DISSERTATION | DOCTORAL THESIS

Titel | Title

Validation in unsupervised Computational Text Analysis Methods

verfasst von | submitted by

Jana Bernhard-Harrer BA Bakk.phil. MSc MSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Doktorin der Philosophie (Dr.phil.)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 796 310 301
Dissertationsgebiet lt. Studienblatt | Field of study as it appears on the student record sheet:	Publizistik- und Kommunikationswissenschaft
Betreut von | Supervisor:	Univ.-Prof. Hajo Boomgaarden PhD

Acknowledgements

Although this Dissertation is the culmination of years of hard work on my part, it would not have been possible without the support, guidance, and encouragement of many people, to whom I am deeply grateful.

First and foremost, I would like to thank my supervisor, **Hajo G. Boomgaarden**, who believed in me from the very beginning. He allowed me to work in my own way and provided great support from the very start by introducing me to academia already as his Teaching Assistant. His openness to all my concerns and questions has been invaluable, and his guidance has profoundly influenced both this dissertation and my development as a researcher. I am also immensely grateful to my co-supervisors, **Jakob-Moritz Eberl** and **Petro Tolochko**, who always had an open door for me. Their realistic optimism and readiness to answer every question helped carry me through this process. They consistently supported my ideas, encouraged my independence, and helped me make my own decisions, for which I am incredibly thankful. Besides my advisors, I would like to thank the rest of my thesis committee, **Prof. Anne Kroon** and **Prof. Annett Heft**, for having offered to review my thesis, for their insightful comments and encouragements.

To all my wonderful co-workers, and especially you **Ahrabhi**, thank you all for keeping me sane over the past three and a half years. Your constant presence, encouragement, and positivity lifted me through all the highs and lows. Whether it was smiling at me during conference presentations or celebrating milestones, you were there for me every step of the way. And, of course, thank you for keeping the excitement alive with endless new project ideas that we'll tackle as soon as we get our those unlimited contracts!

I'd also like to say a big thank you to **my family**, who have been immensely supportive and encouraging. A special thanks to my brother, who was always there for the programming language questions and to everyone for being ready to listen when it all felt overwhelming. But, most importantly, thank you all for being the best family support system I could have asked for. Finally, to my best friend and husband, **Lukas**, thank you for your endless patience through the long work hours, the stress, and the demands of pursuing a PhD. You've listened to my presentations more times than anyone should, ~~tried to take~~ *took* an interest in my endless ramblings about science and the truth, and always had my back no matter what. I couldn't have done this without you!

There were so many more colleagues at conferences, journals, workshops and meetings, so to everyone who has supported me, whether directly or indirectly, throughout this journey—thank you.

Abstract

This dissertation explores the validation of unsupervised computational text analysis methods, focusing specifically on word embeddings and topic modeling in the field of computational social science. The need for reliable automated text analysis methods has increased as digitization expands access to textual data. This work explores the methodological challenges of validating these methods to ensure they produce credible and consistent results.

The first study examines the validation of word embedding models by assessing the impact of hyperparameter settings on their performance and stability when trained on large text corpora. It highlights the critical role of validation in model selection and shows how different settings can lead to different interpretations of semantic relationships.

The second study systematically reviews validation practices in topic modeling across 792 studies, revealing a lack of standardization in validation approaches. It emphasizes the importance of adopting more qualitative and context-specific validation methods to increase the credibility of topic modeling studies.

The third study evaluates the influence of different validation strategies on selecting and evaluating topic models, clearly showing the need for transparent and objective validation practices to reduce researcher bias and improve model reliability.

Through these studies, the dissertation identifies gaps in current validation practices and proposes best practices for ensuring the rigor and validity of computational text analysis. The findings aim to provide actionable guidelines for improving the accuracy and credibility of research findings in the social sciences, emphasizing the importance of aligning validation tasks with specific research objectives. Overall, this work contributes to developing more robust methodologies in the computational analysis of social and cultural phenomena.

Kurzfassung

Diese Dissertation untersucht die Validierung unüberwachter computergestützter Textanalysemethoden *(unsupervised computational text analysis methods)* und konzentriert sich dabei speziell auf Worteinbettungen *(word embeddings)* und Themenmodellierung *(topic modeling)* im Bereich der computergestützten Sozialwissenschaften. Der Bedarf an zuverlässigen automatisierten Textanalysemethoden ist mit der Digitalisierung und der damit verbundenen Erweiterung des Zugangs zu Textdaten gestiegen. Diese Arbeit untersucht die methodischen Herausforderungen bei der Validierung dieser Methoden, um sicherzustellen, dass sie glaubwürdige und konsistente Ergebnisse liefern.

Die erste Studie untersucht die Validierung von Worteinbettungsmodellen, indem sie die Auswirkung von Hyperparametereinstellungen auf ihre Leistung und Stabilität beim Training auf großen Textkorpora bewertet. Sie unterstreicht die entscheidende Rolle der Validierung bei der Modellauswahl und zeigt, wie unterschiedliche Einstellungen zu unterschiedlichen Interpretationen semantischer Beziehungen führen können.

Die zweite Studie gibt einen systematischen Überblick über die Validierungspraktiken bei der Themenmodellierung in 792 Studien und zeigt einen Mangel an Standardisierung bei den Validierungsansätzen auf. Sie unterstreicht, wie wichtig es ist, qualitativere und kontextspezifischere Validierungsmethoden anzuwenden, um die Glaubwürdigkeit von Studien zur Themenmodellierung zu erhöhen.

Die dritte Studie bewertet den Einfluss verschiedener Validierungsstrategien auf die Auswahl von Themenmodellen und macht deutlich, dass transparente und objektive Validierungsverfahren erforderlich sind, um die Voreingenommenheit der Forschenden zu verringern und die Zuverlässigkeit der Modelle zu verbessern.

Anhand dieser Studien werden in der Dissertation Lücken in den derzeitigen Validierungsverfahren aufgezeigt und bewährte Verfahren zur Gewährleistung der Strenge und Validität der computergestützten Textanalyse vorgeschlagen. Die Ergebnisse zielen darauf ab, umsetzbare Richtlinien für die Verbesserung der Genauigkeit und Glaubwürdigkeit von Forschungsergebnissen in den Sozialwissenschaften bereitzustellen, wobei die Bedeutung der Abstimmung von Validierungsaufgaben auf spezifische Forschungsziele betont wird. Insgesamt trägt diese Arbeit dazu bei, validere Methoden für die computergestützte Analyse sozialer und kultureller Phänomene zu entwickeln.

Information on Publication Status of the Studies included in this Dissertation

Comprehensive Validation of Word Embeddings for Social Science Research

Submitted to Social Science Computer Review (SSