for frame mining in text-based content based on AMR, comprising a pipeline of three main components (i.e., contribution *C1*):

1. *AMR parsing* with a pretrained BART model and Penman decoder.
2. *Mining narrative elements*, such as characters, plot, setting, and the moral of the story.
3. *Analysis of narrative information* concerning differences in word usage, embedding spaces, and subgraphs.

We first describe the main components in detail, before providing a complete conceptual *description of the pipeline* from a technical perspective.

AMR Parsing

For AMR parsing, i.e., the conversion from text to AMR graphs, we use the AMRlib[4] with a pretrained BART-based model (i.e., *parse_xfm_bart_base-v0_1_0*; based on Lewis et al., 2019). The model was trained on the AMR Annotation Release 3.0 (LDC2020T02; Knight et al., 2021) based on the PropBank annotations (Palmer et al., 2005) and has a SMATCH score of 82.3, which is a semantic matching score based on F1-measure (refer to Cai & Knight, 2013, for details). AMR parsing also has the advantage of applying multiple linguistic tasks simultaneously, such as co-reference resolutions via reentrants in the graph (refer to Szubert, Damonte, Cohen, & Steedman, 2020, for an overview of different types of reentrants) and named entity recognition (via the name attribute). Hence, it alleviates the need for building sophisticated processing pipelines. The output of the AMR parser is in PENMAN notation, which is transformed into a graph for mining via the Penman library (Goodman, 2020).

Mining Narrative Elements

We first introduce the narrative policy framework (Jones & McBeth, 2010), which describes an empirical approach to studying policy narratives. Thereby, a narrative structure consists of characters (e.g., heroes/villains or victims), a plot (i.e., actions), a setting or context, as well as the moral of the story. The narrative policy framework provides the theoretical grounding for mining the narrative information. Specifically, we extract the narrative elements, i.e., characters and plots, by considering the AMR edge information.

Characters and plots are described as simple (<subject>, <verb/predicate>, <object>) triples, such as, e.g., *we protect them*. A more general representation of the plot and its corresponding characters is as a variable-length tuple of the format: (<predicate>, <argument0>, <argument1>, . . . , <argumentN>), which resembles PropBank frames. Frame arguments can be other frames (e.g., vaccinate-01 or spread-03), concepts (e.g., nouns such as doctor, company, or virus), or attributes (e.g., named entities such as Pfizer or a year such as 2021).

Characters. For the characters (orange), we consider instances of *ARG0* or *ARG1* roles. While frames can have more than these two arguments (i.e., *ARG2* and beyond), they tend to appear less often and hence play a less important role, as the highest-ranked (i.e., the lowest number) argument precedes according to the PropBank guidelines (Babko-Malaya, 2005). Hence, we focus on the first two arguments for simplicity. Due to reentrants in the graph, characters can assume multiple (possibly even different) roles. This is exemplified in Figure 1 as seen by the doctor in the example, who acts twice as a character.

According to the narrative policy framework, characters can be categorized as heroes, villains, or victims. In the present work, we do not distinguish between these subtypes of characters[5].

Plot. For the plot (blue), we use the predicates of the semantic frames directly. To find the *frames* (i.e., predicates), we reverse the traversal of the graph (i.e., go up from *ARG0* or *ARG1* arguments to parent nodes and towards their instances). One observation is that the plot is driven by verbs and indicated by other words that can be encoded in frames. In the example given in Figure 1, the spread is part of the plot, as it suggests the distribution of a virus (i.e., *ARG1*) but does not detail who the spreader is (i.e., misses an *ARG0*).

Setting. For the setting (green), we consider the special *time* and *location* relations. These represent the context in which the narration is embedded and are typically associated with attributes (purple), such the specific *year*. Compared to the characters and plot, the attributes can be more diverse as they are not bounded by the number of common words and their normalization. For instance, considering the range of pharmaceutical companies researching vaccinations for COVID-19, some associated named entities are more commonly portrayed (e.g., Pfizer), while others are only rarely mentioned (e.g., Sanofi). Similarly, certain temporal or spatial information might appear more frequently in relation to particular topics (e.g., 2021 for COVID-19). Nevertheless, these types of information are unbounded by definition. In the analysis, we thus differentiate between types of the settings, such as the narrative refer to a *year*, and their specific attributes, e.g., 2021.

Moral of the Story. Similar to the setting, the moral of the story (i.e., reason) relies on specific relations, i.e., *purpose* and *cause*. Unlike the setting, these relations often comprise concepts or even complete subgraphs. Here, we use the top element (i.e., root of the subgraph), which carries the most meaningful information. Moreover, as many sentences do neither include a purpose nor cause, such relations are only sparsely available. Nevertheless, they provide important narrative information.

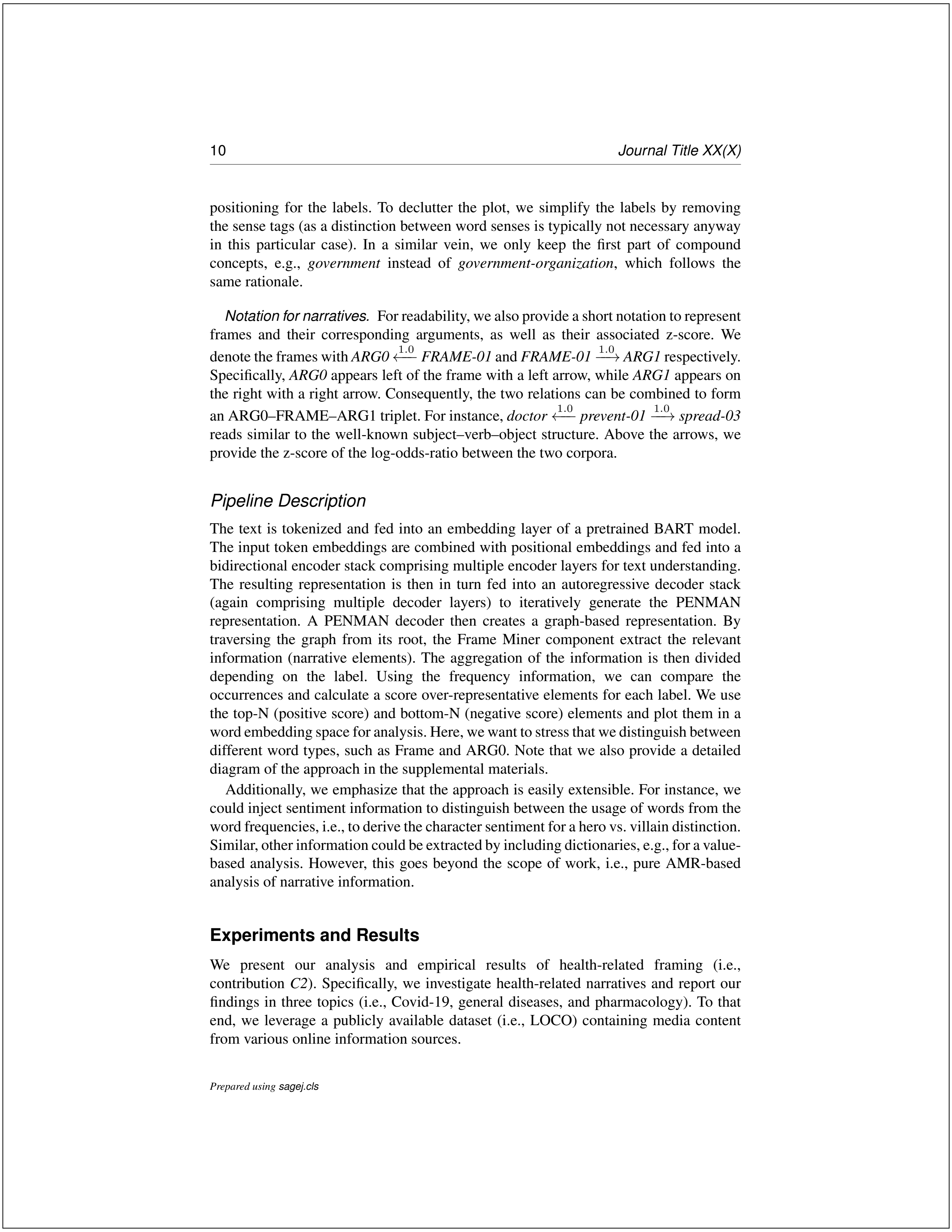
Analysis of Narrative Information

We compare the narrative information extracted between the mainstream and conspiracy corpus. Herein, we use the log-odds ratio to diminish the influence of predominant characters and plot devices in terms of relative frequency to each other, e.g., similar to Jing and Ahn (2021). However, we leverage smoothed log-odds ratio instead of informative Dirichlet priors (Monroe, Colaresi, & Quinn, 2008), and thus do not require a separate background corpus. The complete equation, which also includes Z-score normalization, is given by:

$$z_w = \frac{log\frac{f_i(w)+1}{n_i-f_i(w)+1} - log\frac{f_j(w)+1}{n_j-f_j(w)+1}}{\sqrt{\frac{1}{f_i(w)+1}+\frac{1}{f_j(w)+1}}} \tag{1}$$

, where $f_i(w)$ and $f_j(w)$ represent the frequency of a given word w in its corresponding sub-corpus, while n_i and n_j represent the total number of words per sub-corpus (i.e., $n_i = \sum_{w \in V} f_i(w)$ with V containing all words and similarly for $n_j = \sum_{w \in V} f_j(w)$). Hence, the enumerator of Equation 1 corresponds to a relative probability that is symmetric due to the log transformation, while the denominator accounts for the variance. Consequently, over-represented words in the given sub-corpus get a high absolute value. The sign indicates the dominant sub-corpus, while the magnitude of the score (i.e., absolute value) is equivalent for both sub-corpora. Hence, negative values show the over-representativeness in the alternative sub-corpus without requiring recalculation.

Visualization for elements. We plot the over-represented words (indicating plot, characters, setting, and moral of the story) in a shared two-dimensional embedding space using UMAP-reduced embeddings (McInnes, Healy, & Melville, 2018) of the model's input layer side by side for comparison[6]. The positioning of the plot improves the analysis by positioning semantically similar words in a similar region, and thus improves the subsequent interpretation. For readability, we use a force-adjusted



positioning for the labels. To declutter the plot, we simplify the labels by removing the sense tags (as a distinction between word senses is typically not necessary anyway in this particular case). In a similar vein, we only keep the first part of compound concepts, e.g., *government* instead of *government-organization*, which follows the same rationale.

Notation for narratives. For readability, we also provide a short notation to represent frames and their corresponding arguments, as well as their associated z-score. We denote the frames with $\textit{ARG0} \xleftarrow{1.0} \textit{FRAME-01}$ and $\textit{FRAME-01} \xrightarrow{1.0} \textit{ARG1}$ respectively. Specifically, *ARG0* appears left of the frame with a left arrow, while *ARG1* appears on the right with a right arrow. Consequently, the two relations can be combined to form an ARG0–FRAME–ARG1 triplet. For instance, $\textit{doctor} \xleftarrow{1.0} \textit{prevent-01} \xrightarrow{1.0} \textit{spread-03}$ reads similar to the well-known subject–verb–object structure. Above the arrows, we provide the z-score of the log-odds-ratio between the two corpora.

Pipeline Description

The text is tokenized and fed into an embedding layer of a pretrained BART model. The input token embeddings are combined with positional embeddings and fed into a bidirectional encoder stack comprising multiple encoder layers for text understanding. The resulting representation is then in turn fed into an autoregressive decoder stack (again comprising multiple decoder layers) to iteratively generate the PENMAN representation. A PENMAN decoder then creates a graph-based representation. By traversing the graph from its root, the Frame Miner component extract the relevant information (narrative elements). The aggregation of the information is then divided depending on the label. Using the frequency information, we can compare the occurrences and calculate a score over-representative elements for each label. We use the top-N (positive score) and bottom-N (negative score) elements and plot them in a word embedding space for analysis. Here, we want to stress that we distinguish between different word types, such as Frame and ARG0. Note that we also provide a detailed diagram of the approach in the supplemental materials.

Additionally, we emphasize that the approach is easily extensible. For instance, we could inject sentiment information to distinguish between the usage of words from the word frequencies, i.e., to derive the character sentiment for a hero vs. villain distinction. Similar, other information could be extracted by including dictionaries, e.g., for a value-based analysis. However, this goes beyond the scope of work, i.e., pure AMR-based analysis of narrative information.

Experiments and Results

We present our analysis and empirical results of health-related framing (i.e., contribution *C2*). Specifically, we investigate health-related narratives and report our findings in three topics (i.e., Covid-19, general diseases, and pharmacology). To that end, we leverage a publicly available dataset (i.e., LOCO) containing media content from various online information sources.

Dataset and Preprocessing

We use the LOCO dataset (Miani et al., 2021), which contains documents collected from English-speaking news websites concerning both mainstream and conspiracy media[7]. The documents were collected from May to July 2020 via web scraping (thus, also including older documents dating back to 2004 for the oldest conspiracy document) using a combination of predefined sources and manual seed selection while excluding non-English domains from the collection (refer to Miani et al. (2021) for the complete data collection and processing details). Documents are labeled as *conspiracy* if they originate from a website known to publish "unverifiable information that is not always supported by evidence" (as determined by the Media Bias/Fact Check list[8]) and mainstream otherwise. The corpus comprises $72,806$ mainstream documents from 92 websites and $23,937$ conspiracy documents from 58 websites on 47 seeds.

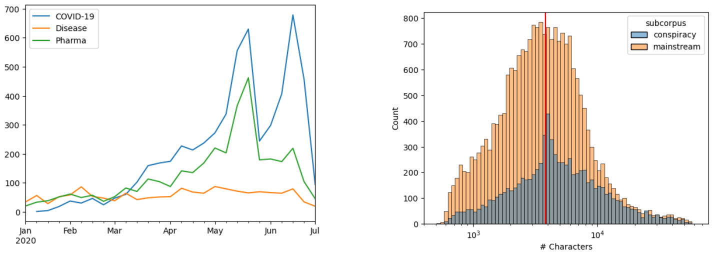
We consider three health-related subcorpora. First, we focus on the documents on COVID-19-related topics, i.e., we use the following seeds as defined by LOCO: *vaccine.covid*, *covid.19*, and *coronavirus*. Second, we consider documents related to disease with the seeds *aids*, *cancer*, *zika.virus*, and *ebola*. Third, the pharmacology LOCO subset comprises documents with the seeds *vaccine*, *pharma*, and *drug*.

Considering the time (see Figure 2a), we observe that the majority of documents for COVID-19 and pharmacology appear in 2020 with peaks in May and June (COVID-19 specific peak) just before the end of data collection on July 3rd, 2020. Note, however, that both disease and pharmacology have more documents overall compared to COVID-19, which is in turn more clumped in 2020. This in turn also result in a greater number of graphs and narrative elements.

In Figure 2b, we observe that both mainstream and conspiracy media resemble a lognormal distribution in terms of document length (we only depict the distribution for the number of characters, but observe similar distributions for the number of words and sentences). On average, each document consists of 5455 characters, 1009 words, and 38 sentences. However, conspiracy documents are more concentrated near the median (i.e., red line at 3805). We extract the AMR graphs using the methodology described in the Method Section[9]. The detailed statistics are described in Table 1.

Analysis of Narrative Information

We contrast the narrative framing of COVID-19 in the mainstream and conspiracy corpus. In Figure 3a (we also provide the Table with the top 15 over-represented words per corpus per narrative element type with their associated score in the supplementary materials.), we observe that conspiracy media tends to focus on argumentation frames as plot, such as *believe*, *claim*, *lie*, and *oppose*. Conversely, mainstream media focuses on action-oriented frames like *develop*, *spread*, and *reopen*. Similarly, mainstream media uses science-related characters such as *scientists*, *vaccines*, *antibodies*, and *proteins*. In comparison, conspiracy media use typical characters that suggest large-scale conspiracies, such as *world*, *elites*, *truth*, and *power*. When considering the contexts, we note that conspiracy media is more focused on the *now*, such as *today* or *tomorrow*, rather than specific *weeks* or *months* as is the case for mainstream media. Finally, when considering the moral of the story, conspiracy media reasons more concerning *alarm*, while mainstream uses *information* in its narratives.



(a) Number of documents in LOCO per subcorpus since 2020. We observe an increasing trend for COVID-19 and Pharmacology, with two separate peaks in May and June. Hence, it resembles the COVID-19 waves in English-speaking countries (Mathieu et al., 2020). The relative peaks are about equal for COVID-19, while the first peak is far more pronounced in Pharmacology. No such trend is present for the disease dataset. We omitted plotting the data before 2020 due to a lack of noteworthy peaks.

(b) Histogram of document length, i.e., number of characters, in the three subcorpora of LOCO. The X-axis is on a log scale. Hence, both conspiracy and mainstream media resemble lognormal distributions, but conspiracy media is more pronounced in the median (as depicted by the red line). This observation is consistent both when considering individual subcorpora, and concerning the number of words and sentences. Overall, there are more mainstream than conspiracy documents.

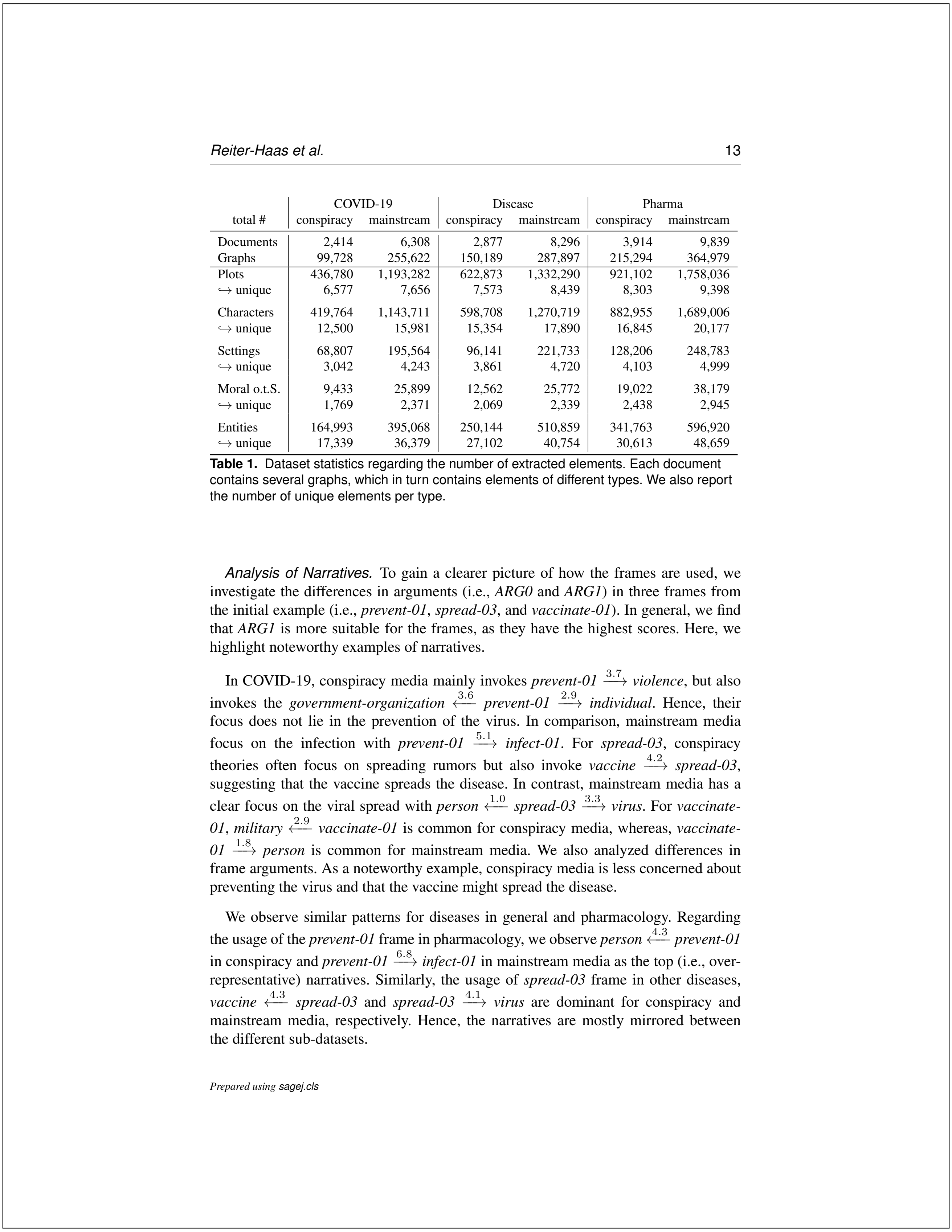
Figure 2. Details of the LOCO dataset in terms of temporality and distribution.

Another difference is the war-focused framing in conspiracy media (e.g., using *destroy* as frame, *military* as character, and *counter* as moral of the story). Whereas mainstream media has a more health-oriented framing (e.g., *infect* being used both for the plot and as character, while *treat* acts as rationale). Besides, we also briefly investigated the associated attributes (see Table in supplementary materials[10]), such as named entities, where we observe that narratives in conspiracy media revolve about people, such as *Gates* and *Trump*, as well as religion (e.g., *Jews* and *Christians*) and have a focus on *US*/*China*. In comparison, mainstream media focus on institutions, such as *universities* and the *NHS*.

In the disease dataset (see Figure 3b), we observe many similarities to the COVID-19 dataset. However, we also notice a shift, especially in the entities of mainstream media, toward global south countries where the diseases are more prevalent. Furthermore, in conspiracy media, the narrations shift toward non-natural origins such as *engineering*, *weapons*, and *chemical*.

In the pharmacology dataset (see Figure 3c), the mainstream media uses the drug company names as entities and the *development* and *manufacturing* as plots with treatment-related characters such as *dose*. Conspiracy media shows its mistrust with terms like *corrupt*, *kill*, and *control*.

While all three datasets exhibit similarities, we also observe specific differences. Most notably, the mainstream disease dataset has a stronger emphasis on the role of women due to female-associated elements (e.g., *she*, *woman*, *pregnancy*, *care*).



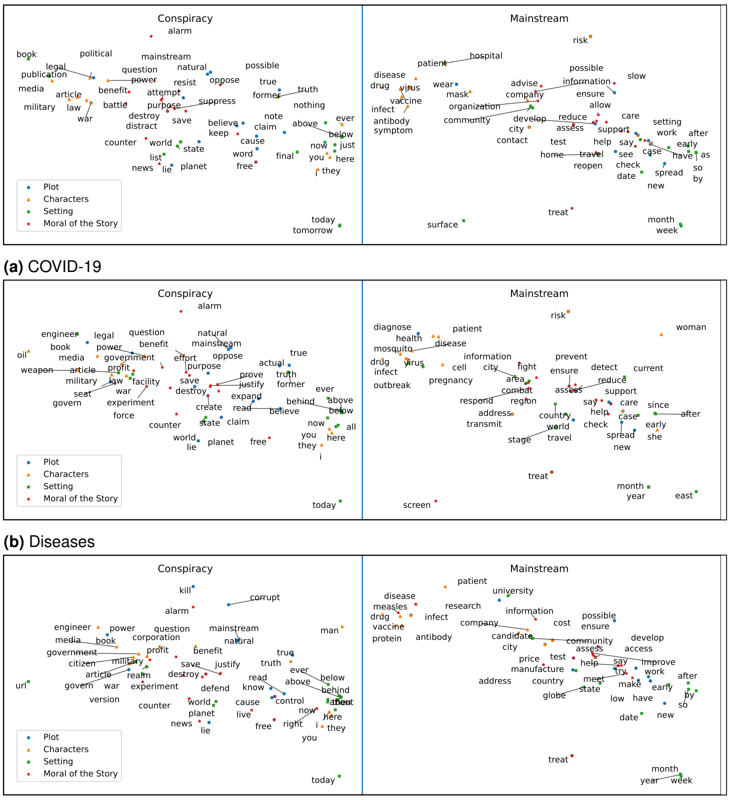
	COVID-19		Disease		Pharma	
total #	conspiracy	mainstream	conspiracy	mainstream	conspiracy	mainstream
Documents	2,414	6,308	2,877	8,296	3,914	9,839
Graphs	99,728	255,622	150,189	287,897	215,294	364,979
Plots	436,780	1,193,282	622,873	1,332,290	921,102	1,758,036
↪ unique	6,577	7,656	7,573	8,439	8,303	9,398
Characters	419,764	1,143,711	598,708	1,270,719	882,955	1,689,006
↪ unique	12,500	15,981	15,354	17,890	16,845	20,177
Settings	68,807	195,564	96,141	221,733	128,206	248,783
↪ unique	3,042	4,243	3,861	4,720	4,103	4,999
Moral o.t.S.	9,433	25,899	12,562	25,772	19,022	38,179
↪ unique	1,769	2,371	2,069	2,339	2,438	2,945
Entities	164,993	395,068	250,144	510,859	341,763	596,920
↪ unique	17,339	36,379	27,102	40,754	30,613	48,659

Table 1. Dataset statistics regarding the number of extracted elements. Each document contains several graphs, which in turn contains elements of different types. We also report the number of unique elements per type.

Analysis of Narratives. To gain a clearer picture of how the frames are used, we investigate the differences in arguments (i.e., *ARG0* and *ARG1*) in three frames from the initial example (i.e., *prevent-01*, *spread-03*, and *vaccinate-01*). In general, we find that *ARG1* is more suitable for the frames, as they have the highest scores. Here, we highlight noteworthy examples of narratives.

In COVID-19, conspiracy media mainly invokes *prevent-01* $\xrightarrow{3.7}$ *violence*, but also invokes the *government-organization* $\xleftarrow{3.6}$ *prevent-01* $\xrightarrow{2.9}$ *individual*. Hence, their focus does not lie in the prevention of the virus. In comparison, mainstream media focus on the infection with *prevent-01* $\xrightarrow{5.1}$ *infect-01*. For *spread-03*, conspiracy theories often focus on spreading rumors but also invoke *vaccine* $\xrightarrow{4.2}$ *spread-03*, suggesting that the vaccine spreads the disease. In contrast, mainstream media has a clear focus on the viral spread with *person* $\xleftarrow{1.0}$ *spread-03* $\xrightarrow{3.3}$ *virus*. For *vaccinate-01*, *military* $\xleftarrow{2.9}$ *vaccinate-01* is common for conspiracy media, whereas, *vaccinate-01* $\xrightarrow{1.8}$ *person* is common for mainstream media. We also analyzed differences in frame arguments. As a noteworthy example, conspiracy media is less concerned about preventing the virus and that the vaccine might spread the disease.

We observe similar patterns for diseases in general and pharmacology. Regarding the usage of the *prevent-01* frame in pharmacology, we observe *person* $\xleftarrow{4.3}$ *prevent-01* in conspiracy and *prevent-01* $\xrightarrow{6.8}$ *infect-01* in mainstream media as the top (i.e., over-representative) narratives. Similarly, the usage of *spread-03* frame in other diseases, *vaccine* $\xleftarrow{4.3}$ *spread-03* and *spread-03* $\xrightarrow{4.1}$ *virus* are dominant for conspiracy and mainstream media, respectively. Hence, the narratives are mostly mirrored between the different sub-datasets.



(a) COVID-19

(b) Diseases

(c) Pharmacology

Figure 3. Over-represented **narrative elements** (i.e., plot, characters, setting, moral of the story) on COVID-19 in conspiracy versus mainstream media. Positioning is according to 2-dimensional UMAP embedding of the AMR input layer (i.e., semantically similar words appear in similar locations), and labels are force-adjusted for readability (with lines indicating their associated positioning if moved beyond a threshold).

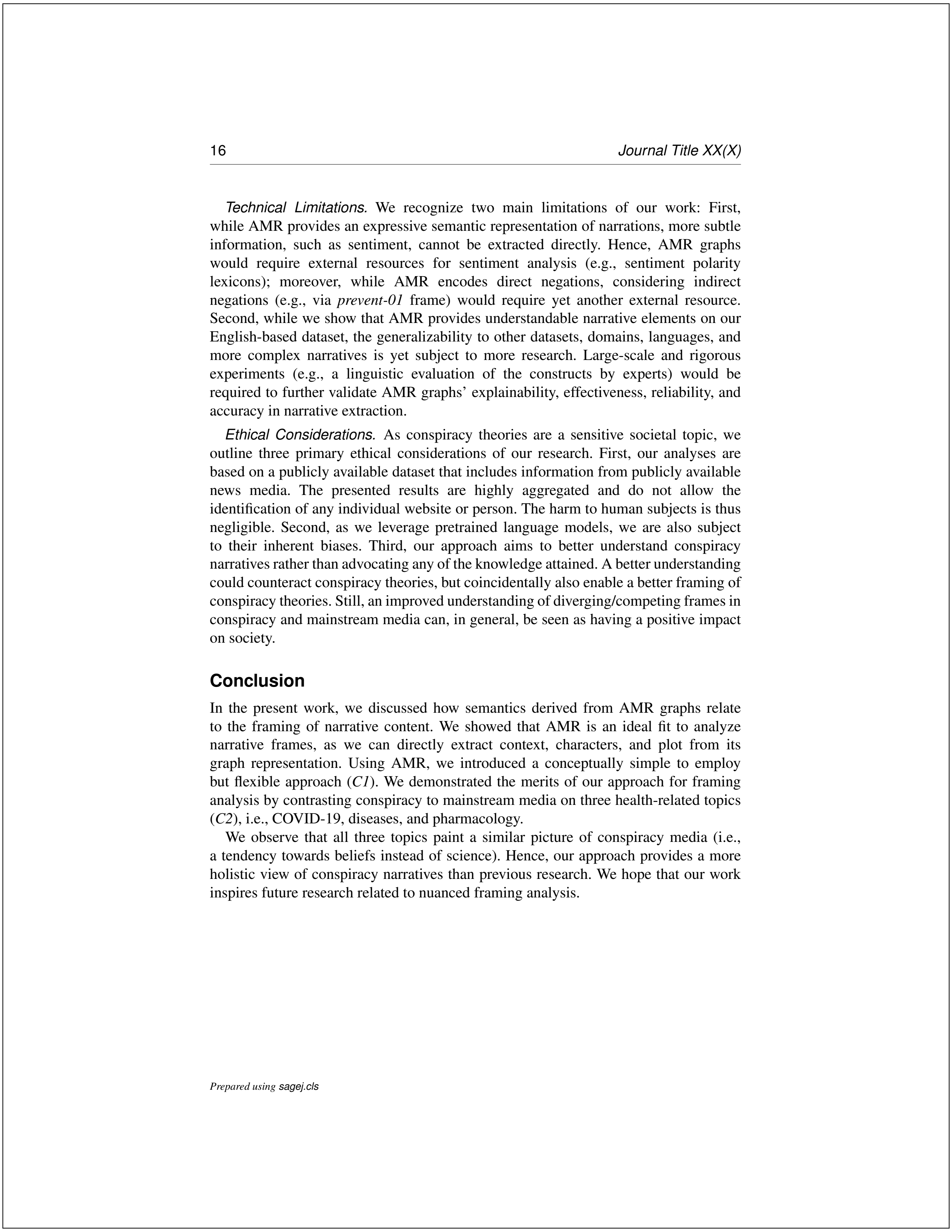
Discussion

We now briefly discuss the implications of our findings on five distinct aspects.

Narrative Themes. Apart from the well-known belief and faith focus of conspiracy outlets, our analyses of the different narratives highlights that the conspiracy sites emphasize an urgency of the social problem ("Today", "Now", etc.) and also an immediacy of the issue ("You", "I", and "We"). These characterizations are in line with the description of conspiracy adherer (Douglas et al., 2019). Interestingly, conspiracy sites also see the mainstream and media as characters and part of the game, whereas mainstream does not refer to conspiracists - which points to a non-reciprocity. Similarly, we find a corroborative emphasis ("truth", "true", "actual"), which might suggest another demarcation to mainstream media. Finally, while war-framing is often present in health-related discourse (e.g., as shown in Wicke & Bolognesi, 2020), we observe a one-sided tendency towards war-framing in conspiracy media.

Societal Implications. Our work shows a clear distinction between the narratives in conspiracy and mainstream media. While this finding on its own is expected, we can draw parallels to prior studies. For instance, Shelton (2020) suggests that the COVID-19 pandemic was the first *post-truth pandemic*. Our work complements this finding, as a belief-oriented framing in conspiracy media competes with a science-oriented framing in mainstream media. Thus, we see that the media have different lines of argumentation on the same issue, which can influence and bias people's attitudes and drive polarization, e.g., through diverging assessments of the consensus in society regarding the seriousness of the COVID-19 pandemic (Logemann & Tomczyk, 2023). A deeper understanding of the main competing narratives is therefore essential for improving the ability to spot fake news (Porshnev, Miltsov, Lokot, & Koltsova, 2021) and for automatic detection to identify and combat conspiracy theories in the media (Shahsavari, Holur, Wang, Tangherlini, & Roychowdhury, 2020). Knowledge of conspiracy narratives can thus be used to improve the dissemination guidelines of social media platforms or handbooks developed by policymakers, e.g., the 'Check Before You Share Toolkit' in the UK (Bloomfield, Magnusson, Walsh, & Naylor, 2021). It can also be used to educate society at an individual level, e.g., through 'fake news games' where players learn to identify manipulation techniques commonly used in conspiracy theories (Basol et al., 2021). Yet, we need to acknowledge that our analysis is only highlighting the structure of arguments, but does not consider the reach of these sources. However, based on previous research (Reiter-Haas, Klösch, Hadler, & Lex, 2022), we can estimate that COVID-19 conspiracy beliefs are more widespread than belief in other conspiracy theories (Uscinski et al., 2022). Assessing content and arguments in online media is thus of utmost importance.

Methodological Advancements. Our work is based on the premise that text analyses are challenging, especially when considering more abstract concepts, such as framing (partly also due to a lack of a clear definition (Entman, 1993)). Hence, graph representations such as AMR allow for a more comprehensive analysis, which is supported by our approach. Most text processing tasks are directly handled by AMR parsing, which is conceptually easy to employ using pretrained models based on the Transformer architecture (Vaswani et al., 2017). Our approach demonstrates that narrative elements can directly be mapped onto AMR and thus extracted. For instance, the four elements of the narrative policy framework (Jones & McBeth, 2010) are directly applicable. Moreover, similar elements as described by Piper et al. (2021) can be extracted (besides the perspective, as it is typically not part of the textual content). Finally, we show that we can perform a wide range of analyses on the extracted information.



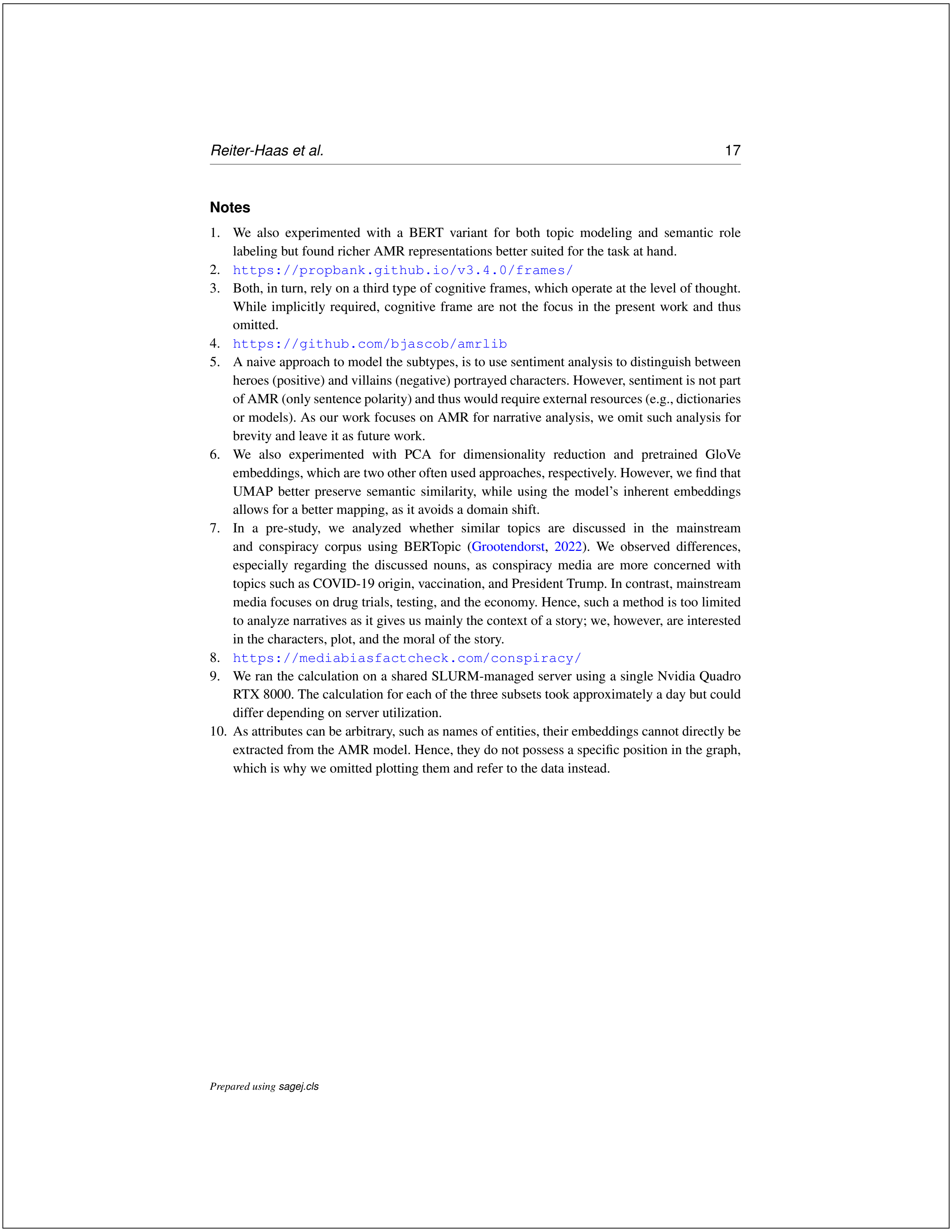
Technical Limitations. We recognize two main limitations of our work: First, while AMR provides an expressive semantic representation of narrations, more subtle information, such as sentiment, cannot be extracted directly. Hence, AMR graphs would require external resources for sentiment analysis (e.g., sentiment polarity lexicons); moreover, while AMR encodes direct negations, considering indirect negations (e.g., via *prevent-01* frame) would require yet another external resource. Second, while we show that AMR provides understandable narrative elements on our English-based dataset, the generalizability to other datasets, domains, languages, and more complex narratives is yet subject to more research. Large-scale and rigorous experiments (e.g., a linguistic evaluation of the constructs by experts) would be required to further validate AMR graphs' explainability, effectiveness, reliability, and accuracy in narrative extraction.

Ethical Considerations. As conspiracy theories are a sensitive societal topic, we outline three primary ethical considerations of our research. First, our analyses are based on a publicly available dataset that includes information from publicly available news media. The presented results are highly aggregated and do not allow the identification of any individual website or person. The harm to human subjects is thus negligible. Second, as we leverage pretrained language models, we are also subject to their inherent biases. Third, our approach aims to better understand conspiracy narratives rather than advocating any of the knowledge attained. A better understanding could counteract conspiracy theories, but coincidentally also enable a better framing of conspiracy theories. Still, an improved understanding of diverging/competing frames in conspiracy and mainstream media can, in general, be seen as having a positive impact on society.

Conclusion

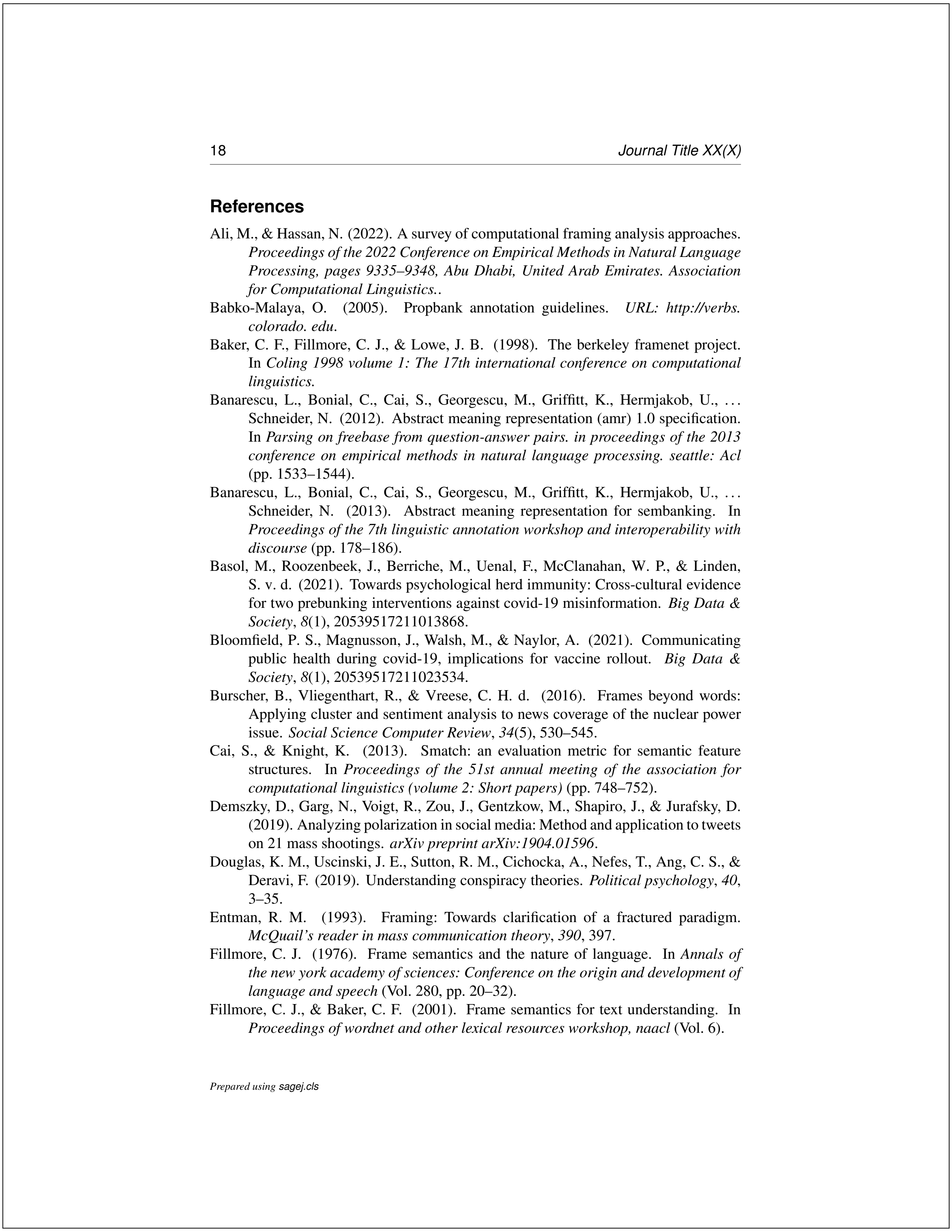
In the present work, we discussed how semantics derived from AMR graphs relate to the framing of narrative content. We showed that AMR is an ideal fit to analyze narrative frames, as we can directly extract context, characters, and plot from its graph representation. Using AMR, we introduced a conceptually simple to employ but flexible approach (*C1*). We demonstrated the merits of our approach for framing analysis by contrasting conspiracy to mainstream media on three health-related topics (*C2*), i.e., COVID-19, diseases, and pharmacology.

We observe that all three topics paint a similar picture of conspiracy media (i.e., a tendency towards beliefs instead of science). Hence, our approach provides a more holistic view of conspiracy narratives than previous research. We hope that our work inspires future research related to nuanced framing analysis.



Notes

1. We also experimented with a BERT variant for both topic modeling and semantic role labeling but found richer AMR representations better suited for the task at hand.
2. https://propbank.github.io/v3.4.0/frames/
3. Both, in turn, rely on a third type of cognitive frames, which operate at the level of thought. While implicitly required, cognitive frame are not the focus in the present work and thus omitted.
4. https://github.com/bjascob/amrlib
5. A naive approach to model the subtypes, is to use sentiment analysis to distinguish between heroes (positive) and villains (negative) portrayed characters. However, sentiment is not part of AMR (only sentence polarity) and thus would require external resources (e.g., dictionaries or models). As our work focuses on AMR for narrative analysis, we omit such analysis for brevity and leave it as future work.
6. We also experimented with PCA for dimensionality reduction and pretrained GloVe embeddings, which are two other often used approaches, respectively. However, we find that UMAP better preserve semantic similarity, while using the model's inherent embeddings allows for a better mapping, as it avoids a domain shift.
7. In a pre-study, we analyzed whether similar topics are discussed in the mainstream and conspiracy corpus using BERTopic (Grootendorst, 2022). We observed differences, especially regarding the discussed nouns, as conspiracy media are more concerned with topics such as COVID-19 origin, vaccination, and President Trump. In contrast, mainstream media focuses on drug trials, testing, and the economy. Hence, such a method is too limited to analyze narratives as it gives us mainly the context of a story; we, however, are interested in the characters, plot, and the moral of the story.
8. https://mediabiasfactcheck.com/conspiracy/
9. We ran the calculation on a shared SLURM-managed server using a single Nvidia Quadro RTX 8000. The calculation for each of the three subsets took approximately a day but could differ depending on server utilization.
10. As attributes can be arbitrary, such as names of entities, their embeddings cannot directly be extracted from the AMR model. Hence, they do not possess a specific position in the graph, which is why we omitted plotting them and refer to the data instead.



References

Ali, M., & Hassan, N. (2022). A survey of computational framing analysis approaches. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 9335–9348, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics..*

Babko-Malaya, O. (2005). Propbank annotation guidelines. *URL: http://verbs. colorado. edu.*

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The berkeley framenet project. In *Coling 1998 volume 1: The 17th international conference on computational linguistics.*

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., … Schneider, N. (2012). Abstract meaning representation (amr) 1.0 specification. In *Parsing on freebase from question-answer pairs. in proceedings of the 2013 conference on empirical methods in natural language processing. seattle: Acl* (pp. 1533–1544).

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., … Schneider, N. (2013). Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse* (pp. 178–186).

Basol, M., Roozenbeek, J., Berriche, M., Uenal, F., McClanahan, W. P., & Linden, S. v. d. (2021). Towards psychological herd immunity: Cross-cultural evidence for two prebunking interventions against covid-19 misinformation. *Big Data & Society*, *8*(1), 20539517211013868.

Bloomfield, P. S., Magnusson, J., Walsh, M., & Naylor, A. (2021). Communicating public health during covid-19, implications for vaccine rollout. *Big Data & Society*, *8*(1), 20539517211023534.

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

Cai, S., & Knight, K. (2013). Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 748–752).

Demszky, D., Garg, N., Voigt, R., Zou, J., Gentzkow, M., Shapiro, J., & Jurafsky, D. (2019). Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. *arXiv preprint arXiv:1904.01596*.

Douglas, K. M., Uscinski, J. E., Sutton, R. M., Cichocka, A., Nefes, T., Ang, C. S., & Deravi, F. (2019). Understanding conspiracy theories. *Political psychology*, *40*, 3–35.

Entman, R. M. (1993). Framing: Towards clarification of a fractured paradigm. *McQuail's reader in mass communication theory*, *390*, 397.

Fillmore, C. J. (1976). Frame semantics and the nature of language. In *Annals of the new york academy of sciences: Conference on the origin and development of language and speech* (Vol. 280, pp. 20–32).

Fillmore, C. J., & Baker, C. F. (2001). Frame semantics for text understanding. In *Proceedings of wordnet and other lexical resources workshop, naacl* (Vol. 6).

Fong, A., Roozenbeek, J., Goldwert, D., Rathje, S., & van der Linden, S. (2021). The language of conspiracy: A psychological analysis of speech used by conspiracy theorists and their followers on twitter. *Group Processes & Intergroup Relations*, *24*(4), 606–623.

Goodman, M. W. (2019). Amr normalization for fairer evaluation. *arXiv preprint arXiv:1909.01568*.

Goodman, M. W. (2020, July). Penman: An open-source library and tool for AMR graphs. In *Proceedings of the 58th annual meeting of the association for computational linguistics: System demonstrations* (pp. 312–319). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.acl-demos.35 doi: 10.18653/v1/2020.acl-demos.35

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Huguet Cabot, P.-L., Dankers, V., Abadi, D., Fischer, A., & Shutova, E. (2020, November). The Pragmatics behind Politics: Modelling Metaphor, Framing and Emotion in Political Discourse. In *Findings of the association for computational linguistics: Emnlp 2020* (pp. 4479–4488). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.findings-emnlp.402 doi: 10.18653/v1/2020.findings-emnlp.402

Jing, E., & Ahn, Y.-Y. (2021). Characterizing partisan political narrative frameworks about covid-19 on twitter. *EPJ data science*, *10*(1), 53.

Jones, M. D., & McBeth, M. K. (2010). A narrative policy framework: Clear enough to be wrong? *Policy studies journal*, *38*(2), 329–353.

Kasper, R. T. (1989). A flexible interface for linking applications to Penman's sentence generator. In *Speech and natural language: Proceedings of a workshop held at philadelphia, Pennsylvania, February 21-23, 1989.* Retrieved from https://aclanthology.org/H89-1022

Knight, K., Badarau, B., Baranescu, L., Bonial, C., Bardocz, M., Griffitt, K., … others (2021). *Abstract meaning representation (amr) annotation release 3.0.* Abacus Data Network.

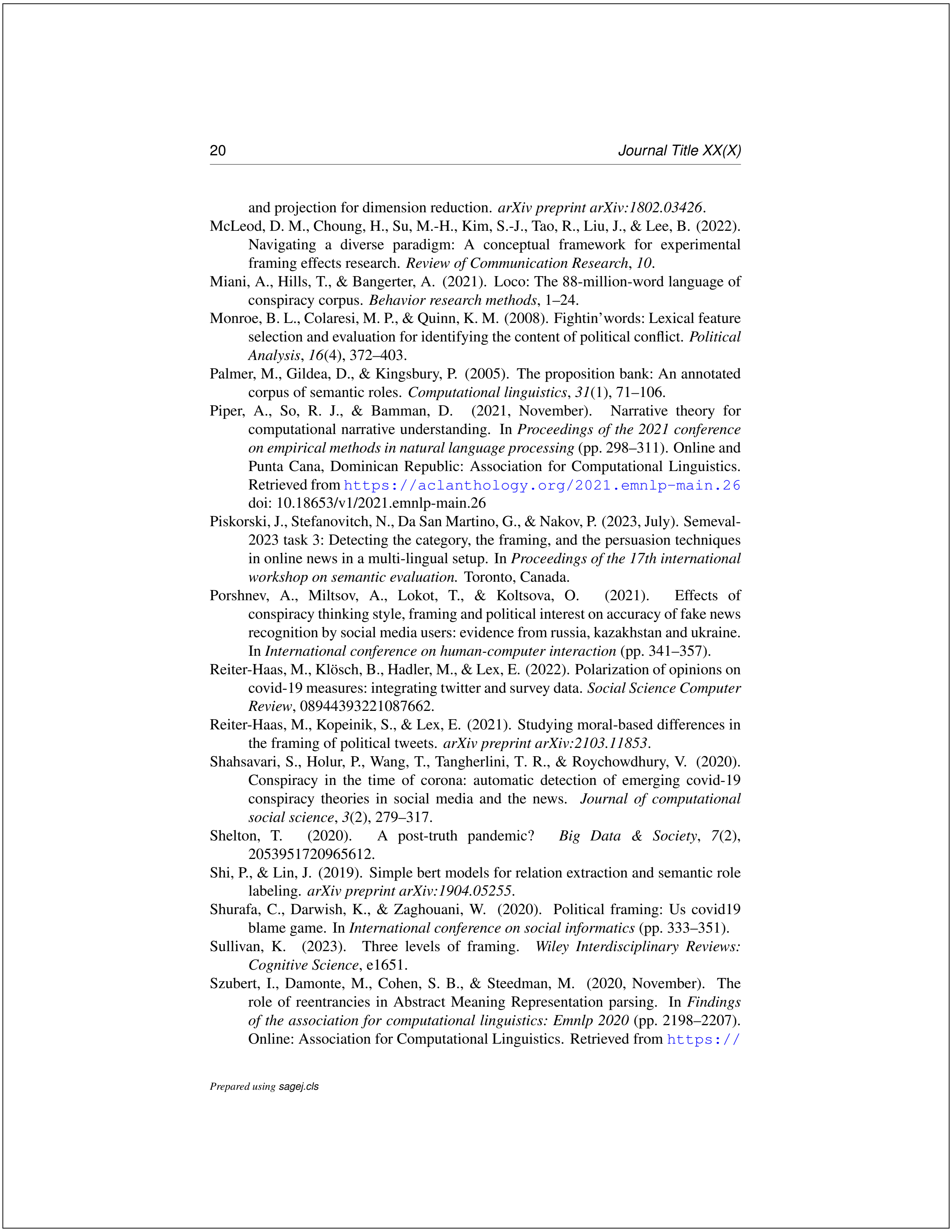
Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., … Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., … Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Logemann, H. T., & Tomczyk, S. (2023). How media reports on covid-19 conspiracy theories impact consensus beliefs and protective action: A randomized controlled online trial. *Science Communication*, *45*(2), 145–171.

Mathieu, E., Ritchie, H., Rodés-Guirao, L., Appel, C., Giattino, C., Hasell, J., … Roser, M. (2020). Coronavirus pandemic (covid-19). *Our World in Data*. (https://ourworldindata.org/coronavirus)

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation



and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

McLeod, D. M., Choung, H., Su, M.-H., Kim, S.-J., Tao, R., Liu, J., & Lee, B. (2022). Navigating a diverse paradigm: A conceptual framework for experimental framing effects research. *Review of Communication Research*, *10*.

Miani, A., Hills, T., & Bangerter, A. (2021). Loco: The 88-million-word language of conspiracy corpus. *Behavior research methods*, 1–24.

Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2008). Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, *16*(4), 372–403.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, *31*(1), 71–106.

Piper, A., So, R. J., & Bamman, D. (2021, November). Narrative theory for computational narrative understanding. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 298–311). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.emnlp-main.26 doi: 10.18653/v1/2021.emnlp-main.26

Piskorski, J., Stefanovitch, N., Da San Martino, G., & Nakov, P. (2023, July). Semeval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of the 17th international workshop on semantic evaluation.* Toronto, Canada.

Porshnev, A., Miltsov, A., Lokot, T., & Koltsova, O. (2021). Effects of conspiracy thinking style, framing and political interest on accuracy of fake news recognition by social media users: evidence from russia, kazakhstan and ukraine. In *International conference on human-computer interaction* (pp. 341–357).

Reiter-Haas, M., Klösch, B., Hadler, M., & Lex, E. (2022). Polarization of opinions on covid-19 measures: integrating twitter and survey data. *Social Science Computer Review*, 08944393221087662.

Reiter-Haas, M., Kopeinik, S., & Lex, E. (2021). Studying moral-based differences in the framing of political tweets. *arXiv preprint arXiv:2103.11853*.

Shahsavari, S., Holur, P., Wang, T., Tangherlini, T. R., & Roychowdhury, V. (2020). Conspiracy in the time of corona: automatic detection of emerging covid-19 conspiracy theories in social media and the news. *Journal of computational social science*, *3*(2), 279–317.

Shelton, T. (2020). A post-truth pandemic? *Big Data & Society*, *7*(2), 2053951720965612.

Shi, P., & Lin, J. (2019). Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*.

Shurafa, C., Darwish, K., & Zaghouani, W. (2020). Political framing: Us covid19 blame game. In *International conference on social informatics* (pp. 333–351).

Sullivan, K. (2023). Three levels of framing. *Wiley Interdisciplinary Reviews: Cognitive Science*, e1651.

Szubert, I., Damonte, M., Cohen, S. B., & Steedman, M. (2020, November). The role of reentrancies in Abstract Meaning Representation parsing. In *Findings of the association for computational linguistics: Emnlp 2020* (pp. 2198–2207). Online: Association for Computational Linguistics. Retrieved from https://

aclanthology.org/2020.findings-emnlp.199 doi: 10.18653/v1/2020.findings-emnlp.199

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, *29*(1), 24–54.

Tourni, I., Guo, L., Daryanto, T. H., Zhafransyah, F., Halim, E. E., Jalal, M., . . . Wijaya, D. T. (2021, November). Detecting frames in news headlines and lead images in U.S. gun violence coverage. In *Findings of the association for computational linguistics: Emnlp 2021* (pp. 4037–4050). Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.findings-emnlp.339 doi: 10.18653/v1/2021.findings-emnlp.339

Tversky, A., & Kahneman, D. (1985). The framing of decisions and the psychology of choice. In *Behavioral decision making* (pp. 25–41). Boston, MA: Springer.

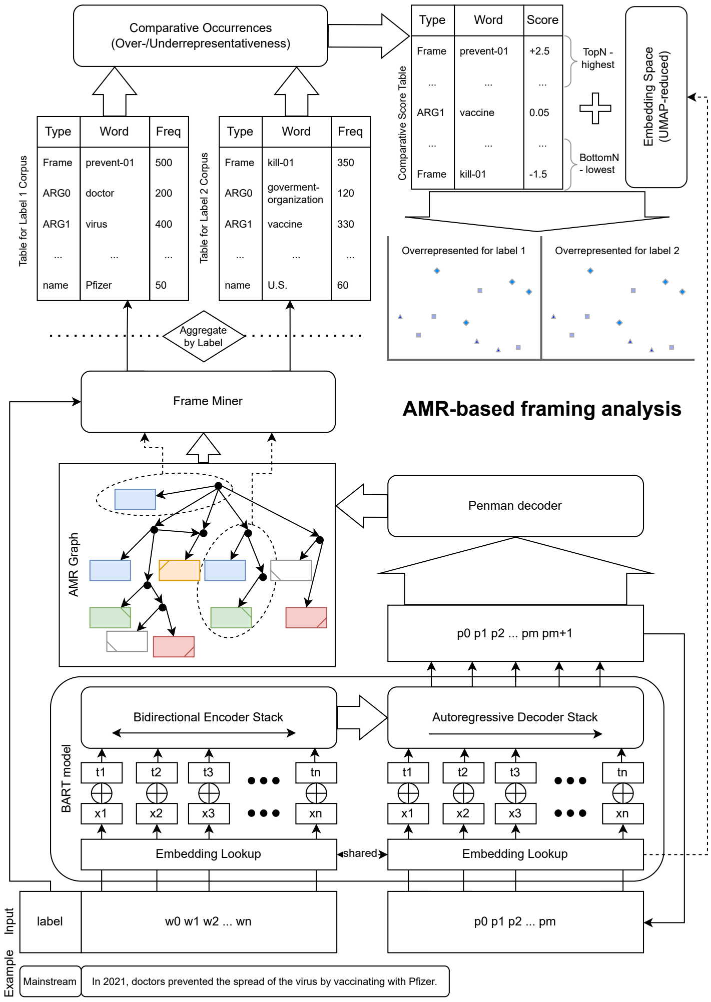
Uscinski, J., Enders, A., Klofstad, C., Seelig, M., Drochon, H., Premaratne, K., & Murthi, M. (2022). Have beliefs in conspiracy theories increased over time? *PLoS One*, *17*(7), e0270429.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wicke, P., & Bolognesi, M. M. (2020). Framing covid-19: How we conceptualize and discuss the pandemic on twitter. *PloS one*, *15*(9), e0240010.

Xu, D., Li, J., Zhu, M., Zhang, M., & Zhou, G. (2020). Improving amr parsing with sequence-to-sequence pre-training. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 2501–2511).

Supplemental Material



Comparative Score Table

Type	Word	Score
Frame	prevent-01	+2.5
...	...	...
ARG1	vaccine	0.05
...	...	...
Frame	kill-01	-1.5

Table for Label 1 Corpus

Type	Word	Freq
Frame	prevent-01	500
ARG0	doctor	200
ARG1	virus	400
...	...	...
name	Pfizer	50

Table for Label 2 Corpus

Type	Word	Freq
Frame	kill-01	350
ARG0	goverment-organization	120
ARG1	vaccine	330
...	...	...
name	U.S.	60

Figure 4. AMR-based Framing Analysis Approach Overview

		Plot		Characters		Setting		Moral of the Story		Entities	
		Conspiracy	Mainstream	Conspiracy	Mainstream	Conspiracy	Mainstream	Conspiracy	Mainstream	Conspiracy	Mainstream
COVID-19	1	destroy (24.91)	say (43.95)	i (39.65)	case-04 (29.38)	ever (27.06)	early (11.73)	counter-01 (9.03)	ensure-01 (7.48)	"America" (35.88)	"COVID-19" (47.36)
	2	claim (23.10)	test (29.09)	book (29.93)	test-01 (28.29)	now (17.16)	pandemic (10.15)	alarm-01 (8.62)	information (7.21)	"Gates" (34.28)	"Covid-19" (23.09)
	3	free (20.90)	develop (21.72)	military (25.37)	vaccine (26.00)	planet (14.88)	week (10.14)	free-01 (6.07)	test-01 (6.83)	"God" (28.94)	"SARS-CoV-2" (21.57)
	4	state (20.59)	infect (21.17)	article (24.81)	patient (24.83)	today (13.47)	by (9.25)	keep-up-05 (5.97)	treat-03 (6.83)	"Wuhan" (28.81)	"University" (19.35)
	5	believe (20.14)	possible (20.98)	amr-unknown (23.68)	disease (24.56)	above (12.25)	date-entity (8.95)	news (5.14)	help-01 (5.95)	"Bill" (28.22)	"Health" (18.17)
	6	true (19.88)	work (18.38)	you (23.17)	risk-01 (22.47)	book (12.15)	hospital (8.93)	attempt-01 (5.01)	reduce-01 (4.43)	"US" (24.03)	" coronavirus" (16.85)
	7	lie (18.90)	reopen (17.32)	law (22.56)	infect-01 (21.64)	below (11.67)	after (8.68)	benefit-01 (4.52)	assess-01 (3.85)	"Israel" (23.92)	"Oxford" (14.73)
	8	war (18.86)	case (16.54)	media (21.88)	company (21.43)	world (11.49)	home (8.35)	possible-01 (4.50)	allow-01 (3.84)	"China" (21.81)	"Zealand" (13.30)
	9	mainstream (18.55)	wear (16.28)	truth (21.42)	symptom (20.21)	final (11.30)	month (8.10)	battle-01 (4.45)	advise-01 (3.80)	"Earth" (19.88)	"Brazil" (12.72)
	10	legal (17.90)	risk (16.20)	political-party (21.35)	drug (19.70)	just (10.46)	city (8.07)	save-02 (4.44)	slow-01 (3.69)	"Trump" (19.84)	"Moderna" (12.29)
	11	cause (17.18)	care (15.91)	world (19.66)	virus (18.30)	publication (9.94)	as-of (7.24)	purpose (4.26)	see-01 (3.48)	"Big" (18.91)	"and" (11.91)
	12	oppose (16.70)	new (15.61)	power (19.32)	mask (18.23)	tomorrow (9.90)	community (7.23)	resist-01 (4.22)	organization (3.28)	"Christian" (18.78)	"NHS" (11.38)
	13	note (16.50)	spread (15.59)	war-01 (19.26)	antibody (17.48)	former (9.39)	surface (6.97)	suppress-01 (4.12)	support-01 (3.18)	"Iran" (18.76)	"England" (11.19)
	14	natural (16.45)	have (15.58)	they (19.10)	care-03 (17.08)	list (8.97)	so-far (6.68)	destroy-01 (4.05)	travel-01 (3.07)	"Bible" (18.35)	"UK" (10.57)
	15	word (16.39)	contact (15.37)	nothing (19.04)	contact-01 (16.62)	here (8.89)	setting (6.59)	distract-01 (3.99)	check-01 (3.04)	"West" (17.90)	"AstraZeneca" (10.36)
Disease	1	govern (28.01)	say (49.44)	you (43.43)	disease (80.28)	ever (26.35)	country (16.90)	alarm-01 (7.38)	treat-03 (11.21)	"US" (36.60)	"HIV" (63.26)
	2	claim (27.98)	infect (47.20)	government-organization (37.47)	infect-01 (40.25)	here (20.92)	early (15.86)	free-01 (7.23)	prevent-01 (8.93)	"America" (32.58)	"Ebola" (49.01)
	3	true (25.36)	treat (35.50)	military (30.23)	virus (38.44)	now (18.49)	after (14.60)	counter-01 (6.31)	fight-01 (6.30)	"Israel" (30.29)	"Zika" (47.78)
	4	state (25.32)	transmit (31.94)	i (29.88)	treat-03 (35.35)	planet (17.88)	outbreak-29 (13.49)	effort-01 (5.55)	ensure-01 (6.24)	"Russia" (30.14)	"cancer" (39.39)
	5	natural (25.06)	spread (30.48)	media (26.96)	outbreak-29 (33.95)	below (16.17)	area (12.38)	profit-01 (5.53)	information (6.05)	"God" (24.22)	"AIDS" (27.75)
	6	engineer (23.70)	diagnose (26.98)	article (26.71)	case-04 (33.57)	former (14.31)	world-region (12.28)	purpose (5.40)	assess-01 (5.85)	"Earth" (24.21)	"Congo" (26.01)
	7	actual (23.19)	risk (25.07)	amr-unknown (26.68)	risk-01 (31.23)	above (14.23)	region (11.82)	save-02 (5.40)	reduce-01 (5.65)	"Trump" (23.01)	"HIV/AIDS" (25.46)
	8	legal (23.01)	impregnate (23.84)	book (25.71)	woman (29.97)	world (13.63)	pregnancy (10.37)	justify-01 (5.39)	help-01 (4.88)	"Jew" (22.44)	"Health" (24.94)
	9	destroy (22.19)	prevent (23.63)	they (25.33)	drug (28.02)	all-over (11.31)	year (10.14)	benefit-01 (4.95)	address-02 (4.82)	"China" (21.80)	"WHO" (23.82)
	10	mainstream (21.16)	outbreak (23.26)	truth (23.57)	transmit-01 (27.80)	article (10.90)	east (9.95)	create-01 (4.88)	screen-01 (4.78)	"Iraq" (21.63)	"Z Zika" (21.38)
	11	lie (20.86)	care (23.14)	weapon (22.89)	patient (27.74)	war (10.65)	since (9.51)	experiment-01 (4.87)	check-01 (4.67)	"Gates" (20.82)	"Africa" (19.98)
	12	create (20.72)	detect (22.88)	law (22.08)	she (26.49)	book (10.62)	stage (9.50)	seat (4.74)	respond-01 (4.49)	"Monsanto" (20.80)	"Uganda" (18.89)
	13	oppose (20.43)	case (22.66)	power (21.90)	mosquito (25.80)	today (10.16)	current (9.16)	expand-01 (4.67)	combat-01 (4.38)	"War" (20.75)	"Organization" (17.70)
	14	read (20.19)	new (22.41)	oil (21.62)	health (25.37)	behind (10.01)	city (8.10)	prove-01 (4.66)	support-01 (4.33)	"Syria" (20.50)	"DRC" (17.47)
	15	believe (20.19)	travel (22.32)	war-01 (21.30)	cell (25.03)	facility (9.91)	month (7.94)	force-01 (4.61)	care-03 (4.27)	"Iran" (19.55)	"Dengue" (16.10)
Pharma	1	destroy (33.30)	say (71.64)	you (58.11)	company (53.50)	ever (34.17)	year (23.63)	profit-01 (9.52)	disease (9.50)	"America" (44.91)	"COVID-19" (39.24)
	2	natural (28.75)	develop (45.22)	i (38.97)	drug (52.56)	now (20.76)	date-entity (17.03)	alarm-01 (7.13)	ensure-01 (8.98)	"Big" (35.03)	"UK" (23.26)
	3	mainstream (28.57)	price (37.93)	media (31.46)	vaccine (37.57)	below (19.15)	week (16.19)	news (7.06)	treat-03 (7.90)	"God" (32.06)	"India" (23.20)
	4	govern (26.98)	cost (26.99)	government-organization (30.76)	price-01 (36.62)	here (18.45)	month (16.12)	free-01 (6.70)	develop-02 (7.40)	"Gates" (30.30)	"Covid-19" (21.92)
	5	know (26.79)	new (25.10)	truth (30.75)	disease (30.94)	planet (16.44)	by (13.70)	destroy-01 (6.57)	test-01 (6.80)	"CDC" (27.00)	"NHS" (21.81)
	6	true (26.55)	infect (23.70)	article (30.21)	cost-01 (29.57)	above (15.82)	city (10.91)	right-05 (5.66)	information (5.98)	"Bush" (26.98)	"University" (21.73)
	7	lie (25.39)	possible (21.89)	they (29.98)	develop-02 (26.54)	today (14.25)	pandemic (10.82)	save-02 (5.15)	help-01 (5.82)	"Bill" (26.41)	"Pfizer" (21.67)
	8	free (24.79)	have (20.15)	amr-unknown (29.50)	patient (23.93)	url-entity (13.70)	early (10.57)	defend-01 (5.13)	address-02 (5.55)	"Russia" (23.84)	"Moderna" (20.98)
	9	read (23.66)	low (19.72)	corporation (29.00)	try-01 (23.70)	behind (12.70)	country (9.95)	benefit-01 (5.02)	meet-01 (5.04)	"Iraq" (22.77)	"measles" (20.54)
	10	corrupt (23.49)	treat (19.61)	military (28.80)	protein (23.00)	realm (10.38)	after (9.05)	experiment-01 (4.98)	improve-01 (4.71)	"Obama" (22.74)	"Wakefield" (20.49)
	11	kill (23.08)	test (19.28)	book (28.78)	vaccinate-01 (22.19)	article (10.28)	university (8.69)	cause-01 (4.88)	measles (4.56)	"Iran" (22.64)	"Medicare" (20.26)
	12	control (22.97)	access (19.01)	man (24.60)	antibody (21.78)	war (10.03)	so-far (8.47)	justify-01 (4.70)	immunize-01 (4.31)	"Wuhan" (22.23)	"Health" (20.12)
	13	cause (22.95)	research (18.82)	war-01 (24.60)	treat-03 (21.75)	then (9.94)	community (8.00)	live-01 (4.68)	assess-01 (4.21)	"CIA" (21.48)	"Oxford" (19.84)
	14	engineer (21.77)	work (18.47)	power (22.68)	infect-01 (21.70)	world (9.83)	globe (7.98)	version (4.67)	try-02 (4.05)	"Pharma" (21.01)	"AstraZeneca" (19.11)
	15	war (21.60)	manufacture (18.22)	citizen (22.56)	candidate (21.61)	about-to (9.52)	state (7.62)	counter-01 (4.66)	make-01 (4.00)	"War" (20.92)	"SARS-CoV-2" (18.22)

Table 2. Overrepresented elements.

FrameFinder: Explorative Multi-Perspective Framing Extraction from News Headlines

Markus Reiter-Haas
reiter-haas@tugraz.at
TU Graz, Austria

Beate Klösch
beate.kloesch@uni-graz.at
University of Graz, Austria

Markus Hadler
markus.hadler@uni-graz.at
University of Graz, Austria

Elisabeth Lex
elisabeth.lex@tugraz.at
TU Graz, Austria

ABSTRACT

Revealing the framing of news articles is an important yet neglected task in information seeking and retrieval. In the present work, we present FrameFinder, an open tool for extracting and analyzing frames in textual data. FrameFinder visually represents the frames of text from three perspectives, i.e., (i) frame labels, (ii) frame dimensions, and (iii) frame structure. By analyzing the well-established gun violence frame corpus, we demonstrate the merits of our proposed solution to support social science research and call for subsequent integration into information interactions.

CCS CONCEPTS

• **Information systems → Content analysis and feature selection**; World Wide Web; Language models; • **Computing methodologies** → *Information extraction.*

KEYWORDS

Computational Framing Extraction, Exploratory Content Analysis, Media Bias, Text Representations, Online News

ACM Reference Format:
Markus Reiter-Haas, Beate Klösch, Markus Hadler, and Elisabeth Lex. 2024. FrameFinder: Explorative Multi-Perspective Framing Extraction from News Headlines. In *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '24), March 10–14, 2024, Sheffield, United Kingdom.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3627508.3638308

1 INTRODUCTION

Cognitive biases, such as framing effects, influence information seeking and retrieval behaviors [3]. In this vein, biased search results have been shown to affect user attitudes due to exposure [7]. Moreover, it has been well established in psychology that framing also affects the behavior and choices of people [32]. Detecting and understanding the framing of online news is thus important due to its influence on readers, but also very challenging [21]. While there are several approaches for computational framing analysis (see [2] for an overview), many rely on annotated data and train a classifier. However, framing is defined as *the selection and salience of aspects in a communicating text* [8] and thus requires a deeper understanding than just doing predictions. Moreover, even in such supervised settings, the amount of available data is typically rather sparse. The sparsity issue was therefore one of the main challenges in the recent shared framing detection task at SemEval 2023 [22], where only a few or even zero samples were available per language. Notably, the best performing teams all used pretrained Transformers to tackle the task at hand [16, 25, 35]. Due to these challenges, the landscape of framing detection tools is still shallow, especially regarding openly available ones (e.g., [5]).

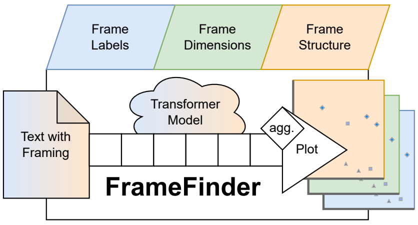


Figure 1: Schematic overview of the framing detection tool.

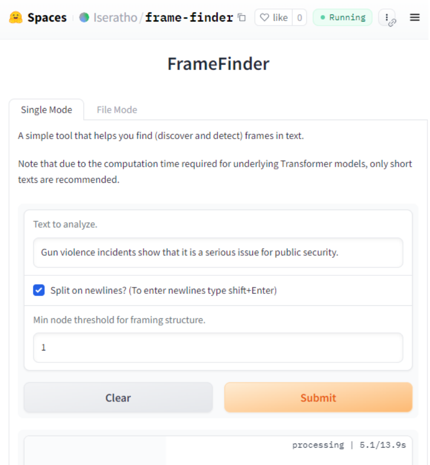
In the present work, we expand upon existing computational framing research by providing a novel tool to discover and extract frames from texts with a focus on online news. *FrameFinder* extracts frames from three distinct perspectives using Transformer models [33]. As described in [24], frames can be analyzed (i) by their associated *frame labels*, (ii) their *frame dimensions*, and (iii) their *frame structure*. To showcase the benefits of the tool, we conducted an analysis on the gun violence frame corpus (GVFC) [17]. There we find that the discussion is mostly framed regarding security rather than health, despite the names of involved people being a major structural element. Besides, the **openly available library and online demonstration**[1] allows both social science researchers and novice users to analyze the framing of texts without requiring technical (e.g., programming) skills. For future research, we strive to incorporate framing analyses directly into the retrieval process of online news to accomplish more balanced media consumption of users, either by informing them about the framing bias or by adapting, e.g., reranking, the retrieved results.

2 FRAMEFINDER: FRAMING DETECTION

Framing has multiple definitions across various scientific disciplines [31]. In this work, we consider communicative frames following Entman [8] regarding the *selection and salience of aspects in a communicating text to promote a specific interpretation.* As a result,

[1] The demo is available at: **https://huggingface.co/spaces/Iseratho/frame-finder** and accompanied by a brief video introduction: **https://iseratho.github.io/external/frame-finder-video.html** The underlying code is also available as a standalone Python library for full customization of algorithms and configuration: **https://github.com/Iseratho/framefinder** that can be installed via: *pip install framefinder*

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
CHIIR '24, March 10–14, 2024, Sheffield, United Kingdom
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0434-5/24/03.
https://doi.org/10.1145/3627508.3638308



A simple tool that helps you find (discover and detect) frames in text.

Note that due to the computation time required for underlying Transformer models, only short texts are recommended.

Text to analyze.

Gun violence incidents show that it is a serious issue for public security.

☑ Split on newlines? (To enter newlines type shift+Enter)

Min node threshold for framing structure.

1

```
processing | 5.1/13.9s
```

Figure 2: Truncated Screenshot of the Online Demo. For an overview of the generated plots, refer to experimental results in Figure 3.

framing deals with the presentation on both a micro and macro level [27]. Due to such nuances, framing is difficult to identify for algorithms [21]. Therefore, conceptualizations of framing are often only partially considered in automatic text processing [2].

FrameFinder is a tool to discover and extract frames from textual data using multiple distinct perspectives. As depicted in Figure 1, the tool takes texts as input that are (deliberately or undeliberately) framed in a certain way. The aim is to extract those frames in a human-comprehensible manner. To that end, we use the expressive power of Transformer models [33]. Internally, Transformers use embeddings, i.e., numerical vectors, to create rich representations of texts and parts thereof. The output representations, which can be probability vectors, alignment scores, or graph representations, are then aggregated and plotted. As previously identified in [24], we consider three distinct types of representations for framing analysis, i.e., frame labels, frame dimensions, and frame structure. For each type of representation, FrameFinder aggregates the result (when analyzing more than 1 sample) and visualizes them in a suitable format. Taken together, such a multi-perspective view of the data allows for a more nuanced framing analysis and the customizability of the library enables an explorative way to not only detect established but also discover novel frames.

Online Demonstration. For the online demonstration, we built the core part using HuggingFace Transformers [34] library and models. Together with Gradio [1], we deployed it as a HuggingFace space (see Figure 2), which also allows for discussions and feedback in the community tab. The demo runs on a CPU-only instance with 16 GB of RAM. The basic interface comprises two modes, a text-based and a file-based mode. The first allows entering example(s) in a text box, while the latter requires the upload of a text file. In both modes, the text is by default split on newlines into individual documents that are analyzed and aggregated. This option can be disabled to analyze the corpus as a single document (which is only recommended for short texts, as both the probability of frames being present and text structure tend to increase with text length). Additionally, there is a filtering option for the structural visualization based on node occurrence within graphs (i.e., the degree-weighted frequency across individual graphs). Finally, in the text-based mode, a few examples are provided that are cached (i.e., pre-computed) and thus evaluated instantly. For the deployed configuration (i.e., models and definitions of labels/dimensions) refer to the detailed description in Section 3 that was conducted with the same settings.

In the following, we describe the basic approaches of the three types of framing perspectives. Afterward, we discuss the relation of the tool to social science research.

2.1 Frame Labels

Framing detection can be approached as a classification task, in which specific *frame labels* are predicted to be either present or absent. This typically requires an annotated corpus. However, such corpora are scarce, with notable examples including the media frame corpus [6], the gun violence frame corpus [17], and the SemEval 2023 Task 3 Subtask 2 corpus [22]. Moreover, the number of samples within these corpora are typically rather small[2]. Alternatively, when given label definitions, the label prediction can also be modeled as a zero-shot prediction task.

Recent efforts to predict frame labels include contributions to the SemEval tasks (e.g., [16, 25, 35]) and the OpenFraming tool [5]. The latter differentiates between *frame discovery* using topic models and *frame prediction*, which involves training a classification model. Due to this explorative nature, it is similar in spirit to FrameFinder but requires expert knowledge and labor to annotate the data through content analysis. In contrast, we strive to avoid manual annotations, by considering multiple perspectives instead.

For aggregation of the prediction, we consider the mean and standard error of the label probabilities per sample. We then visualize the aggregated scores using a bar chart, and typically consider a threshold of 0.5 (denoted by color) to be indicative of which frame labels to assign to the corpus as a whole.

2.2 Frame Dimensions

Some frames are defined antagonistically, such as concerning moral foundations [11]. Considering the antagonistic care/harm pair, a text can be framed either positively emphasizing care (i.e., as a virtue) or negatively with harm in mind (i.e., as a vice), but not both. Such dimensions can be analyzed by considering the alignment within the embedding spaces of words and documents. Example approaches of dimensional framing analysis are moral framing in news [20], political framing on social media [13], or both, i.e., moral framing of political messages on social media [26].

The framing of documents can be analyzed either on a per-word basis using e.g. Word2Vec [19] or on a per-document basis using e.g. Sentence-Transformers [23]. In both scenarios, the position of an embedding (of a word or document) concerning the anchor

[2] The media frame corpus version 2 contains three subcorpora with 6327 on average but is deprecated due to changes in LexisNexis interface. The gun violence frame corpus contains 2990 samples, while the shared task in SemEval contains 2,049 split among train/dev/test set and nine languages (with three languages only in test).

embeddings (from the antagonistic pair) is determined. Herein, the FrameAxis method [14] scores the *frame bias* and intensity by projecting embeddings onto the axis formed by the antagonistic pair. The frame bias is defined as the mean of the scores, while the intensity considers the variance. Hence, the former specifies the leaning towards a frame, while the latter determines the activity along an axis. In the present work, we use FrameAxis for aggregating alignment scores but apply it to documents rather than words. The dimensions are plotted using horizontal lines, with the position of the projected points specifying the bias and their size specifying the intensity after aggregation.

2.3 Frame Structure

Some frames within a text are even more nuanced and require the consideration of the semantic structure. In this regard, the relations between the parts of text (e.g., words or phrases) are vital to extract the framing. One potential method for structural analysis is semantic role labeling (SRL) [10] that assigns tags that identify the type of argument in relation to a predicate. Two common examples of semantic roles are the *agent* tag, which is typically the subject, and *patients*, which are usually objects. An example approach for framing analysis is detailed in [13], where the agents and patients are visualized as tree stumps.

Alternatively, abstract meaning representations (AMR) [4] explicitly capture the semantic relations as rooted, directed, acyclic graphs[3]. In addition to extracting the semantic roles, these semantic graphs transform words and phrases into simplified semantic concepts, which improves comparability and subsequent transformations. Therefore, and in line with [24], we use AMR in the present work. When aggregating multiple semantic graphs, we create a weighted metagraph by superimposition of individual graphs. Thus, more pronounced concepts and relations get more emphasis, while additionally allowing filtering operations to only retain the most common elements of the metagraph.

2.4 Relation to Social Science Research

In the social sciences, such as sociology or communication studies, the analysis of frames also plays an important role, particularly in qualitative social research such as content analysis. Here, texts are typically coded manually, either deductively, i.e., on the basis of predetermined theoretical aspects [18], or inductively derived from the data material, as in the case of grounded theory [30]. FrameFinder works similarly to deductive content analysis by assigning predefined frames, i.e., frame labels (2.1) or moral dimensions (2.2), as codes to text passages. The detection of frame structures (2.3) is comparable to the basic principles of axial coding in the grounded theory approach, where identified codes and concepts, i.e., frames, are interpretatively contrasted and linked to each other. A tool like FrameFinder can help to get a first impression of the frames used in the text corpora and to decide on the further way of analysis. The frames found can then be integrated into MAXQDA [28] or other qualitative coding software for more in-depth analysis. However, social researchers need to consider the pre-defined labels and dimensions that underlie this tool in order to interpret and extend their manual analyses accordingly. The adoption of frame detection tools such as FrameFinder in social science research will depend on the choice of underlying framing concepts and their adaptability to various contexts and research goals.

GVFC Themes	# Events	# Issues
Total headlines	1269	1339
Economic consequences	3	92
Gun control/regulation	16	306
Gun/2nd Amendment rights	7	59
Mental health	51	29
Politics	32	401
Public opinion	18	244
Race/ethnicity	84	50
School or public space safety	28	156
Society/culture	4	44
Total labels	243	1381

Table 1: Statistics of the annotated GVFC.

3 DEMONSTRATION WITH THE GVFC

To demonstrate the merits of the framing extraction tool, we analyze the gun violence frame corpus (GVFC) [17]. The corpus consists of 2990 news headlines about gun violence in the United States. Figure 3 shows the results extracted with FrameFinder[4].

Models and Configuration. In the code, the models and their configuration can be adapted before computation. For the analysis of the GVFC, we use the same configuration (i.e., definitions of labels and dimensions), as well as models that are deployed in the online demonstration for consistency's sake. We choose three popular models, together with the well-established labels of the media frame corpus as labels and moral foundation theory as dimensions.

For the frame label extraction, we use a zero-shot classification model based on BART [15], i.e., *facebook/bart-large-mnli*. For zero-shot labels, we used the 14 specific media frames (and 1 unspecific other category) defined by their keyword list in [6].

For the frame dimensions, we use an encoder model based on MPNet [29], i.e., *sentence-transformers/all-mpnet-base-v2*. For the poles of the dimensions, we use the instructions from the moral foundation dictionary [9] (i.e., version two) of the five axes: harm/care, cheating/fairness, betrayal/loyalty, subversion/authority, and degradation/sanctity.

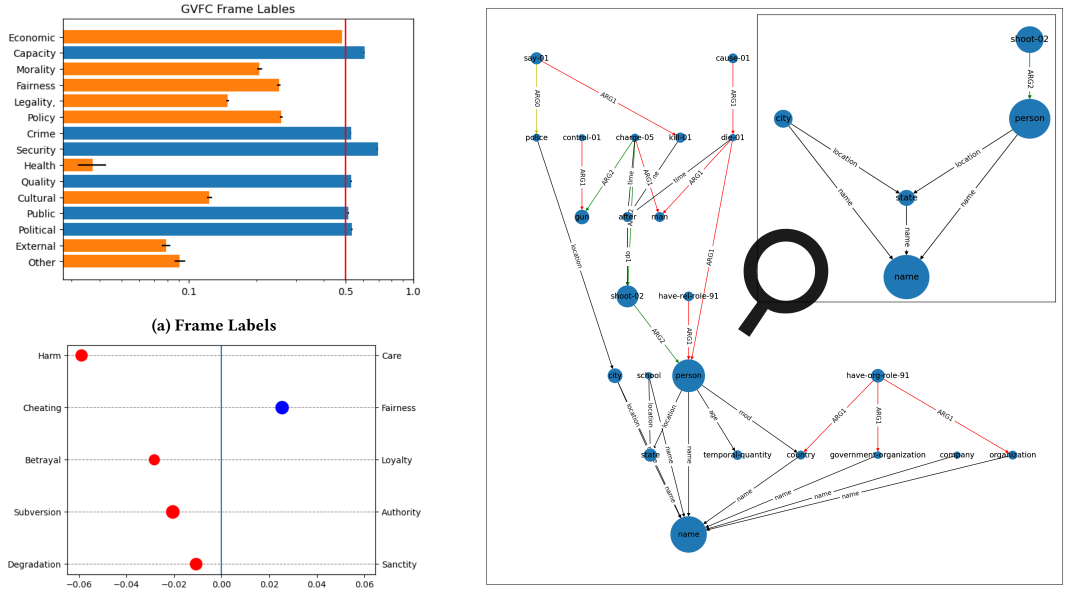
For the frame structure, we use another BART-based based model trained on abstract meaning representations (AMR) [4], i.e., *model_parse_xfm_bart_base-v0_1_0*[5]. We set the threshold for nodes to 300 and only plot the largest weakly connected component together with another zoomed-in version using a threshold of 1000.

Framing Analysis. From Figure 3a, we observe that gun violence headlines are mostly framed from a security viewpoint. Other important frames are about resources (i.e., capacity), crime, quality of life, public opinion, as well as political frames. In comparison, it is not seen as a health issue. Interestingly, there appears to be an

[3] For details of the node and edge types refer to the guidelines: `https://github.com/amrisi/amr-guidelines/blob/master/amr.md`

[4] Note that while the results were computed using the same underlying code, for efficiency, we extracted the frames using a GPU rather than using the free CPU-only online demo interface.

[5] The model can be downloaded from the AMRlib model GitHub repo: `https://github.com/bjascob/amrlib-models`



(a) Frame Labels

(b) Frame Dimensions

(c) Frame Structure (with zoomed-in substructure at the top right).

Figure 3: Framing visualizations of the GVFC.

absence of certain frames regarding morality and fairness, which can be investigated with the framing dimensions.

From a moral standpoint (see Figure 3b), it revolves most about the harm caused. Overall, the moral framing is rather negative with betrayal, subversion, and degradation residing on the vice side. In contrast, the fairness frame is the only positive (i.e., virtue) frame invoked in the headlines. While the bias of the frames differs noticeably (i.e., regarding their positions), the differences in intensity (i.e., point size) are much less pronounced. In this regard, the subversion/authority axis appears to be more emphasized compared to the betrayal/loyalty axis. While we clearly observe differences, with the overall negativity and fairness being less biased compared to harm, it shows that these moral values are of lesser concern when framing the news headlines.

Considering the structural view of the arguments (i.e., Figure 3c) shows that, while complex in nature, the headlines have a common theme. Specifically, as shown in the zoomed-in version, the headlines typically refer to the name of the victim of the shooting rather than the shooter (which is specified by the ARG2 role). Noteworthy is that guns and police have a subordinate role in the headlines.

To summarize, gun violence headlines frame the topic as a security issue that causes harm, with specific persons, such as the victims (mentioned by their names), being a focal point.

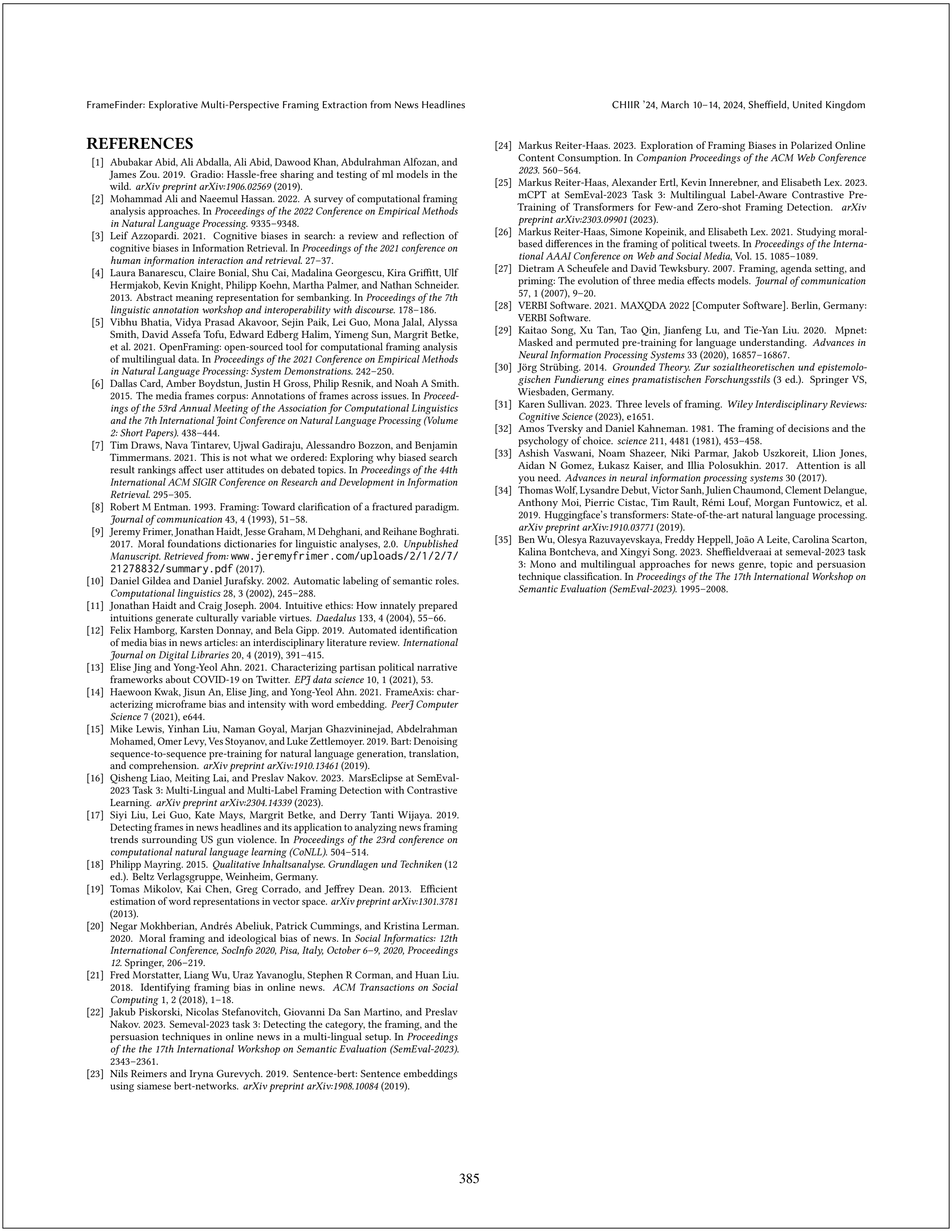
Comparison to GVFC Annotations. Here, we compare our results with the ground truth labels of the GVFC. In GVFC, headlines can either be assigned to singular events/incidents or issues of gun violence as an ongoing problem. Additionally, each headline gets assigned zero to two labels that determine the theme of the news story. We provide an aggregated overview in Table 1 (refer to [17] for further details). Both types (i.e., events and issues) appear roughly equally, but issues are far more often associated with labels.

This highlights a limitation of non-exploratory framing analysis, which involves first creating a codebook and then applying it to a corpus. Our use of FrameFinder reveals that the corpus often emphasizes the victims in event headlines. Similar to the annotations, we observe that politics and public opinion are common themes, while mental health gets neglected. In sum, while the annotations and findings from the exploratory framing analysis using FrameFinder largely align, the latter offers additional insights, e.g., emphasis on victims, which was not explicitly annotated in the corpus.

4 CONCLUSION

Framing analysis is intrinsically explorative and spans multiple disciplines. To advance research in this complex field, we present FrameFinder: an *explorative multi-perspective framing extraction* tool. Our user-friendly online demo offers insights into three distinct types of framing present in a text.

Currently, FrameFinder is designed to serve as a support tool for social science researchers. However, we recommend extending its application to information retrieval systems in future work. With media biases being a societal concern [12], we advocate for the development of more refined automatic models for media analysis.



REFERENCES

[1] Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. 2019. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569* (2019).

[2] Mohammad Ali and Naeemul Hassan. 2022. A survey of computational framing analysis approaches. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 9335–9348.

[3] Leif Azzopardi. 2021. Cognitive biases in search: a review and reflection of cognitive biases in Information Retrieval. In *Proceedings of the 2021 conference on human information interaction and retrieval*. 27–37.

[4] Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*. 178–186.

[5] Vibhu Bhatia, Vidya Prasad Akavoor, Sejin Paik, Lei Guo, Mona Jalal, Alyssa Smith, David Assefa Tofu, Edward Edberg Halim, Yimeng Sun, Margrit Betke, et al. 2021. OpenFraming: open-sourced tool for computational framing analysis of multilingual data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 242–250.

[6] Dallas Card, Amber Boydstun, Justin H Gross, Philip Resnik, and Noah A Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. 438–444.

[7] Tim Draws, Nava Tintarev, Ujwal Gadiraju, Alessandro Bozzon, and Benjamin Timmermans. 2021. This is not what we ordered: Exploring why biased search result rankings affect user attitudes on debated topics. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 295–305.

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

[9] Jeremy Frimer, Jonathan Haidt, Jesse Graham, M Dehghani, and Reihane Boghrati. 2017. Moral foundations dictionaries for linguistic analyses, 2.0. *Unpublished Manuscript. Retrieved from:* `www.jeremyfrimer.com/uploads/2/1/2/7/21278832/summary.pdf` (2017).

[10] Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational linguistics* 28, 3 (2002), 245–288.

[11] Jonathan Haidt and Craig Joseph. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus* 133, 4 (2004), 55–66.

[12] Felix Hamborg, Karsten Donnay, and Bela Gipp. 2019. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries* 20, 4 (2019), 391–415.

[13] Elise Jing and Yong-Yeol Ahn. 2021. Characterizing partisan political narrative frameworks about COVID-19 on Twitter. *EPJ data science* 10, 1 (2021), 53.

[14] Haewoon Kwak, Jisun An, Elise Jing, and Yong-Yeol Ahn. 2021. FrameAxis: characterizing microframe bias and intensity with word embedding. *PeerJ Computer Science* 7 (2021), e644.

[15] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* (2019).

[16] Qisheng Liao, Meiting Lai, and Preslav Nakov. 2023. MarsEclipse at SemEval-2023 Task 3: Multi-Lingual and Multi-Label Framing Detection with Contrastive Learning. *arXiv preprint arXiv:2304.14339* (2023).

[17] Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019. Detecting frames in news headlines and its application to analyzing news framing trends surrounding US gun violence. In *Proceedings of the 23rd conference on computational natural language learning (CoNLL)*. 504–514.

[18] Philipp Mayring. 2015. *Qualitative Inhaltsanalyse. Grundlagen und Techniken* (12 ed.). Beltz Verlagsgruppe, Weinheim, Germany.

[19] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[20] Negar Mokhberian, Andrés Abeliuk, Patrick Cummings, and Kristina Lerman. 2020. Moral framing and ideological bias of news. In *Social Informatics: 12th International Conference, SocInfo 2020, Pisa, Italy, October 6–9, 2020, Proceedings 12*. Springer, 206–219.

[21] Fred Morstatter, Liang Wu, Uraz Yavanoglu, Stephen R Corman, and Huan Liu. 2018. Identifying framing bias in online news. *ACM Transactions on Social Computing* 1, 2 (2018), 1–18.

[22] Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. Semeval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of the the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. 2343–2361.

[23] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[24] Markus Reiter-Haas. 2023. Exploration of Framing Biases in Polarized Online Content Consumption. In *Companion Proceedings of the ACM Web Conference 2023*. 560–564.

[25] Markus Reiter-Haas, Alexander Ertl, Kevin Innerebner, and Elisabeth Lex. 2023. mCPT at SemEval-2023 Task 3: Multilingual Label-Aware Contrastive Pre-Training of Transformers for Few-and Zero-shot Framing Detection. *arXiv preprint arXiv:2303.09901* (2023).

[26] Markus Reiter-Haas, Simone Kopeinik, and Elisabeth Lex. 2021. Studying moral-based differences in the framing of political tweets. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 15. 1085–1089.

[27] Dietram A Scheufele and David Tewksbury. 2007. Framing, agenda setting, and priming: The evolution of three media effects models. *Journal of communication* 57, 1 (2007), 9–20.

[28] VERBI Software. 2021. MAXQDA 2022 [Computer Software]. Berlin, Germany: VERBI Software.

[29] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems* 33 (2020), 16857–16867.

[30] Jörg Strübing. 2014. *Grounded Theory. Zur sozialtheoretischen und epistemologischen Fundierung eines pramatistischen Forschungsstils* (3 ed.). Springer VS, Wiesbaden, Germany.

[31] Karen Sullivan. 2023. Three levels of framing. *Wiley Interdisciplinary Reviews: Cognitive Science* (2023), e1651.

[32] Amos Tversky and Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *science* 211, 4481 (1981), 453–458.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[34] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).

[35] Ben Wu, Olesya Razuvayevskaya, Freddy Heppell, João A Leite, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2023. Sheffieldveraai at semeval-2023 task 3: Mono and multilingual approaches for news genre, topic and persuasion technique classification. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*. 1995–2008.

CEUR-WS.org/Vol-3660/paper18.pdf

The Framing Loop: Do Users Repeatedly Read Similar Framed News Online?

Markus Reiter-Haas[1,*], Elisabeth Lex[1]

[1] *Graz University of Technology, Institute of Interactive Systems and Data Science, 8010 Graz, Sandgasse 36/III*

Abstract

It is well established in psychology that framing of content affects the behavior of people. This effect is, however, only sparsely explored in information-seeking and retrieval behavior. In the present work, we consider the diversity of consumed content and repetition patterns regarding their framing. We conduct a framing analysis in the MIcrosoft News Dataset (MIND) comprising textual content and user interaction behaviors. By extracting the frames of the item sequences, we uncover a tendency of users to consume similar framed news repeatedly when sticking to the same type of content. Consequently, framing biases are important to consider in information systems. We hope that our work inspires future research on corresponding debiasing methods.

Keywords

Framing Theory, User Behavior, Empirical Study, Content Bias, Repeat Consumption, Viewpoint Diversity

1. Introduction

The effects of framing on peoples' choices have been well established in psychology and can be traced back to the notable work of Tversky and Kahneman [1]. While the grounding of framing effect as a cognitive bias is solid, research on its effects on information seeking and retrieval behavior has only recently emerged [2]. Besides this sparsely explored area resides a vast body of research on both framing theory (see [3] for an overview) and biases in online information systems ([4] provides an overview of biases in Web data) to draw from. Regarding framing theory, a wide variety of computational methods are available to extract the framing of content [5]. Whereas, for analyzing biased behavior patterns, several approaches have been studied for information systems, such as to understand repeat consumption [6] and assessing viewpoint diversity [7] regarding web searches. Hence, the conflation of the two research strands to expand the framing research in information systems seems reasonable.

In the present work, we investigate the content consumption regarding the framing in the MIcrosoft News Dataset (MIND) [8], which is well researched and sparked an influx for news recommendation research [9]. As depicted in Figure 1, each user has a sequential history of consumed items, as well as impressions and interactions for a specific timestamp. Additionally, each news item is assigned a specific category, which can be used to represent the sequence

Joint Proceedings of the ACM IUI Workshops 2024, March 18-21, 2024, Greenville, South Carolina, USA

[*] Corresponding author.

reiter-haas@tugraz.at (M. Reiter-Haas); elisabeth.lex@tugraz.at (E. Lex)
https://iseratho.github.io/ (M. Reiter-Haas); https://elisabethlex.info/ (E. Lex)
0000-0001-9852-8206 (M. Reiter-Haas); 0000-0001-5293-2967 (E. Lex)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

regarding consumed categories. Similarly, we can extract the framing based on the content and assign it to the items, which results in sequences of consumed frames. We consider such sequences to uncover biased behavioral patterns regarding the framing. For frame extraction, we use the FrameFinder library [10], which extracts three types of frames, i.e., media frames, moral frames, and semantic frames.

We find that frame consumption depends on the consumed categories, the types of frames and the information system itself. In particular, users repeatedly consume the same frames when sticking to the same category, which could be counteracted by the information system. Overall, the consumption behavior is more balanced concerning moral frames compared to semantic frames, whereas media frames depend on the categories the most.

In sum, our main contributions are:

1. We connect two separate strands of research in computer science (i.e., computational framing analysis and biases in information systems) that are both *rooted in psychology*.
2. We introduce an approach to analyze biased behavior patterns based on *sequences of consumed frames*.
3. We provide *empirical evidence of behavioral biases* due to framing on a well-established recommendation dataset.

To the best of our knowledge, we are the first to directly investigate this *link between the framing of content and the consumption behavior* of users. For reproducibility reasons, we additionally open source the code (also containing the supplementary materials referenced in the paper), as well as the framing dataset used for our study[1].

2. Related Work

Framing Theory: Framing has long been considered as a fractured paradigm in literature [12]. According to [3], there are three types of framing relating to language, cognition, and communication, respectively. While our study touches all three types, its focus lies on *communicative frames* present in media. Herein, framing as a form of bias in media has identified [13] and been thoroughly studied. For example, Morstatter et al. [14] train a classifier to detect the framing bias in news articles and relate it to opinion bias. This already indicates the relation to *cognitive frames*, which is an explicit requirement of communicative frames [3]. Finally, *semantic frames* were established by Fillmore and Baker [15] and depend on the language structure, but also on cognitive frames.

Recently, a vast amount of research uses computational methods for framing detection on wide range of frames, e.g., war [16], terrorists [17], morality [18], or blame [19] frames. The range of **computational framing analysis** approaches mainly span topic modeling and neural networks models (see Ali and Hassan [5] for a comprehensive survey). Neural networks are especially suitable in supervised settings, such as at the SemEval Challenge of 2023 [20], where every best-performing team used Transformer models [21, 22, 23]. Besides, open-source libraries like OpenFraming [24] and FrameFinder [10] support the extraction of frames. Our approach

[1] Code: https://github.com/Iseratho/frameloop Dataset: https://zenodo.org/records/10509498 [11]

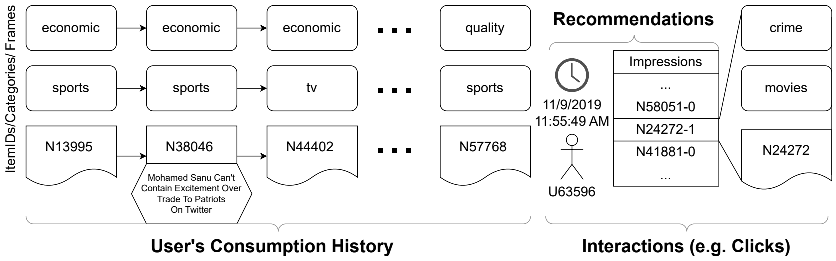


Figure 1: Real example from the dataset that shows how users interact with the system. Each user has a history of consumed items. Each item in the system contains textual content (depicted for the second item in the sequence), from which the frames can be extracted. At the top, the calculated media frame labels are represented. Here, the user shows a low viewpoint diversity regarding both categories and labels but even higher repeat consumption behavior regarding framing compared to categories. On the right, the user impression log is represented together with the clicked item (typically one). The impressions or clicked items can be seen as a continuation of the user history.

uses the latter to extract the framings present in news articles, as it also employs Transformer models [25] to extract frame representation in an unsupervised manner.

Biased Behavior Patterns: Cognitive psychology plays a vital role in information systems, which also provides the inspiration for various recommendation approaches [26]. As an example, a cognitive model of human memory (ACT-R) can predict music genre preferences [27]. Moreover, it has been shown that the cognitive-inspired ACT-R model also effectively predicts music relistening behavior [28], while also increasing the diversity of genres [29]. The relistening behavior is a type of **repeat consumption**, defined as "the act of consuming an enjoyable stimulus that one has already consumed in full in the past" in psychology [30]. Such biased repetition patterns have been found in a variety of domains and platforms, such as on Wikipedia, Google Maps, and YouTube [6]. Regarding diversity, assessing the viewpoints presented to users is another important bias in information systems to consider [7]. Herein, algorithmic diversification plays a key role in opinion forming domains, e.g., the news domain [31]. Moreover, the presence of distinct frames as a proxy for **viewpoint diversity** in news discourse is vital for high-quality debates [32].

In the present work, we investigate biased behavior patterns in news consumption sequences due to framing concerning both repeat consumption and viewpoint diversity with frame labels.

3. Problem Formulation and Notation

In an information system, a set of users U interacts with a set of items I. Each user $u \in U$ has a consumption history H_u, which consists of a sequence of n_h consumed items $i \in I$ by the user. For simplicity, we consider all user histories from the same specified time, thus omitting an additional time index ($i^H_{u,t,1} = i^H_{u,1}$). To access the information, a user might be presented with a list of n_r potential items $i \in I$ given by the function $\mathcal{R}(u, t)$. The function takes as input the

Notation	Description	# in *MIND-small*	$\overline{AVG}$
U	set of users, represented by their user IDs: $u \in U$	$\lvert U\rvert = 50,000$	$\lvert R\rvert/\lvert U\rvert = 3.14$
I	set of items, represented by their item IDs: $i \in I$	$\lvert I\rvert = 51,282$	$\overline{w} = 10.75$
H	set of user histories, $H = \bigcup_{u \in U} H_u$	$\lvert H\rvert = 49,108$	$\overline{H} = 18.52$
R	set of impression logs from the function $\mathcal{R}(u,t)$	$\lvert R\rvert = 156,965$	$\overline{R} = 37.23$
C	set of click logs from the function $\mathcal{C}(u,t)$	$\lvert C\rvert = 236,344$	$\overline{C} = 1.51$
L	set of label spaces; $l \in \bigcup_{L_j \in L_1, L_2, \ldots} L_j$; # category labels: $\lvert L_{Cat}\rvert = 17$		
$f_j(\cdot)$	mapping function for label space L_j from a set of mapping functions $f_j(\cdot) \in F$		
n_h, n_r, n_c	lengths of specific item set (i.e., logs of history, impression, and click, respectively)		
i_x^H, i_x^R, i_x^C	item lookup from logs (i.e., H, R, and C) with x providing required indices		

Table 1
Description of symbols used throughout the paper and their according statistics in *MIND-small*. $\overline{w}$ denotes the average number of words in the title.

user u and a specific time t. After evaluating the potential items in $\mathcal{R}(u,t)$, a user then interacts (i.e., consumes) one (or more) items of the list of potential items $i_c \in \mathcal{R}(u,t)$. This interaction can be formalized by the function $\mathcal{C}(u,t)$. The number of interacted items is denoted by n_c, which is $n_c = 1$ in most cases (i.e., where we can omit the positional index: $\mathcal{C}(u,t) = \{i_{u,t}^C\}$). The three described equations are thus given by (a summary of the main symbols is in Table 1):

$$
\begin{aligned}
H_u &= [i_{u,1}^H, i_{u,2}^H, \ldots, i_{u,n_h}^H] \\
\mathcal{R}(u,t) &= [i_{u,t,1}^R, i_{u,t,2}^R, \ldots, i_{u,t,n_r}^R] \\
\mathcal{C}(u,t) &= \{i_{u,t,1}^C, i_{u,t,2}^C, \ldots, i_{u,t,n_c}^C\}
\end{aligned}
$$

Each item i contains some content and can additionally be assigned some metadata, such as labels. For instance, we can assign a category label l to each item i based on its content $f_j(i) = l$, where $f_j(\cdot)$ is the mapping function from the content to the label space from a list of potential categories $l \in L_j$. Note that the system can have multiple label spaces $L = L_1, L_2, \ldots$, each with their corresponding mapping function. Consequently, we can transform the previous equations to the label space L_j for analysis, as shown in Figure 1:

$$
\begin{aligned}
H_u^{L_j} &= [f_j(i_{u,1}^H), f_j(i_{u,2}^H), \ldots, f_j(i_{u,n_h}^H)] &&= [l_{i_{u,1}^H,j}, l_{i_{u,2}^H,j}, \ldots, l_{i_{u,n_h}^H,j}] \\
\mathcal{R}^{L_j}(u,t) &= [f_j(i_{u,t,1}^R), f_j(i_{u,t,2}^R), \ldots, f_j(i_{u,t,n_r}^R)] &&= [l_{i_{u,t,1}^R,j}, l_{i_{u,t,2}^R,j}, \ldots, l_{i_{u,t,n_h}^R,j}] \\
\mathcal{C}^{L_j}(u,t) &= \{f_j(i_{u,t,1}^C), f_j(i_{u,t,2}^C), \ldots, f_j(i_{u,t,n_c}^C)\} &&= \{l_{i_{u,t,1}^C,j}, l_{i_{u,t,2}^C,j}, \ldots, l_{i_{u,t,n_h}^C,j}\}
\end{aligned}
$$

4. Data and Methods

We employ a two-step approach to identify biased behavior patterns regarding framing in the MIND dataset [8]. Specifically, we first construct sequences of labels (see Figure 1), which we then use to calculate four metrics on the sequence of categorical data for the behavior analysis. To ensure a fair comparison, we implement several simplifications on the data representation and evaluation setting (described below).

4.1. MIND Dataset

The MIND dataset [8] is a *large-scale dataset for news recommendation research* released in 2020, which follows the structure outlined in Figure 1. We use the smaller version *MIND-small*, which is a subset consisting of $50,000$ randomly sampled users and their associated data. The most important statistics of the dataset are provided in Table 1. The dataset has a high sparsity of $\frac{|H|\times\overline{H}+|R|\times\overline{R}}{|I|\times|U|} = 2.63 \times 10^{-3}$. In the dataset, each item (i.e., news article) consists of a single category that was manually assigned. Note that while a timestamp is available for the impression log, neither the individual interactions nor the sequential items in the history have been assigned any temporal data besides the order.

4.2. Constructing Label Sequences

We use a two-step procedure to construct the label sequences. First, we use metadata assigned to the items to construct sequences of categories. Second, we extract framing representations from the textual data (specifically the titles, as the short text is partially incomplete). Here, we employ the FrameFinder library [10], which allows the extraction of three distinct types, i.e., (i) media frames, (ii) moral frames, and (iii) semantic frames. Each representation uses a Transformer [25] model from Hugging Face [33] as a basis, where we use the default setting for all three types (details below). As these representations are not directly comparable, we simplify them by only considering the most pronounced feature per item and using that as a label.

Categories: For each item in the sequence (e.g., user history), we look up the category as there is always exactly one and assign it. Thus, the sequence is transformed into a sequence of labels. In *MIND-small*, there are 17 distinct labels, which are: ’lifestyle’, ’health’, ’news’, ’sports’, ’weather’, ’entertainment’, ’autos’, ’travel’, ’foodanddrink’, ’tv’, ’finance’, ’movies’, ’video’, ’music’, ’kids’, ’middleeast’, and ’northamerica’.

Media Frames: For the media frames: we use the *facebook/bart-large-mnli* model for zero-shot learning [34, 35, 36] with label definitions from the media frame corpus [37]. This model transforms the textual data to label probability scores, where we take the label with the maximum score. It is thus similar to the categories, but the labels are computed automatically rather than assigned manually. The set of 15 labels comprises: ’morality’, ’economic’, ’quality’, ’capacity’, ’crime’, ’security’,’health’, ’political’, ’public’, ’other’, ’cultural’, ’fairness’, ’policy’, ’legality,’, and ’external’.

Moral Frames: We use the *sentence-transformers/all-mpnet-base-v2* encoder model [38, 39] to extract the moral frames with the definitions derived from the moral foundation theory [40, 41]. Here, the textual data is transformed into alignment scores, which can be positive or negative, as each dimension is formed by an antagonistic label pair. Therefore, we take the maximum absolute value with a corresponding label (i.e., positive or negative, depending on the original sign). This forms a set of 10 labels: ’authority’, ’cheating’, ’subversion’, ’degredation’, ’harm’, ’fairness’, ’care’, ’betrayal’, ’loyalty’, and ’sanctity’.

Semantic Frames: The model *Iseratho/model_parse_xfm_bart_base-v0_1_0*, which is a copy on Hugging Face of an AMRLib[2] model. The model is based on BART using abstract meaning

[2] https://amrlib.readthedocs.io/en/latest/

Name	Example sequence	DRR	$RRdist$	$Uniq$	$Gini$
Specific	$[a, b, a, b, b, c]$	0.2	0.5	0.4 if $\lvert L_j \rvert \geq 6$	$0.6\dot{1}$
All same	$[a, a, \ldots, a]$	1.0	1.0	0.0	0.0
Alternating	$[a, b, a, b, \ldots, a, b]$	0.0	0.5	$1/(\lvert L_j \rvert - 1)$	0.5
All different	$[a, b, c, \ldots, j]$	0.0	0.0	1.0	$lim_{n \to \infty}$ = 1
Encased	$[a, b, b, \ldots, b, a]$	$lim_{n \to \infty}$ = 1	0.5	$1/(\lvert L_j \rvert - 1)$	$lim_{n \to \infty}$ = 0
Random	$[rng(L_j), rng(L_j),$ $\ldots, rng(L_j)]$	$lim_{n \to \infty} \mathbb{E}[S]$ $= 1/\lvert L_j \rvert$	$lim_{n \to \infty} \mathbb{E}[S]$ $= 1/\lvert L_j \rvert$	$lim_{n \to \infty} \mathbb{E}[S]$ $= 1$	$lim_{n \to \infty} \mathbb{E}[S]$ $= 1-(1/\lvert L_j \rvert)$

Table 2
Metrics on example sequences DRR and $RRdist$ tend to behave mostly opposite to $Uniq$ and $Gini$.

representations [42, 34] that transforms texts to semantic graphs comprising semantic frames[3]. From the semantic graphs, we extract the most pronounced frames. Due to the large size of the label space, we only consider frames that appear at least 200 times at the root (i.e., the most pronounced position). The resulting set contains 23 frames: 'say-01', 'possible-01', 'report-01', 'cause-01', 'die-01', 'find-01', 'have-degree-91', 'watch-01', 'contrast-01', 'get-01', 'arrest-01', 'be-located-at-91', 'charge-05', 'open-01', 'show-01', 'kill-01', 'have-03', 'reveal-01', 'recommend-01', 'announce-01', 'want-01', 'close-01', and 'win-01'. We then use the first frame of the set in the serialized form of the graph. If none of the frames are present, we insert a special 'other' frame (similar to how the media frames have an 'other' label) instead.

4.3. Behavior Sequence Analysis

For the behavior analysis, we use two metrics each (one coarse- and one fine-grained) as a proxy to measure repeat consumption behavior and viewpoint diversity, respectively. All metrics are normalized to fall in the range of $[0, 1]$. For repeat consumption behavior, a high value means that the same items are repeatedly consumed and thus indicate a less balanced consumption pattern. For viewpoint diversity, a high value means more diversity in consumed items and thus indicates a balanced consumption diet. For repeat consumption metrics, the sequence order is relevant while the label distribution is secondary, whereas for viewpoint diversity metrics, the sequential orderings are irrelevant.

The metrics are defined to work on arbitrary sequences S containing categorical data. In the most basic case, we evaluate the sequence of a user's history of a particular label space, i.e., $S = H_u^{L_j}$. For simplicity, we omit the details of the indices besides the positional index (i.e., $[l_{i^H_{u,1},j}, l_{i^H_{u,2},j}, \ldots, l_{i^H_{u,n_h},j}]$ becomes $[l_1, l_2, \ldots, l_n]$). Besides, we use $\mathbb{1}_{condition}$ as the indicator function, which returns 1 if the $condition$ is true and 0 otherwise.

Direct Repetition Ratio (DRR) measures the ratio of sequential item pairs having the same labels.

$$DRR(S) = 1/(n-1) \sum_{i=1}^{n-1} \mathbb{1}_{l_i = l_{i+1}} \tag{1}$$

[3] The representation also contains additional data beyond the scope of this work. The list of frames is available at: https://propbank.github.io/v3.4.0/frames/

When considering the example sequences in Table 2, we observe the in the *specific* example sequence one out of five sequential pairs is a direct repetition (i.e., $1/5$). Note that higher order patterns (e.g., *alternating* sequences) do not impact the value. Therefore, singular outliers (e.g., in the *encased* sequence) will only marginally affect the value. The convergence behavior of *random* sequences depends on the size of the label space.

Reciprocal Repeat Distance (RRdist) measures the average distance between neighboring repetitions (i.e., same labels while every label between them is different) and is normalized by the reciprocal value. Therefore, it can be seen as a sort of probability score that labels are repeated.

$$RRdist(S) = \frac{\sum_{i=1}^{n-1}\sum_{j=2}^{n} \mathbb{1}_{i<j,l_i=l_j \wedge l_i \neq l_k, \forall k, i<k<j}}{\sum_{i=1}^{n-1}\sum_{j=2}^{n}(j-i)\mathbb{1}_{i<j,l_i=l_j \wedge l_i \neq l_k, \forall k, i<k<j}} \tag{2}$$

Concerning the *specific* example in Table 2, a has a distances of one and two, while b has a distance of three, which results in an average distance of two (i.e., reciprocal value of 0.5). Note that metric capture higher order patterns, such as both the *alternating* and *encased* sequence having a distance of 0.5. In the former case, the distance is always two, while in the latter case, $n-3$ times a distance of one and one time a distance of $n-1$ resulting of $n-2$ repetition events (i.e., $\frac{n-2}{(n-3)*1+1*(n-1)}$). Similar to DDR, the limit of a *random* sequence approaches the reciprocal value of the label space.

Uniqueness Index (Uniq) determines how much of unique labels are present compared to the theoretical maximum. The maximum depends on the sequence length and label space and is bounded by whichever is smaller. Therefore, if $|L_j| < n$, then the maximum is reached when all labels are present, whereas if $n < |L_j|$, the maximum is reached when *all* labels are *different*.

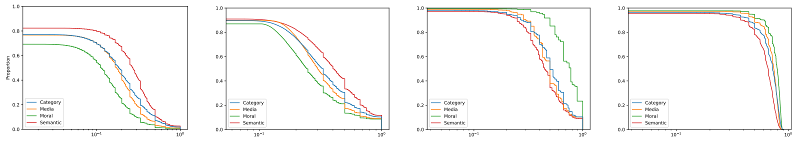
$$Uniq(S) = \frac{|\{S\}| - 1}{min(|S|, |L_j|) - 1} \tag{3}$$

In Table 2, the *specific* sequence is $(3-1)/(6-1)$ as three of potentially six labels are present. If *all* items are the *same*, then the minimum of zero is reached (which is why one is deducted from both the enumerator and denominator). The value tends towards one for long *random* sequences. Therefore, the metric is a form of coverage on the sequence level rather than system level.

Distribution Imbalance (Gini) uses the Gini index, which considers the probabilities of label occurrence. Therefore, uniform distribution lead to higher values than skewed distributions.

$$Gini(S) = 1 - \sum_{l \in L_j}(p_l)^2, \quad p_l = \frac{1}{|S|}\sum_{i=1}^{n}\mathbb{1}_{l_i=l} \tag{4}$$

The *specific* example of Table 2 is thus the result of $1-((1/6)^2+(2/6)^2+(3/6)^2)$. Gini is 0 with *all same* sequence, has 0.5 with two labels equally distributed (e.g., *alternating* sequence), and tends towards 1 as long sequence of *all different* labels. Similar to DDR, singular outliers do



(a) DRR ECDF Plot (b) RRdist ECDF (c) Uniq ECDF (d) Gini ECDF

Figure 2: Complementary empirical cumulative distribution function (ECDF) plots on a log scale (x-axis) on S_H. Category = blue, Media Frames = orange, Moral Frames = green, Semantic Frames = red.

not significantly affect the outcome on long sequences (e.g., consider *encased*). For long *random* sequences, the value depends on the size of the label space.

5. Experiments

We want to answer the following research questions by analyzing their corresponding label sequences (denoted by $\rightarrow$):

RQ1: **How is the repeat consumption behavior and viewpoint diversity of frames compared to categories?**
$\rightarrow S_H = H$: the set of user histories; also used for comparison in RQ2 and RQ3.

RQ2: **What is the interplay between frames and categories?**
Whether more of the same frames are consumed in per-category sub-sequences?
$\rightarrow S_{H/L_{Cat}}$: the subsets from the user histories per category

RQ3: **What are the effects of framing with regard to (a) retrieved, i.e., with impressions,**
($\rightarrow S_{H \oplus R}$: the user history enhanced with a single impression)
and (b) consumed, i.e., with clicked, content
($\rightarrow S_{H \oplus C}$: the user history enhanced with a single click)?

RQ1: Comparison of Framing Behavior

Concerning the user history S_H, we observe that categories and media frames are closely related (Table 3 and Figure 2), which can be the result of the set of media frames being defined in terms of topics (for which they were already criticized [5]). On the other hand, moral and semantic frames deviate notably and have the opposite tendency towards each other. Users show a low repeat consumption behavior (according to DRR and $RRdist$) in terms of moral frames and high viewpoint diversity (according to $Uniq$ and $Gini$). The effect is most pronounced for the uniqueness index, which becomes visually apparent in Figure 2c. Concerning the overall distribution of values (Figure 2), repeat behavior metrics are lower for all label types compared to viewpoint diversity. In fact, around 20% of sequences do not have any direct repetitions (left starting point in Figure 2a), and around 10% of sequences have all different labels, which results in an $RRdist = 0$ indicated in Figure 2b. Herein, the results first increase much quicker for DRR, whereas for $RRdist$, there is a noticeable jump at the end to the value of 1. For $Gini$ the values appear clustered around a high value close to 1 without actually reaching it (Figure 2d).

	Categories (L_{Cat})	Media Frames (L_{Media})	Moral Frames (L_{Moral})	Semantic Frames (L_{Sem})
User History (RQ1, $\|S_H\| = 49,108$, $\overline{S_H} = 18.85$)				
DRR	0.2194 ± 0.21	0.1908 ± 0.17	$\underline{0.1336} \pm 0.14$	0.2861 ± 0.22
$RRdist$	0.3880 ± 0.28	0.3641 ± 0.26	0.3164 ± 0.26	0.4532 ± 0.29
$Uniq$	0.5453 ± 0.23	0.5288 ± 0.21	0.7459 ± 0.20	**0.4883** ± 0.23
$Gini$	0.6715 ± 0.19	0.6956 ± 0.16	0.7464 ± 0.16	0.6115 ± 0.18
Per Category (RQ2, $\|S_{H/L_{Cat}}\| = 687,054$, $\overline{S_{H/L_{Cat}}} = 6.84$)				
DRR	-	$\underline{0.2990} \pm 0.35 \uparrow$	$0.1591 \pm 0.26 \uparrow$	**0.3217** $\pm 0.35 \uparrow$
$RRdist$	-	$\underline{0.4703} \pm 0.40 \uparrow$	$0.3354 \pm 0.37 \uparrow$	**0.4979** $\pm 0.40 \uparrow$
$Uniq$	-	$0.5523 \pm 0.35 \uparrow$	$0.7385 \pm 0.28 \downarrow$	$0.5454 \pm 0.35 \uparrow$
$Gini$	-	$\underline{0.5058} \pm 0.24 \downarrow$	$0.6051 \pm 0.19 \downarrow$	**0.4886** $\pm 0.23 \downarrow$
With Impressions (RQ3a, $\|S_{H \oplus R}\| = 5,723,002$, $\overline{S_{H \oplus R}} = 37.26$)				
DRR	0.2182 ± 0.16	$0.1959 \pm 0.13 \upuparrows$	**0.1331** ± 0.11	$0.2978 \pm 0.17 \upuparrows$
$RRdist$	$0.3126 \pm 0.23 \downarrow$	$0.3007 \pm 0.20 \downarrow$	$\underline{0.2533} \pm 0.20 \downarrow$	$0.3813 \pm 0.24 \downarrow$
$Uniq$	$0.5563 \pm 0.19 \uparrow$	$0.5075 \pm 0.17 \downdownarrows$	$\underline{0.8077} \pm 0.18 \uparrow$	$\underline{0.4902} \pm 0.19$
$Gini$	$0.7291 \pm 0.13 \upuparrows$	$0.7473 \pm 0.10 \uparrow$	$\underline{0.8061} \pm 0.09 \uparrow$	$0.6515 \pm 0.14 \uparrow$
With Clicks (RQ3b, $\|S_{H \oplus C}\| = 231,530$, $\overline{S_{H \oplus C}} = 41.09$)				
DRR	$0.2227 \pm 0.17 \uparrow$	$0.1946 \pm 0.13 \uparrow$	$\underline{0.1336} \pm 0.10$	$0.2914 \pm 0.16 \uparrow$
$RRdist$	$0.3093 \pm 0.22 \downdownarrows$	$0.2942 \pm 0.20 \downdownarrows$	**0.2490** $\pm 0.20 \downdownarrows$	$0.3674 \pm 0.25 \downdownarrows$
$Uniq$	$0.5580 \pm 0.19 \upuparrows$	$0.5128 \pm 0.17 \downarrow$	**0.8113** $\pm 0.18 \upuparrows$	$0.5080 \pm 0.20 \upuparrows$
$Gini$	$0.7253 \pm 0.14 \uparrow$	$0.7497 \pm 0.10 \upuparrows$	**0.8066** $\pm 0.09 \upuparrows$	$0.6597 \pm 0.13 \upuparrows$

Table 3
Mean metrics of sequences with standard deviation ($\pm$). $\uparrow$ indicates a statistically significant increase ($p < 0.0005$ according to a t-test) in metric compared to user history sequences S_H, while $\downarrow$ indicates the opposite direction. In case that both impressions and clicks have the same effect on the direction, we denote the stronger effect with a double arrow (i.e., $\upuparrows$ or $\downdownarrows$). The overall highest and lowest values per metric are highlighted in **bold**, while the second highest/lowest are underlined. For each set of sequences, we denote the amount and average length.

In sum, the repeat consumption and viewpoint diversity are frame-specific. Moral frames appear to be consumed in a more balanced way compared to semantic frames. Furthermore, categories and media frames seem to be closely related in terms of consumption behavior. Therefore, we investigate this relation in RQ2.

RQ2: Relation between Categories and Framing

All three types of frames are correlated with the categories on all four metrics (plots are provided in the code repository). The consideration of the subsequences per category ($S_{H/L_{Cat}}$) leads to statistically significant changes in all metrics and frames (Table 3). Specifically, the repeat consumption always increases (both DRR and $RRdist$), while $Gini$ always decreases. In fact, this results in the highest (bold in semantic frames) and second highest (underlined in media frames) values overall in terms of repeat consumption and similarly the lowest and second lowest for $Gini$. Therefore, the consumption behavior appears less balanced when considering individual categories. In other words, a balanced consumption behavior regarding framing appears to be partially the result of a more diverse set of categories consumed. Interestingly, the $Uniq$, while still affected, does not show such a tendency. Moreover, it even increases for media and semantic frames, thus indicating a still broad range of frames in these shorter sequences.

Overall, we can conclude that categories play a vital role in the consumption behavior of frames, as the same frames are consumed even more repeatedly. As information systems are also prone to narrow the content shown to users [43], e.g., by repeatedly recommending similar items in terms of categories, we investigate these effects more closely in RQ3.

RQ3: Framing Effects in Information Systems
To start, we investigate whether shown and click items are a mere repetition of the last item's label in the user history (i.e., whether DRR increases in Table 3). Apparently, the last category is not used to determine the shown items, while the users themselves, more often than not, stick to the same category. Here, user intent might play a role (see [44] for an example of intent modeling in sequential recommendation), which is beyond the scope of the current study. Nevertheless, the system seems to repeat the media and semantic frames, which also affects the user click behavior. The effect is more pronounced in $S_{H \oplus R}$ compared to $S_{H \oplus C}$, which might indicate that the system is the source of the bias rather than the users themselves. Interestingly, moral frames do not seem that affected (no statistically significant change of $p < 0.0005$) and stay low (being the lowest values of DRR overall). In comparison, $RRDist$ decreases for both sets of sequences, while viewpoint diversity tends to increase. This effect is most pronounced regarding the moral frames, especially on the click behavior. In general, the click behavior is more affected regarding $RRdist$, $Uniq$, and $Gini$. One outlier here is the uniqueness of media frames, which decreases and is more pronounced in the impressions rather than click behavior.

The results suggest that, although information systems tend to promote sticking to the same type of content, the effects on consumption behavior might be a net positive, as users could be supported in balancing their media consumption. Please note that the current study cannot deduce long-term effects and therefore urges for future work.

6. Conclusion

In the present, study we relate the framing of content to consumption behavior in information systems. Herein, we investigate the repeat consumption behavior and viewpoint diversity for three types of frames (i.e., media, moral, and semantic frames). Our findings suggest the relation to behavior is different per frame type, with media frames closely following categories. The repetition of frames also increases when investigating the categories separately, whereas the diversity tends to increase due to the effects of information systems.

Our study has broad implications for the design of information systems, as it suggests considering user behavior within particular types of content rather than diversifying through recommending a broad spectrum of types.

Limitations. Our study has two main limitations. First, the scope of the study is narrow, as we consider only a single dataset in the news domain, which was designed for recommendation research, with three specific models. Second, we performed a simplified analysis for better comparison, which omitted fine-grained details in content (e.g., the graph structure of semantic frames), metrics (e.g., the influence due to number of labels), and behavior (e.g., user intent).

Future Work. We hope our work sparks interest in considering framing as a form of bias in information systems. Most of all, we call for the development of debiasing methods concerning user behavior due to framing. Specifically, we see personalized user interfaces that support a balanced consumption diet, e.g., through transparency, as a promising research direction for future work.

References

[1] A. Tversky, D. Kahneman, The framing of decisions and the psychology of choice, science 211 (1981) 453–458.

[2] L. Azzopardi, Cognitive biases in search: a review and reflection of cognitive biases in information retrieval, in: Proceedings of the 2021 conference on human information interaction and retrieval, 2021, pp. 27–37.

[3] K. Sullivan, Three levels of framing, Wiley Interdisciplinary Reviews: Cognitive Science (2023) e1651.

[4] R. Baeza-Yates, Bias on the web, Communications of the ACM 61 (2018) 54–61.

[5] M. Ali, N. Hassan, A survey of computational framing analysis approaches, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 9335–9348.

[6] A. Anderson, R. Kumar, A. Tomkins, S. Vassilvitskii, The dynamics of repeat consumption, in: Proceedings of the 23rd international conference on World wide web, 2014, pp. 419–430.

[7] T. Draws, N. Tintarev, U. Gadiraju, Assessing viewpoint diversity in search results using ranking fairness metrics, ACM SIGKDD Explorations Newsletter 23 (2021) 50–58.

[8] F. Wu, Y. Qiao, J.-H. Chen, C. Wu, T. Qi, J. Lian, D. Liu, X. Xie, J. Gao, W. Wu, et al., Mind: A large-scale dataset for news recommendation, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 3597–3606.

[9] S. Raza, C. Ding, News recommender system: a review of recent progress, challenges, and opportunities, Artificial Intelligence Review (2022) 1–52.

[10] M. Reiter-Haas, B. Klösch, M. Hadler, E. Lex, Framefinder: Explorative multi-perspective framing extraction from news headlines, arXiv preprint arXiv:2312.08995 (2023).

[11] M. Reiter-Haas, Mind-small frames, 2024. URL: https://doi.org/10.5281/zenodo.10509498. doi:10.5281/zenodo.10509498.

[12] R. M. Entman, Framing: Toward clarification of a fractured paradigm, Journal of communication 43 (1993) 51–58.

[13] R. M. Entman, Framing bias: Media in the distribution of power, Journal of communication 57 (2007) 163–173.

[14] F. Morstatter, L. Wu, U. Yavanoglu, S. R. Corman, H. Liu, Identifying framing bias in online news, ACM Transactions on Social Computing 1 (2018) 1–18.

[15] C. J. Fillmore, C. F. Baker, Frame semantics for text understanding, in: Proceedings of WordNet and Other Lexical Resources Workshop, NAACL, volume 6, 2001.

[16] P. Wicke, M. M. Bolognesi, Framing covid-19: How we conceptualize and discuss the pandemic on twitter, PloS one 15 (2020) e0240010.

[17] D. Demszky, N. Garg, R. Voigt, J. Zou, M. Gentzkow, J. Shapiro, D. Jurafsky, Analyzing

polarization in social media: Method and application to tweets on 21 mass shootings, arXiv preprint arXiv:1904.01596 (2019).

[18] M. Reiter-Haas, S. Kopeinik, E. Lex, Studying moral-based differences in the framing of political tweets, arXiv preprint arXiv:2103.11853 (2021).

[19] C. Shurafa, K. Darwish, W. Zaghouani, Political framing: Us covid19 blame game, in: International Conference on Social Informatics, Springer, 2020, pp. 333–351.

[20] J. Piskorski, N. Stefanovitch, G. Da San Martino, P. Nakov, Semeval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup, in: Proceedings of the the 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023, pp. 2343–2361.

[21] Q. Liao, M. Lai, P. Nakov, Marseclipse at semeval-2023 task 3: Multi-lingual and multi-label framing detection with contrastive learning, arXiv preprint arXiv:2304.14339 (2023).

[22] M. Reiter-Haas, A. Ertl, K. Innerebner, E. Lex, mcpt at semeval-2023 task 3: Multilingual label-aware contrastive pre-training of transformers for few-and zero-shot framing detection, arXiv preprint arXiv:2303.09901 (2023).

[23] B. Wu, O. Razuvayevskaya, F. Heppell, J. A. Leite, C. Scarton, K. Bontcheva, X. Song, Sheffieldveraai at semeval-2023 task 3: Mono and multilingual approaches for news genre, topic and persuasion technique classification, in: Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023, pp. 1995–2008.

[24] V. Bhatia, V. P. Akavoor, S. Paik, L. Guo, M. Jalal, A. Smith, D. A. Tofu, E. E. Halim, Y. Sun, M. Betke, et al., Openframing: open-sourced tool for computational framing analysis of multilingual data, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2021, pp. 242–250.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[26] E. Lex, D. Kowald, P. Seitlinger, T. N. T. Tran, A. Felfernig, M. Schedl, et al., Psychology-informed recommender systems, Foundations and Trends® in Information Retrieval 15 (2021) 134–242.

[27] D. Kowald, M. Reiter-Haas, S. Kopeinik, M. Schedl, E. Lex, Transparent music preference modeling and recommendation with a model of human memory theory (2024).

[28] M. Reiter-Haas, E. Parada-Cabaleiro, M. Schedl, E. Motamedi, M. Tkalcic, E. Lex, Predicting music relistening behavior using the act-r framework, in: Proceedings of the 15th ACM Conference on Recommender Systems, 2021, pp. 702–707.

[29] M. Moscati, C. Wallmann, M. Reiter-Haas, D. Kowald, E. Lex, M. Schedl, Integrating the act-r framework with collaborative filtering for explainable sequential music recommendation, in: Proceedings of the 17th ACM Conference on Recommender Systems, 2023, pp. 840–847.

[30] E. O'Brien, A mind stretched: The psychology of repeat consumption, Consumer Psychology Review 4 (2021) 42–58.

[31] N. Tintarev, E. Sullivan, D. Guldin, S. Qiu, D. Odjik, Same, same, but different: algorithmic diversification of viewpoints in news, in: Adjunct publication of the 26th conference on user modeling, adaptation and personalization, 2018, pp. 7–13.

[32] C. Baden, N. Springer, Conceptualizing viewpoint diversity in news discourse, Journalism 18 (2017) 176–194.

[33] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, 2020, pp. 38–45.

[34] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, arXiv preprint arXiv:1910.13461 (2019).

[35] A. Williams, N. Nangia, S. Bowman, A broad-coverage challenge corpus for sentence understanding through inference, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, pp. 1112–1122.

[36] W. Yin, J. Hay, D. Roth, Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach, in: 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Association for Computational Linguistics, 2019, pp. 3914–3923.

[37] D. Card, A. Boydstun, J. H. Gross, P. Resnik, N. A. Smith, The media frames corpus: Annotations of frames across issues, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2015, pp. 438–444.

[38] K. Song, X. Tan, T. Qin, J. Lu, T.-Y. Liu, Mpnet: Masked and permuted pre-training for language understanding, Advances in Neural Information Processing Systems 33 (2020) 16857–16867.

[39] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, 2019, pp. 3982—-3992.

[40] J. Frimer, J. Haidt, J. Graham, M. Dehghani, R. Boghrati, Moral foundations dictionaries for linguistic analyses, 2.0, Unpublished Manuscript. Retrieved from: www. jeremyfrimer. com/uploads/2/1/2/7/21278832/summary. pdf (2017).

[41] J. Graham, J. Haidt, B. A. Nosek, Liberals and conservatives rely on different sets of moral foundations., Journal of personality and social psychology 96 (2009) 1029.

[42] L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, N. Schneider, Abstract meaning representation for sembanking, in: Proceedings of the 7th linguistic annotation workshop and interoperability with discourse, 2013, pp. 178–186.

[43] T. T. Nguyen, P.-M. Hui, F. M. Harper, L. Terveen, J. A. Konstan, Exploring the filter bubble: the effect of using recommender systems on content diversity, in: Proceedings of the 23rd international conference on World wide web, 2014, pp. 677–686.

[44] Y. Chen, Z. Liu, J. Li, J. McAuley, C. Xiong, Intent contrastive learning for sequential recommendation, in: Proceedings of the ACM Web Conference 2022, 2022, pp. 2172–2182.

Computational Narrative Framing: Towards Identifying Frames through Contrasting the Evolution of Narrations

Markus Reiter-Haas[1,*], Beate Klösch[2], Markus Hadler[2] and Elisabeth Lex[1]

[1]Graz University of Technology, Institute of Interactive Systems and Data Science, 8010 Graz, Sandgasse 36/III
[2]University of Graz, Department of Sociology, 8010 Graz, Universitätsstraße 15/G4

Abstract

Our understanding of the world is fundamentally shaped by language, with narrations being a central point, and influenced by its framing. Recent advancements in language models gave rise to computational methods for both narrative understanding and framing analysis. Although given their overlap, these two strands are mostly researched independently. In this position paper, we argue for their consolidation in the form of narrative framing, i.e., the framing process driven by narrations. Herein, we outline similarities between both based on semantic elements. Besides, we discuss how different narratives might compete with each other, as well as evolve over time. Thereby, narratives inevitably change the framing, exemplarily depicted on the issue of climate change. We believe that the analysis of narrative frames will lead to a broader understanding of textual corpora as a whole rather than individual pieces of text.

Keywords

Framing Theory, Narrative Frames, Competing Narrations, Climate Change Framing, Semantic Graphs

1. Introduction

Experiences in the real-world and narrative perception are inextricably linked in humans, even on a neurological level [1]. In a similar vein, the framing of narratives can act as a device to blend fiction and reality [2], consequently suggesting certain solutions to specific problems [3] and affect the people's choices [4]. Unlike other types of frames, the pool of options concerning narratives for framing is essentially endless. Although some works on computationally extracting narrative framing have already been conducted [e.g., 5, 6, 7], the still sparse body of research tends to favor one strand of research, i.e., either narrative understanding or framing analysis, over the other.

In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia, M. Litvak (eds.): Proceedings of the Text2Story'24 Workshop, Glasgow (Scotland), 24-March-2024

[*]Corresponding author.

reiter-haas@tugraz.at (M. Reiter-Haas); beate.kloesch@uni-graz.at (B. Klösch); markus.hadler@uni-graz.at (M. Hadler); elisabeth.lex@tugraz.at (E. Lex)
https://iseratho.github.io/ (M. Reiter-Haas); https://elisabethlex.info/ (E. Lex)
0000-0001-9852-8206 (M. Reiter-Haas); 0000-0002-8061-6088 (B. Klösch); 0000-0002-0359-5789 (M. Hadler); 0000-0001-5293-2967 (E. Lex)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3671/paper11.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

In this position paper, we present a basic theoretical framework for computational narrative framing analysis, effectively combining computational narrative understanding and computational framing analysis research. We identify the commonalities between the two strands to form an elementary understanding of the necessities for emerging approaches in this direction. Moreover, we explore how such a framework enables contrasting the evolution of different lines of narrative frames across important issues. Herein, we exemplarily discuss the narrative change regarding climate change, i.e., the evolution from *global warming* to the more urgent naming of *climate catastrophe* and similar [8].

As our main contribution, we want to provide an impulse towards further exploration of how narrations are being used to frame long-term discourses. We hope that our work bridges the gap between two similar but still distinct communities.

2. Background

The present work comprises two strands of computational research, based on narrations and frames, respectively. Specifically, we focus on the parts where computational narrative understanding and computational framing analysis mostly overlap.

Computational Narrative Understanding (CNU) Narrations, being defined by their content and structure, are used to study many topics, with the policy process in the narrative policy framework being a well-known example [9]. Herein, the elements of narrativity have first been fully formalized by Piper et al. [10], with the minimal definition being structured as *"Someone tells someone somewhere that || someone did something(s) [to someone] somewhere at some time for some reason"*. Here, the left part (before the ||) is the perspective of *narrating* the story, while the right part concerns the story itself (i.e., *diegesis*). In a similar vein, some strides have already been made towards analyzing narrative frames [5]. Overall, we observe that actors and events are central components of narrations, which provides overlap with some computational framing analysis approaches.

Computational Framing Analysis (CFA) Framing deals with salience in communication [3] and is concerned "how" a text is presented rather than "what" is apparent [11]. The analysis of framing can be seen as a task of natural language understanding (e.g., similar to tasks in the GLUE benchmark [12]). The notion of framing is very distinctively conceptualized in computational literature, comprising supervised and unsupervised, as well as mixed-method based approaches [11]. As supervised approaches depend on corpora and codebooks, unsupervised approaches are more in line with narrative understanding. For instance, DiMaggio et al. [13] use topic modeling for framing analysis and equate certain topics with frames. Besides, they define frames as comprising narratives among other cues, and also find narratives as being part of a particular topic. Other works consider semantic information, such as semantic role labels [6] and semantic graphs [7] to analyze narratives directly.

In the remainder of the paper, we use the theory presented by Piper et al. [10] for CNU and the survey by Ali and Hassan [11] for CFA as cornerstones in their respective areas. Also, when using the term *narrative framing*, we refer to the framing using narratives as device, thus compounding both CNU and CFA. Herein, we focus on framing through semantic structure (i.e., following Fillmore and Baker [14]) rather than other forms of framing.

3. Computational Narrative Framing

As a starting point for better understanding narrative framing, we analyze how their research directions are entangled. Both, Piper et al. [10] and Ali and Hassan [11], identify a set of future research endeavors by stating core challenges and open questions, respectively. We provide an overview of these future directions in Table 1. Comparing them, we observe remarkable overlap between the two strands that we summarize as key requirements.

First (*R1*), there is the improvement of methods by considering fine-grained nuanced features, e.g., latent features (CNU) and semantic relations (CFA). Herein, CNU focuses on understanding deep stories via narrative structuring of higher-order organizing principles, while CFA focuses on semantic relations going beyond words with the aim to better explore frames. Here, we identify *narrative structure* as a key direction for future research.

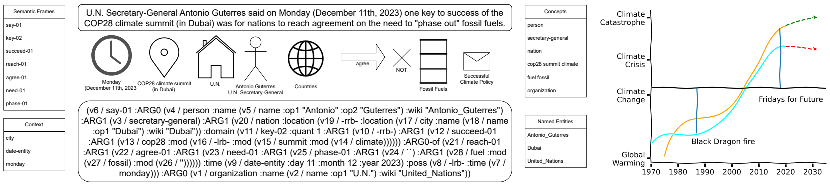
Second (*R2*), the relation between multiple documents (potentially even for distinct types) for a broader understanding are established. CNU aims to understand narrative discourse by studying the interaction of narrative features, even between different narrative products (e.g., movies vs. books). CFA questions how different documents can be connected or inform each other. We reason that the understanding of narratives must go beyond individual narratives and shift towards a focus on *competing narratives*.

Third (*R3*), both emphasize the incorporation of more nuanced knowledge sources, e.g., past events like wars (CNU), culture, and omission (CFA). CNU argues for more robust classification of narrative types via interdisciplinary large-scale registers. CFA calls for a computational model to construct frames via salience through various framing devices. We see the modeling of the *temporal evolution* as a good starting point to capture more nuances.

Based on these suggestions, we reason that computational narrative framing approaches must go beyond simple feature analysis (e.g., on the word-level) of individual documents, but rather analyze the corpus as a whole considering the nuances within. Specifically, we argue that narrative frames emerge from the temporal evolution of collections of documents comprising structural elements. In the following, we aim to synthesize these requirements from the bottom-up.

Table 1
Overview of core challenges in CNU [10] and open question in CFA [11].

	CNU core challenges	CFA open questions (abbreviated)	Synthesized requirements
R1	Narrative beliefs	Capture all relevant semantic relations?	Narrative Structure
R2	Narrative responses	Frames across multiple documents?	Competing Narratives
R3	Narrative economies	Salience through framing devices?	Temporal Evolution



U.N. Secretary-General Antonio Guterres said on Monday (December 11th, 2023) one key to success of the COP28 climate summit (in Dubai) was for nations to reach agreement on the need to "phase out" fossil fuels.

```
(v6 / say-01 :ARG0 (v4 / person :name (v5 / name :op1 "Antonio" :op2 "Guterres") :wiki "Antonio_Guterres")
:ARG1 (v3 / secretary-general) :ARG1 (v20 / nation :location (v19 / -rrb- :location (v17 / city :name (v18 / name
:op1 "Dubai") :wiki "Dubai")) :domain (v11 / key-02 :quant 1 :ARG1 (v10 / -rrb-) :ARG1 (v12 / succeed-01
:ARG1 (v13 / cop28 :mod (v16 / -lrb- :mod (v15 / summit :mod (v14 / climate)))))) :ARG0-of (v21 / reach-01
:ARG1 (v22 / agree-01 :ARG1 (v23 / need-01 :ARG1 (v25 / phase-01 :ARG1 (v24 / ``) :ARG1 (v28 / fuel :mod
(v27 / fossil) :mod (v26 / ")))))) :time (v9 / date-entity :day 11 :month 12 :year 2023) :poss (v8 / -lrb- :time (v7 /
monday))) :ARG0 (v1 / organization :name (v2 / name :op1 "U.N.") :wiki "United_Nations"))
```

(a) Narrative Framing Representation

(b) Temporal Evolution

Figure 1: Exemplary plot on how narratives on climate change could be depicted.

3.1. Narrative Structure

To start, we establish narrative frames that go beyond word frequency, with structure being a focal point. We base the analysis on our prior work [7] using semantic graphs based on abstract meaning representations [15].

In Figure 1a, we depict an example that shows how complex such representations can be, even for short sentences. Specifically, we used a sentence from a recent news article on the COP28[1]:

> U.N. Secretary-General Antonio Guterres said on Monday (December 11th, 2023) one key to success of the COP28 climate summit (in Dubai) was for nations to reach agreement on the need to "phase out" fossil fuels.

We transformed the text to a graph using [16][2] and present its linearized form for brevity. While a detailed explanation is beyond the scope of this paper[3], the key elements that such model extracts are semantic frames (comprising verbs and senses) [14], concepts (nouns), contextual information (time and location), as well as named entities. We want to highlight that the model implicitly performs both simplifications (e.g., singularization of nations to nation) and generalizations (e.g., wikification of U.N. to United Nations), potentially in unison (e.g., stemming and verbification of agreement to agree-01), to improve the resulting representations.

Therefore, this or similar representations are necessary to fulfill the first requirement for computational narrative framing (R1). Note that, we used a straightforward parser here for demonstration, but more recent language models, e.g., BART [17], might be better suited for the task at hand.

3.2. Competing Narratives

After having extracted the narratives of individual documents, we might compare them. In most scenarios, narratives will cluster together and compete with each other, with narratives

[1] Taken from: https://www.reuters.com/business/environment/phasing-out-fossil-fuels-is-key-cop28-success%2Dsays-uns-guterres-2023-12-11/ where we enhanced the text with meta-data from the article, i.e., time and location, which we put in parentheses.

[2] Available as open tool at: https://bollin.inf.ed.ac.uk/amreager.html

[3] The guidelines are available at: https://github.com/amrisi/amr-guidelines/blob/master/amr.md

of conspiracy theories being an obvious instance. Especially regarding the topic of climate change, conspiracy thinking seems larger than anticipated [18]. Even for COP28, conspiracy narratives are spreading, such as relating to the fear of keeping the population captive[4]. Besides considering conspiracies, many intra- and inter-corpus dependencies should also be considered, with polarization [19] being another noteworthy example.

To identify such competing narratives, we can rely on established methods for corpus analysis (e.g., [20]). However, beyond applying them on lexical features (e.g., words), considering the semantic level as established in 3.1 is important for the second requirement (R2).

3.3. Temporal Evolution

While the third requirement (R3) contains many distinct points, we focus on the temporal aspects that we see as the most common factor. Hence, the present should depend on the past, while also account for irregularities like notable omissions of specific narratives. Furthermore, the evolution will depend on the competing narratives established in 3.2. For example, the overall narrative framing might show a similar trend but at a different pace depending on the cultural context, which we visually illustrate using an artificial example in Figure 1b. Notably, certain events could lead to sudden shifts in trajectories that need to be accounted for.

While methods like time-series analyses seems sound at first glance, we believe that due to discreteness of narrative frames, sequential modeling approaches [21] are a better fit. In such models, side-information such as relevant events could be utilized as well.

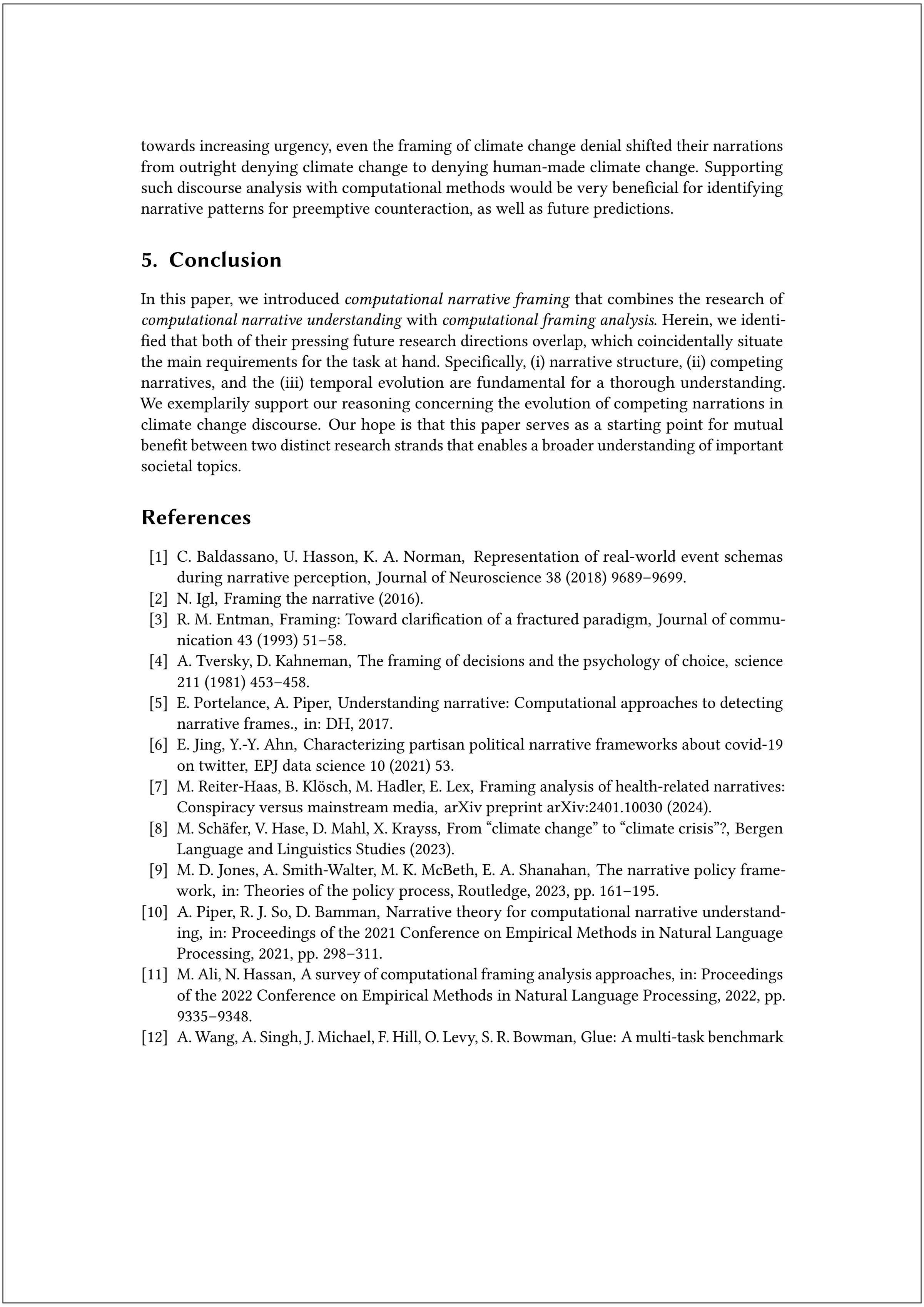
3.4. Challenges in Narrative Framing Analysis

Foremost, we acknowledge that the main challenges identified still remain unsolved. Beyond that, detecting narrative frames is even more difficult to achieve than both CNU and CFA individually. While data is sparse in both domains, there is a complete lack of ground truth data to train algorithms for predicting the narrative framing. Moreover, classical machine learning setups like classification would not work at all, as there is no complete set of narrative frames due to their emergent properties. Finally, the validation, especially quantitatively, is unsolved as the evolving nature of narrative frames hinders most (static) measures.

4. Learning from Evolving Narratives: The Case of Global Warming to Climate Catastrophe

Following up on the topic of the example provided in Figure 1, we now briefly discuss how considering computational narrative framing would support understanding the discourse on climate change. The framing of climate change has gradually shifted from *global warming* to *climate change*, and more recently towards *climate crisis* or even *climate catastrophe* [8]. While anecdotally obvious, such patterns are notoriously challenging to detect computationally when they are not known in advance. Climate change, in particular, is a long-term issue where changes are noticeable even for laymen. Besides the reframing of the scientific consensus

[4] https://phys.org/news/2023-11-climate-conspiracy-theories-flourish-cop28.html



towards increasing urgency, even the framing of climate change denial shifted their narrations from outright denying climate change to denying human-made climate change. Supporting such discourse analysis with computational methods would be very beneficial for identifying narrative patterns for preemptive counteraction, as well as future predictions.

5. Conclusion

In this paper, we introduced *computational narrative framing* that combines the research of *computational narrative understanding* with *computational framing analysis*. Herein, we identified that both of their pressing future research directions overlap, which coincidentally situate the main requirements for the task at hand. Specifically, (i) narrative structure, (ii) competing narratives, and the (iii) temporal evolution are fundamental for a thorough understanding. We exemplarily support our reasoning concerning the evolution of competing narrations in climate change discourse. Our hope is that this paper serves as a starting point for mutual benefit between two distinct research strands that enables a broader understanding of important societal topics.

References

[1] C. Baldassano, U. Hasson, K. A. Norman, Representation of real-world event schemas during narrative perception, Journal of Neuroscience 38 (2018) 9689–9699.

[2] N. Igl, Framing the narrative (2016).

[3] R. M. Entman, Framing: Toward clarification of a fractured paradigm, Journal of communication 43 (1993) 51–58.

[4] A. Tversky, D. Kahneman, The framing of decisions and the psychology of choice, science 211 (1981) 453–458.

[5] E. Portelance, A. Piper, Understanding narrative: Computational approaches to detecting narrative frames., in: DH, 2017.

[6] E. Jing, Y.-Y. Ahn, Characterizing partisan political narrative frameworks about covid-19 on twitter, EPJ data science 10 (2021) 53.

[7] M. Reiter-Haas, B. Klösch, M. Hadler, E. Lex, Framing analysis of health-related narratives: Conspiracy versus mainstream media, arXiv preprint arXiv:2401.10030 (2024).

[8] M. Schäfer, V. Hase, D. Mahl, X. Krayss, From “climate change” to “climate crisis”?, Bergen Language and Linguistics Studies (2023).

[9] M. D. Jones, A. Smith-Walter, M. K. McBeth, E. A. Shanahan, The narrative policy framework, in: Theories of the policy process, Routledge, 2023, pp. 161–195.

[10] A. Piper, R. J. So, D. Bamman, Narrative theory for computational narrative understanding, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 298–311.

[11] M. Ali, N. Hassan, A survey of computational framing analysis approaches, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 9335–9348.

[12] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, S. R. Bowman, Glue: A multi-task benchmark

and analysis platform for natural language understanding, arXiv preprint arXiv:1804.07461 (2018).

[13] P. DiMaggio, M. Nag, D. Blei, Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of us government arts funding, Poetics 41 (2013) 570–606.

[14] C. J. Fillmore, C. F. Baker, Frame semantics for text understanding, in: Proceedings of WordNet and Other Lexical Resources Workshop, NAACL, volume 6, 2001.

[15] L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, N. Schneider, Abstract meaning representation for sembanking, in: Proceedings of the 7th linguistic annotation workshop and interoperability with discourse, 2013, pp. 178–186.

[16] M. Damonte, S. B. Cohen, G. Satta, An incremental parser for abstract meaning representation, in: 15th EACL 2017 Software Demonstrations, Association for Computational Linguistics (ACL), 2017, pp. 536–546.

[17] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, arXiv preprint arXiv:1910.13461 (2019).

[18] J. E. Uscinski, S. Olivella, The conditional effect of conspiracy thinking on attitudes toward climate change, Research & Politics 4 (2017) 2053168017743105.

[19] P. DiMaggio, J. Evans, B. Bryson, Have american's social attitudes become more polarized?, American journal of Sociology 102 (1996) 690–755.

[20] B. L. Monroe, M. P. Colaresi, K. M. Quinn, Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict, Political Analysis 16 (2008) 372–403.

[21] I. Sutskever, O. Vinyals, Q. V. Le, Sequence to sequence learning with neural networks, Advances in neural information processing systems 27 (2014).

We are prone to overestimate how much we understand about the world and to underestimate the role of chance in events.

(Daniel Kahneman in Thinking, Fast and Slow)

4 Conclusion

This thesis improves the understanding of framing by advancing computational methods for extraction and analysis. The research is supported by scientific theories and approaches from natural language understanding. In sum, my work focused on two particular research questions regarding framing, which I applied to polarized topics online.

RQ1: How to detect differences in the framing of online content at various exploratory levels?

I developed different methods that each provide a distinct frame type based on text representations. I provide an overview of the differences and trade-offs in Table 4.1. As there clearly is no single best approach for all scenarios, I argue for a multi-perspective approach, with the methods complementing each other. To answer the three sub-questions, I focus on distinct settings that are suitable for the target of each type of framing analysis.

RQ1a: How to extract Framing Labels with limited annotated data?

In Reiter-Haas et al. (2023a), we use a contrastive loss function together with a multi-stage training procedure to exploit multi-label information for few- and zero-shot scenarios. We find that our approach optimizes the embedding space such that samples with more labels in common tend to be closer, while those with almost no label overlap tend to be dissimilar. In comparison, in an unoptimized space, most samples tend to show similarity regardless of whether their labels overlap or not. Our approach performs especially well in a zero-shot scenario when trained on similar languages (e.g., training with Latin script for zero-shot Spanish framing detection).

RQ1b: How to extract Framing Dimensions in an unsupervised manner?

In Reiter-Haas et al. (2021b), we consider the alignment of documents (e.g., derived from words) in the embedding space together with definitions (e.g., dictionaries). By using antagonistic points in the label space (e.g., by centroids of dictionary word lists), the tendency towards one centroid can be quantified and used for analysis. Consequently, we can distinguish between the moral values of political parties and their followers on social media. For instance, we find that the moral of the Austrian governing parties seems to be reflected in their followers' tweets regarding COVID-19. Hence, the ruling conservative party emphasizes care in their COVID-19 communication, while the opposing social democratic party emphasizes authority, which seems to be issue-specific and unlike the typical leanings.

	Labels (a)	Dimensions (b)	Structure (c)
Training Data	few* samples	pole information	none
Mode	supervised*	unsupervised	discovery
Extraction	scalar (per label)	n-D	irregular
Exploration	low	medium	high
Validation	high	medium	low
Settings	competition	aggregation	narrative
Application	classical	intuitive	challenging

Table 4.1: Comparison of the three framing detection approaches employed in the thesis: (a) label prediction, (b) dimensional alignment, and (c) structural analysis. The complexity increases from left to right but similarly increases in exploratory potential. * Labels (a) can also be employed in an unsupervised zero-shot fashion.

RQ1c: How to extract Framing Structure without prior conceptualization?

In Reiter-Haas et al. (2024c), we convert textual documents to graph-based representations for structural analysis. This approach does not rely on annotated data nor predefining the target of the study. Using this approach, we find that several argument structures seem to distinguish mainstream and conspiracy media on health-related topics. Using abstract meaning representations, we find that well-established differences, such as science vs. beliefs, are being used as framing devices. Moreover, we also find more subtle differences like a focus on the immediacy of certain issues in conspiracy media compared to mainstream media.

Comparison of Approaches for RQ1a-RQ1c

In Table 4.1, which is based on the initial analysis in Reiter-Haas (2023), I briefly compare the main aspects of the three framing detection approaches used in the thesis. It clearly shows the trade-offs between exploration capabilities and the complexity of the analysis. Classical label prediction is conceptually simple to employ but requires most prior knowledge. While structural analysis enables more freedom for exploration, it is challenging to work with the extracted irregular structures. Due to these trade-offs, the work conducted in the course of this thesis led to an open-source multi-perspective tool (Reiter-Haas et al., 2024b) for the detection of well-established frames, as well as the discovery of novel frames based on structural differences.

RQ2: How does framing relate to online information behavior?

In this thesis, I approach this question from three distinct angles. First, in Reiter-Haas et al. (2023b), we observe that opinions have a similar tendency offline and online when considering the sentiment of social media posts as a proxy for framing (i.e., frame production). Second, as analyzed as part of RQ1, we find that certain types of media and groups have different frames present in their content (i.e., frame prevalence). Third, in Reiter-Haas and Lex (2024), we observe that online behavior (e.g., in terms of consumed news items) seems to be influenced by the frames present (i.e., frame consumption). Taken together, we argue that users produce frames in online systems by writing texts that reflect their real opinions, which leads to noteworthy shifts of frames being prevalent concerning aggregated data such as media types or user groups, which then influences the content that users consume on the platform. As such, I want to emphasize that framing is a critical field of study in online information systems. Nevertheless, with our current understanding, it is still challenging to predict how the framing expressed in online content influences changes in behavior.

4.1 Implications

My work has several important implications for the body of existing computational framing and online behavior research. It extended the state-of-the-art in framing detection with three distinct approaches while also supporting the research community with an open-source tool for conducting framing analysis, which can be used by computer scientists and social scientists alike. I approached the nuanced nature of frames by considering multiple angles rather than exclusively relying on data annotation. Furthermore, my work provides empirical evidence that framing influences online behavior in several aspects. This finding is far-reaching, as online platforms have thus an increased urgency to also consider the framing of information in addition to more established types of content analysis like hate-speech detection (see Schmidt and Wiegand, 2017 for an overview). So far, my research extends mostly to polarized topics, for which I hypothesized framing effects to be more pronounced compared to non-polarizing topics. I base my hypothesis on the linguistic divergence (e.g., regarding usage frequencies, sentiment, or meaning of words) that has been found in polarizing online media (Karjus & Cuskley, 2024). However, framing could likely influence other topics in a similar manner while being more subtle to detect, as such linguistic cues ought to be largely absent. Hence, my hope is that my research sparks increased interest in exploring framing in online content as a research direction, especially since contemporary Transformer-based models allow for more in-depth studies of textual data, thus enabling the extraction of more volatile frames. The implications of two areas are especially noteworthy, namely the exploration-validation trade-off and the research conducted toward frame behavior.

Exploration-Validation Trade-off. My work sheds light on the exploration-validation trade-off, i.e., that more validated frames have less explorative potential and vice versa. Specifically, certain framing approaches lack thorough validation that is established in other fields, e.g., with specific validity criteria, which is known to be challenging for frames (D'Angelo, 2017). In my research, the lower extent of validation arises due to the fundamental trade-off between more explorative and more quantifiable approaches. Hence, the higher the amount of possible validation, the less the approach is suitable for exploration. Nevertheless, I aim to mitigate potential validation issues as best as possible. In the prediction task (Reiter-Haas et al., 2023a), we competed against several other teams on a leaderboard that was used as ground truth. In the structural analysis (Reiter-Haas et al., 2024c), we have the performance score of the model itself while we cross-compare the found patterns with the established body of literature. In sum, I argue for a complementary approach using both more validated and more explorative framing detection methods.

Towards Frame Behavior. My research goes beyond static framing detection, i.e., frame prevalence, and also analyzes framing behavior, i.e., frame production and consumption. I established that the opinions expressed online are similar to opinions offline, using sentiment as a proxy (Reiter-Haas et al., 2023b). Moreover, my work showed that frames play a vital role in users' online consumption behavior (Reiter-Haas & Lex, 2024). Therefore, I see a strong need for further investigation in this particular area that bridges information behavior and framing theory. I suspect that studying framing behavior also leads to valuable insights for an improved understanding of opinion polarization.

4.2 Reflections

I see my work as an important direction in an emergent research area. Nevertheless, my research comes with several shortcomings that should be accounted for and mitigated in subsequent contributions. I briefly discuss three noteworthy limitations of my work.

Relation to Polarization. Although I rely on *polarized topics* for studying frames, I do not explicitly explore the relation between frames and polarization itself. Therefore, considerations such as whether framing enhances polarization or whether framing is the result of polarization are left unexplored. I believe that framing behavior must first be understood more thoroughly before such questions can be properly tackled. Nevertheless, it does raise some concerns that two distinct theories are jointly considered without an established link. Still, I aim to advance the complex research area framing on several fronts, which would not be feasible with a purely reductionist approach or without setting appropriate boundaries.

Vague Conceptualizations. There is no clear definition of framing yet, neither in the social sciences nor computational sciences. Similar to previous computational framing works, I also do not rely on precise conceptualization. Moreover, the three distinct frame types (i.e., labels, dimensions, and structure) that I explore have noteworthy differences between them, and some of them may even rely on imperfect conceptualization (e.g., based on topics rather than frames) out of necessity. Out of this limitation, I see our multi-perspective approach as an essential step toward capturing the broad range of potential frames. Instead of tailoring our approaches to specific conceptualizations (e.g., media, moral, or semantic frames), my work considers types of frames (i.e., frame labels, dimensions, and structure) more broadly.

Causality. My studies are empirically conducted and, thus, do not establish causality. Establishing causality requires an extensive amount of additional work that could result in several publications on its own. Nevertheless, I now briefly describe how causality could be established in framing. For frame prevalence, a causal inference could be established by carefully modeling the causal dependencies and considering potential confounders (see Pearl and Mackenzie, 2018). Conducting such studies could answer whether the material source (e.g., political affiliation or media orientation) is indeed the reason for the framing or other factors that are at play. Besides, the consideration of causality in framing behavior would ideally necessitate user studies for both frame production and frame consumption. For frame production, conducting initial surveys before users start producing content could be established. For framing consumption, an interventional study could be conducted. For instance, assigning users to different treatments that determine what particular framed content is shown to them and subsequently compare the differences in outcome, i.e., their behavior.

4.3 Future Research

Besides resolving the established limitations, I see many promising research directions. Four directions, in particular, seem to be logical successors to my conducted research.

Temporal aspects. In my thesis, similar to other works, I studied frames statically. However, there are several temporal aspects that are worth considering, most notably the evolution of frames over time. Frames could evolve due to language shifts, societal changes, or due to sudden events. First, language tends to evolve over time in itself, which can be analyzed with approaches like word shift graphs (Gallagher et al., 2021). Second, framing is an important consideration for society and is thus affected by changes or could even be a central aspect of the changes themselves (e.g., in the case of social movements; Benford and Snow, 2000). Third, framing could abruptly change when triggered by events, such as conflicts (see Alkaabi, 2024 for a recent example of framing the Israeli-Palestinian conflict). Hence, with my collaborators, we plan a longitudinal study of the evolution of frames regarding climate change in media that comprises all three types of frame evolution. Besides frame evolution, other temporal aspects are changes in behavior, patterns in discourses, and sequential models that would also advance framing analysis.

Framing bias mitigation. Framing can be considered as a form of content bias (Entman, 2007), which should be mitigated similarly to other biases (e.g., how societal biases are mitigated using adversarial learning in Rekabsaz et al., 2021). Moreover, the repeat consumption and diversity of frames could be directly reduced with debiasing strategies, e.g., in the form of a novel loss function. Additionally, the framing should be increased in similar content, e.g., on the same topics, rather than merely diversifying by suggesting different kinds of content. Hence, rather than balancing by interleaving certain kinds of frames in one issue with another kind in a completely different issue, the diversity of frames within both issues should be increased. To that end, one could again leverage improved sequential models as discussed above. Furthermore, behavioral research could be incorporated into information systems, e.g., to better support decision-making via explanations or the nudge theory (Thaler & Sunstein, 2009). However, suggesting content that goes beyond one's belief has the potential to lead to a backfire effect, hence amplifying their beliefs (Bail et al., 2018, being a noteworthy field experiment in this area). As an alternative to suggesting the content, providing explanations is likely to even have a more positive overall impact.

Unification. As another prominent research direction, I see unification in several areas as desirable. First, I hope to support the long-standing issue of harmonizing the different conceptualization and research strands in computer science and social science. Within this context, I see our multi-perspective tool (Reiter-Haas et al., 2024b) as a starting point, as framing is inherently nuanced, preventing single perspectives from capturing the latent aspects. Second, we will strive to consolidate our distinct approaches into a more coherent framework, e.g., by using the same pretrained network and thus using shared embeddings for subsequent tasks (i.e., frame label, frame dimension, and frame structure extraction). Such joint training would improve performance, as well as increase consistency and interpretability when using multiple representations, thus allowing for deeper analysis. Likewise, I strive towards convergence with other areas, with narrative understanding being our first contender (Reiter-Haas et al., 2024a), as it would improve the understanding of narrative frames.

The role of LLMs. Finally, large language models (LLMs) will be another accelerator for computational framing research. The current thesis relies heavily on the Transformer architecture, which is the basis of virtually all LLMs. However, as LLMs have only very recently started to produce good results reliably, they are not a consideration in the current thesis. Still, I see LLMs playing an integral role in the future, especially on two fronts. First, LLMs could be used as a validation tool. LLMs can be used to assign labels, e.g., based on a selection of predefined labels, to text, similar to how humans annotate text but in a more scalable manner. In fact, human annotations will play a lesser role, as users of crowdsourcing platforms are suspected of using LLMs for their tasks (Veselovsky et al., 2023). While LLMs could, in theory, perform exploratory framing detection, they would suffer from similar challenges in aggregating the data points, as the assigned labels could wildly differ and thus have similar validation constraints. Second, content produced by LLMs is likely to influence users' behavior, which is thus an important research direction. In this context, LLMs could be used for investigating causality in framing behavior by first reframing particular texts and then presenting them to users or exploiting causal reasoning directly (Kıcıman et al., 2023). The gained knowledge should then, in turn, be used for framing bias mitigation. Altogether, LLMs can heavily influence framing research, which makes studying how to use LLMs for societal good of utmost importance.

Acronyms

ACT-R Adaptive Control of Thought—Rational
AMR Abstract Meaning Representations

BART Bidirectional and Auto-Regressive Transformers
BERT Bidirectional Encoder Representations from Transformers

CFA Computational Framing Analysis
CNU Computational Narrative Understanding
COVID-19 Coronavirus Disease 2019

ELMo Embeddings from Language Models

GloVe Global Vectors
GPT Generative Pre-trained Transformer
GRU Gated Recurrent Units
GVFC Gun Violence Frame Corpus

HeroCon Heterogeneous Contrastive Learning

LDA Latent Dirichlet Allocation
LOCO Language of Conspiracy Corpus
LSTM Long Short-Term Memory

mCPT multlingual Contrastive Pre-training of Transformers
MFD-2 Moral Foundations Dictionary Version 2
MFT Moral Foundations Theory
MIND Microsoft News Dataset

NLG Natural Language Generation
NLP Natural Language Processing
NLU Natural Language Understanding

PCA Principal Component Analysis

RNN Recurrent Neural Network
RoBERTa Robustly optimized BERT approach

SBERT Sentence-BERT
SetFit Sentence Transformer Fine-tuning
SRL Semantic Role Labeling

t-SNE t-distributed Stochastic Neighbor Embedding

ULMFiT Universal Language Model Fine-tuning
UMAP Uniform Manifold Approximation and Projection

Bibliography

Ajjour, Y., Alshomary, M., Wachsmuth, H., & Stein, B. (2019). Modeling frames in argumentation. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2922–2932. https://doi.org/10.18653/v1/d19-1290 (cit. on p. 14).

Ali, M., & Hassan, N. (2022). A survey of computational framing analysis approaches. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 9335–9348. https://doi.org/10.18653/v1/2022.emnlp-main.633 (cit. on pp. 1, 14).

Alkaabi, N. M. (2024). *Socio-political aspects in framing narratives of conflict* [Doctoral dissertation, University of Leicester]. https://doi.org/10.25392/leicester.data.25061255.v1 (cit. on p. 131).

Allen, L., Scott, J., Brand, A., Hlava, M., & Altman, M. (2014). Publishing: Credit where credit is due. *Nature*, *508*(7496), 312–313. https://doi.org/10.1038/508312a (cit. on p. 18).

Ambros, R., Bernsteiner, A., Bloem, R., Dolezal, D., Garcia, D., Göltl, K., Haagen-Schützenhöfer, C., Hadler, M., Hell, T., Herderich, A., et al. (2023). Two-year progress of pilot research activities in teaching digital thinking project (tdt). *Zeitschrift für Hochschulentwicklung*, *18*(Sonderheft Hochschullehre), 117–136. https://doi.org/10.3217/zfhe-SH-HL/07 (cit. on p. 7).

Anderson, A., Kumar, R., Tomkins, A., & Vassilvitskii, S. (2014). The dynamics of repeat consumption. *Proceedings of the 23rd international conference on World wide web*, 419–430. https://doi.org/10.1145/2566486.2568018 (cit. on p. 10).

Baeza-Yates, R. (2018). Bias on the web. *Communications of the ACM*, *61*(6), 54–61. https://doi.org/10.1145/3209581 (cit. on p. 10).

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*. https://doi.org/10.48550/arXiv.1409.0473 (cit. on p. 13).

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. F., Lee, J., Mann, M., Merhout, F., & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, *115*(37), 9216–9221. https://doi.org/10.31235/osf.io/4ygux (cit. on p. 131).

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The berkeley framenet project. *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*. https://doi.org/10.3115/980451.980860 (cit. on p. 11).

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., & Schneider, N. (2013). Abstract meaning representation for sembanking. *Proceedings of the 7th linguistic annotation workshop and inter-*

operability with discourse, 178–186. https://aclanthology.org/W13-2322 (cit. on pp. 13, 14).

Bates, M. J. (2010). Information behavior. *Encyclopedia of library and information sciences*, *3*, 2381–2391. https://doi.org/10.1081/E-EISA-120053335 (cit. on p. 10).

Benford, R. D., & Snow, D. A. (2000). Framing processes and social movements: An overview and assessment. *Annual review of sociology*, *26*(1), 611–639. https://doi.org/10.1146/annurev.soc.26.1.611 (cit. on p. 131).

Bhatia, V., Akavoor, V. P., Paik, S., Guo, L., Jalal, M., Smith, A., Tofu, D. A., Halim, E. E., Sun, Y., Betke, M., et al. (2021). Openframing: Open-sourced tool for computational framing analysis of multilingual data. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 242–250. https://doi.org/10.18653/v1/2021.emnlp-demo.28 (cit. on p. 15).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022. https://doi.org/10.7551/mitpress/1120.003.0082 (cit. on p. 14).

Bonial, C., Lukin, S., Doughty, D., Hill, S., & Voss, C. (2020). Infoforager: Leveraging semantic search with amr for covid-19 research. *Proceedings of the Second International Workshop on Designing Meaning Representations*, 67–77. https://aclanthology.org/2020.dmr-1.7 (cit. on p. 15).

Bramson, A., Grim, P., Singer, D. J., Berger, W. J., Sack, G., Fisher, S., Flocken, C., & Holman, B. (2017). Understanding polarization: Meanings, measures, and model evaluation. *Philosophy of science*, *84*(1), 115–159. https://doi.org/10.1086/688938 (cit. on p. 11).

Card, D., Boydstun, A., Gross, J. H., Resnik, P., & Smith, N. A. (2015). The media frames corpus: Annotations of frames across issues. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 438–444. https://doi.org/10.3115/v1/p15-2072 (cit. on pp. 1, 15).

Cho, K., van Merriënboer, B., Gulçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder–decoder for statistical machine translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734. https://doi.org/10.3115/v1/d14-1179 (cit. on p. 13).

Clarke, C. L., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., & MacKinnon, I. (2008). Novelty and diversity in information retrieval evaluation. *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 659–666. https://doi.org/10.1145/1390334.1390446 (cit. on p. 10).

D'Angelo, P. (2017). Framing: Media frames. *The international encyclopedia of media effects*, 1–10. https://doi.org/10.1002/9781118783764.wbieme0048 (cit. on p. 129).

Del Vicario, M., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2017). Modeling confirmation bias and polarization. *Scientific reports*, *7*(1), 40391. https://doi.org/10.1038/srep40391 (cit. on pp. 1, 11).

Demszky, D., Garg, N., Voigt, R., Zou, J., Shapiro, J., Gentzkow, M., & Jurafsky, D. (2019). Analyzing polarization in social media: Method and application to tweets on 21 mass shootings, 2970–3005. https://doi.org/10.18653/v1/n19-1304 (cit. on p. 14).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. https://doi.org/10.48550/arXiv.1810.04805 (cit. on p. 13).

DiMaggio, P., Evans, J., & Bryson, B. (1996). Have american's social attitudes become more polarized? *American journal of Sociology*, *102*(3), 690–755. https://doi.org/10.1086/230995 (cit. on p. 11).

DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of us government arts funding. *Poetics*, *41*(6), 570–606. https://doi.org/10.1016/j.poetic.2013.08.004 (cit. on pp. 1, 14).

Draws, T., Tintarev, N., & Gadiraju, U. (2021a). Assessing viewpoint diversity in search results using ranking fairness metrics. *ACM SIGKDD Explorations Newsletter*, *23*(1), 50–58. https://doi.org/10.1145/3468507.3468515 (cit. on p. 10).

Draws, T., Tintarev, N., Gadiraju, U., Bozzon, A., & Timmermans, B. (2021b). This is not what we ordered: Exploring why biased search result rankings affect user attitudes on debated topics. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 295–305. https://doi.org/10.1145/3404835.3462851 (cit. on p. 10).

Ellison, A. M. (1987). Effect of seed dimorphism on the density-dependent dynamics of experimental populations of atriplex triangularis (chenopodiaceae). *American Journal of Botany*, *74*(8), 1280–1288. https://doi.org/10.2307/2444163 (cit. on p. 11).

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 (cit. on pp. 1, 11).

Entman, R. M. (2007). Framing bias: Media in the distribution of power. *Journal of communication*, *57*(1), 163–173. https://doi.org/10.1111/j.1460-2466.2006.00336.x (cit. on p. 131).

Fillmore, C. J., et al. (1976). Frame semantics and the nature of language. *Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech*, *280*(1), 20–32. https://doi.org/10.1111/j.1749-6632.1976.tb25467.x (cit. on p. 11).

Fillmore, C. J., & Baker, C. F. (2001). Frame semantics for text understanding. *Proceedings of WordNet and Other Lexical Resources Workshop, NAACL*, *6* (cit. on p. 11).

Gallagher, R. J., Frank, M. R., Mitchell, L., Schwartz, A. J., Reagan, A. J., Danforth, C. M., & Dodds, P. S. (2021). Generalized word shift graphs: A method for visualizing and explaining pairwise comparisons between texts. *EPJ Data Science*, *10*(1), 4. https://doi.org/10.1140/epjds/s13688-021-00260-3 (cit. on p. 131).

Garimella, K., et al. (2018). Polarization on social media. http://urn.fi/URN:ISBN:978-952-60-7833-5 (cit. on p. 11).

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology* (pp. 55–130, Vol. 47). Elsevier. https://doi.org/10.1016/b978-0-12-407236-7.00002-4 (cit. on pp. 10, 12).

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*. https://doi.org/10.48550/arXiv.2203.05794 (cit. on pp. 14, 17).

Hadler, M., Ertl, A., Klösch, B., Reiter-Haas, M., & Lex, E. (n.d.). Development and framing of climate gluing protests: A media analysis (1986-2023) using sentiment analyses and frame detection models [in Review] (cit. on p. 7).

Hadler, M., Ertl, A., Klösch, B., Reiter-Haas, M., & Lex, E. (2024). The climate gluing protests: Analyzing their development and framing in media since 1986 using sentiment analyses and frame detection models [Accepted at the 16th Conference of the European Sociological Association]. (Cit. on p. 7).

Hadler, M., Klösch, B., Lex, E., & Reiter-Haas, M. (2021). Polarization in public opinion: Combining social surveys and big data analyses of twitter (suf edition) [AUSSDA, V1, UNF:6:jPjxWXqS6RVg4uYo3Zplcw== [fileUNF]. https://doi.org/10.11587/OVHKTR (cit. on p. 7).

Hadler, M., Klösch, B., Reiter-Haas, M., & Lex, E. (2022). Combining survey and social media data: Respondents' opinions on covid-19 measures and their willingness to provide their social media account information. *Frontiers in Sociology*, *7*, 885784. https://doi.org/10.3389/fsoc.2022.885784 (cit. on p. 7).

He, Z., Mokhberian, N., Câmara, A., Abeliuk, A., & Lerman, K. (2021). Detecting polarized topics using partisanship-aware contextualized topic embeddings, 2102–2118. https://doi.org/10.18653/v1/2021.findings-emnlp.181 (cit. on p. 11).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735 (cit. on p. 13).

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification, 328–339. https://doi.org/10.18653/v1/p18-1031 (cit. on p. 13).

Jing, E., & Ahn, Y.-Y. (2021). Characterizing partisan political narrative frameworks about covid-19 on twitter. *EPJ data science*, *10*(1), 53. https://doi.org/10.1140/epjds/s13688-021-00308-4 (cit. on p. 14).

Karjus, A., & Cuskley, C. (2024). Evolving linguistic divergence on polarizing social media. *Humanities and Social Sciences Communications*, *11*(1), 1–14. https://doi.org/10.1057/s41599-024-02922-9 (cit. on pp. 11, 129).

Kingsbury, P. R., & Palmer, M. (2002). From treebank to propbank. *LREC*, 1989–1993. http://www.lrec-conf.org/proceedings/lrec2002/pdf/283.pdf (cit. on pp. 11, 12).

Kıcıman, E., Ness, R., Sharma, A., & Tan, C. (2023). Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*. https://doi.org/10.48550/arXiv.2305.00050 (cit. on p. 132).

Klösch, B., Hadler, M., Reiter-Haas, M., & Lex, E. (2021a). Polarization of opinions on political measures in german-speaking countries – a comparison between the covid19 crisis and the climate crisis [Presented at the 15th Conference of the European Sociological Association]. (Cit. on p. 7).

Klösch, B., Hadler, M., Reiter-Haas, M., & Lex, E. (2022). Social desirability and the willingness to provide social media accounts in surveys. the case of environmental attitudes, 119–127. https://doi.org/10.4995/carma2022.2022.15069 (cit. on p. 7).

Klösch, B., Hadler, M., Reiter-Haas, M., & Lex, E. (2023). Polarized opinions on covid-19 and environmental policy measures. the role of social media use and personal concerns in german-speaking countries. *Innovation: The European Journal of*

Social Science Research, 1–24. https://doi.org/10.1080/13511610.2023.2201877 (cit. on p. 7).

Klösch, B., Reiter-Haas, M., Hadler, M., & Lex, E. (2021b). Erkenntnisse und herausforderungen in der kombination von umfrage- und twitter-daten: Eine untersuchung der gesellschaftlichen polarisierung in der covid-19 debatte im deutschsprachigen raum [Presented at Gemeinsamer Kongress der Deutschen Gesellschaft für Soziologie (DGS) und der Österreichischen Gesellschaft für Soziologie (ÖGS).]. (Cit. on p. 7).

Kowald, D., Reiter-Haas, M., Kopeinik, S., Schedl, M., & Lex, E. (2024). Transparent music preference modeling and recommendation with a model of human memory theory. In *A human-centered perspective of intelligent personalized environments and systems* (pp. 113–136). Springer. https://doi.org/10.1007/978-3-031-55109-3_4 (cit. on p. 7).

Kwak, H., An, J., Jing, E., & Ahn, Y.-Y. (2021). Frameaxis: Characterizing microframe bias and intensity with word embedding. *PeerJ Computer Science*, *7*, e644. https://doi.org/10.7717/peerj-cs.644 (cit. on pp. 1, 3, 14).

Lakoff, G. (2014). *The all new don't think of an elephant!: Know your values and frame the debate*. Chelsea Green Publishing. https://doi.org/10.15718/discog.2015.22.3.125 (cit. on p. 11).

Lawrence, J., & Reed, C. (2020). Argument mining: A survey. *Computational Linguistics*, *45*(4), 765–818. https://doi.org/10.1162/coli_a_00364 (cit. on p. 14).

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703 (cit. on p. 13).

Liao, Q., Lai, M., & Nakov, P. (2023). Marseclipse at semeval-2023 task 3: Multi-lingual and multi-label framing detection with contrastive learning. *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 83–87. https://doi.org/10.18653/v1/2023.semeval-1.10 (cit. on pp. 1, 15).

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. https://doi.org/10.18653/v1/k19-1047 (cit. on pp. 14, 15).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*. https://doi.org/10.48550/arXiv.1907.11692 (cit. on p. 13).

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. *1*, 157. https://doi.org/10.1017/cbo9780511809071 (cit. on p. 13).

Mayring, P., et al. (2004). Qualitative content analysis. *A companion to qualitative research*, *1*(2), 159–176. https://doi.org/10.1016/b978-0-12-818630-5.11031-0 (cit. on p. 13).

McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, *3*(29). https://doi.org/10.21105/joss.00861 (cit. on p. 14).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*. https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf (cit. on p. 12).

Mokhberian, N., Abeliuk, A., Cummings, P., & Lerman, K. (2020). Moral framing and ideological bias of news. *International Conference on Social Informatics*, 206–219. https://doi.org/10.1007/978-3-030-60975-7_16 (cit. on p. 14).

Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2008). Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, *16*(4), 372–403. https://doi.org/10.1093/pan/mpn018 (cit. on p. 14).

Moscati, M., Wallmann, C., Reiter-Haas, M., Kowald, D., Lex, E., & Schedl, M. (2023). Integrating the act-r framework with collaborative filtering for explainable sequential music recommendation, 840–847. https://doi.org/10.1145/3604915.3608838 (cit. on p. 7).

Neal, T., Sundararajan, K., Fatima, A., Yan, Y., Xiang, Y., & Woodard, D. (2017). Surveying stylometry techniques and applications. *ACM Computing Surveys (CSuR)*, *50*(6), 1–36. https://doi.org/10.1145/3132039 (cit. on p. 14).

Opitz, J., & Frank, A. (2022). Sbert studies meaning representations: Decomposing sentence embeddings into explainable semantic features. *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, 625–638. https://aclanthology.org/2022.aacl-main.48 (cit. on p. 15).

Pang, B., Lee, L., et al. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, *2*(1–2), 1–135. https://doi.org/10.1561/9781601981516 (cit. on pp. 1, 11, 14).

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic books. https://doi.org/10.1126/science.aau9731 (cit. on p. 130).

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/d14-1162 (cit. on p. 12).

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018, June). Deep contextualized word representations. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (pp. 2227–2237). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1202 (cit. on p. 12).

Pilehvar, M. T., & Camacho-Collados, J. (2020). *Embeddings in natural language processing: Theory and advances in vector representations of meaning*. Morgan & Claypool Publishers. https://doi.org/10.1007/978-3-031-02177-0 (cit. on p. 12).

Piper, A., So, R. J., & Bamman, D. (2021). Narrative theory for computational narrative understanding. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 298–311. https://doi.org/10.18653/v1/2021.emnlp-main.26 (cit. on p. 10).

Piskorski, J., Stefanovitch, N., Da San Martino, G., & Nakov, P. (2023). Semeval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 2343–2361. https://doi.org/10.18653/v1/2023.semeval-1.317 (cit. on pp. 1, 15).

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., & Stein, B. (2018). A stylometric inquiry into hyperpartisan and fake news. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 231–240. https://doi.org/10.18653/v1/p18-1022 (cit. on p. 14).

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (cit. on p. 13).

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks, 3982–3992. https://doi.org/10.18653/v1/d19-1410 (cit. on p. 13).

Reiter-Haas, M. (2023). Exploration of framing biases in polarized online content consumption. *Companion Proceedings of the ACM Web Conference 2023*, 560–564. https://doi.org/10.1145/3543873.3587534 (cit. on pp. 7, 128).

Reiter-Haas, M. (2024, January). *Mind-small frames*. Zenodo. https://doi.org/10.5281/zenodo.10509498 (cit. on p. 7).

Reiter-Haas, M., Ertl, A., Innerebner, K., & Lex, E. (2023a). Mcpt at semeval-2023 task 3: Multilingual label-aware contrastive pre-training of transformers for few-and zero-shot framing detection, 941–949. https://doi.org/10.18653/v1/2023.semeval-1.130 (cit. on pp. 1, 3, 7, 15, 127, 129).

Reiter-Haas, M., Klösch, B., Hadler, M., & Lex, E. (2020). Bridging the gap of polarization in public opinion on misinformed topics. *12th International Conference on Social Informatics: SocInfo 2020*. https://events.kmi.open.ac.uk/misinformation/assets/accepted-papers/socinfo2020_misinformation_reiterhaas.pdf (cit. on p. 7).

Reiter-Haas, M., Klösch, B., Hadler, M., & Lex, E. (2021a). Opinion polarization on covid-19 measures: Integrating surveys and social media data. https://easychair.org/publications/preprint/4vvZ (cit. on p. 7).

Reiter-Haas, M., Klösch, B., Hadler, M., & Lex, E. (2023b). Polarization of opinions on covid-19 measures: Integrating twitter and survey data. *Social Science Computer Review*, *41*(5), 1811–1835. https://doi.org/10.1177/08944393221087662 (cit. on pp. 7, 11, 128, 130).

Reiter-Haas, M., Klösch, B., Hadler, M., & Lex, E. (2023c). Semantic graphs reveal the narrative framing in news [Presented at NetSci 2023, Vienna, Austria.]. https://socialcomplab.github.io/polarization/publications/2023netsci_narrative.pdf (cit. on p. 7).

Reiter-Haas, M., Klösch, B., Hadler, M., & Lex, E. (2024a). Computational narrative framing: Towards identifying frames through contrasting the evolution of narrations. *Text2Story 2024: Seventh International Workshop on Narrative Extraction from Texts held in conjunction with the 46th European Conference on Information Retrieval*. https://ceur-ws.org/Vol-3671/paper11.pdf (cit. on pp. 7, 132).

Reiter-Haas, M., Klösch, B., Hadler, M., & Lex, E. (2024b). Framefinder: Explorative multi-perspective framing extraction from news headlines. *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval*, 381–385. https://doi.org/10.1145/3627508.3638308 (cit. on pp. 7, 128, 131).

Reiter-Haas, M., Klösch, B., Hadler, M., & Lex, E. (2024c). Framing analysis of health-related narratives: Conspiracy versus mainstream media. *arXiv preprint arXiv:2401.10030*. https://doi.org/10.48550/arXiv.2401.10030 (cit. on pp. 4, 7, 128, 129).

Reiter-Haas, M., Kopeinik, S., & Lex, E. (2021b). Studying moral-based differences in the framing of political tweets. *Proceedings of the International AAAI Conference on Web and Social Media*, *15*, 1085–1089. https://doi.org/10.1609/icwsm.v15i1.18135 (cit. on pp. 3, 7, 127).

Reiter-Haas, M., & Lex, E. (2024). The framing loop: Do users repeatedly read similar framed news online. *Proceedings of the 7th HUMANIZE Workshop*. https://ceur-ws.org/Vol-3660/paper18.pdf (cit. on pp. 7, 128, 130).

Reiter-Haas, M., Parada-Cabaleiro, E., Schedl, M., Motamedi, E., Tkalcic, M., & Lex, E. (2021c). Predicting music relistening behavior using the act-r framework. *Proceedings of the 15th ACM Conference on Recommender Systems*, 702–707. https://doi.org/10.1145/3460231.3478846 (cit. on pp. 7, 10).

Rekabsaz, N., Kopeinik, S., & Schedl, M. (2021). Societal biases in retrieved contents: Measurement framework and adversarial mitigation of bert rankers. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 306–316. https://doi.org/10.1145/3404835.3462949 (cit. on p. 131).

Rumelhart, D. E., Hinton, G. E., Williams, R. J., et al. (1985). Learning internal representations by error propagation. https://doi.org/10.21236/ada164453 (cit. on p. 13).

Salton, G., Wong, A., & Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620. https://doi.org/10.1145/361219.361220 (cit. on p. 12).

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*. https://doi.org/10.48550/arXiv.1910.01108 (cit. on p. 13).

Scheufele, D. A. (1999). Framing as a theory of media effects. *Journal of communication*, *49*(1), 103–122. https://doi.org/10.1093/joc/49.1.103 (cit. on p. 11).

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. *Proceedings of the fifth international workshop on natural language processing for social media*, 1–10. https://doi.org/10.18653/v1/w17-1101 (cit. on p. 129).

Shurafa, C., Darwish, K., & Zaghouani, W. (2020). Political framing: Us covid19 blame game. *International Conference on Social Informatics*, 333–351. https://doi.org/10.1007/978-3-030-60975-7_25 (cit. on p. 14).

Sikder, O., Smith, R. E., Vivo, P., & Livan, G. (2020). A minimalistic model of bias, polarization and misinformation in social networks. *Scientific reports*, *10*(1), 5493. https://doi.org/10.1038/s41598-020-62085-w (cit. on p. 11).

Soroya, S. H., Farooq, A., Mahmood, K., Isoaho, J., & Zara, S.-e. (2021). From information seeking to information avoidance: Understanding the health information behavior during a global health crisis. *Information processing & management*, *58*(2), 102440. https://doi.org/10.1016/j.ipm.2020.102440 (cit. on p. 1).

Sullivan, K. (2023). Three levels of framing. *Wiley Interdisciplinary Reviews: Cognitive Science*, e1651. https://doi.org/10.1002/wcs.1651 (cit. on p. 11).

Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. Penguin. (Cit. on p. 131).

Tunstall, L., Reimers, N., Jo, U. E. S., Bates, L., Korat, D., Wasserblat, M., & Pereg, O. (2022). Efficient few-shot learning without prompts. *arXiv preprint arXiv:2209.11055*. https://doi.org/10.48550/arXiv.2209.11055 (cit. on p. 13).

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 (cit. on p. 11).

Vallejo, G., Baldwin, T., & Frermann, L. (2023). Connecting the dots in news analysis: A cross-disciplinary survey of media bias and framing. *arXiv preprint arXiv:2309.08069*. https://doi.org/10.48550/arXiv.2309.08069 (cit. on p. 2).

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(11). https://jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf (cit. on p. 14).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*. https://papers.nips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (cit. on pp. 1, 9, 12, 13).

Veselovsky, V., Ribeiro, M. H., & West, R. (2023). Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks. *arXiv preprint arXiv:2306.07899*. https://doi.org/10.48550/arXiv.2306.07899 (cit. on p. 132).

Wang, T., & Isola, P. (2020). Understanding contrastive representation learning through alignment and uniformity on the hypersphere. *International Conference on Machine Learning*, 9929–9939. https://proceedings.mlr.press/v119/wang20k/wang20k.pdf (cit. on p. 13).

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, *33*, 5776–5788. https://proceedings.neurips.cc/paper_files/paper/2020/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (cit. on p. 13).

Wardana, R., Klösch, B., Reiter-Haas, M., & Penker, M. (2024). Media representation of environmental movements: Unveiling frames in news articles on the last generation in austria [Accepted at the 16th Conference of the European Sociological Association]. https://socialcomplab.github.io/polarization/publications/2024esa_lastgen.pdf (cit. on p. 7).

Westerwick, A., Johnson, B. K., & Knobloch-Westerwick, S. (2017). Confirmation biases in selective exposure to political online information: Source bias vs. content bias. *Communication Monographs*, *84*(3), 343–364. https://doi.org/10.1080/03637751.2016.1272761 (cit. on p. 1).

White, R. W., & Hassan, A. (2014). Content bias in online health search. *ACM Transactions on the Web (TWEB)*, *8*(4), 1–33. https://doi.org/10.1145/2663355 (cit. on p. 1).

Wicke, P., & Bolognesi, M. M. (2020). Framing covid-19: How we conceptualize and discuss the pandemic on twitter. *PloS one*, *15*(9), e0240010. https://doi.org/10.1371/journal.pone.0240010 (cit. on p. 14).

Wilson, T. D. (2000). Human information behavior. *Informing science*, *3*, 49. https://doi.org/10.28945/576 (cit. on p. 10).

Wilson, T. D. (1981). On user studies and information needs. *Journal of documentation*, *37*(1), 3–15. https://doi.org/10.1108/eb026702 (cit. on p. 10).

Wu, B., Razuvayevskaya, O., Heppell, F., Leite, J. A., Scarton, C., Bontcheva, K., & Song, X. (2023). Sheffieldveraai at semeval-2023 task 3: Mono and multilingual approaches for news genre, topic and persuasion technique classification. *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 1995–2008. https://doi.org/10.18653/v1/2023.semeval-1.275 (cit. on pp. 1, 15).

Xing, Y., Zhang, J. Z., Storey, V. C., & Koohang, A. (2024). Diving into the divide: A systematic review of cognitive bias-based polarization on social media. *Journal of Enterprise Information Management*, *37*(1), 259–287. https://doi.org/10.1108/jeim-09-2023-0459 (cit. on p. 1).

Yin, W., Hay, J., & Roth, D. (2019). Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3914–3923. https://doi.org/10.18653/v1/d19-1404 (cit. on p. 13).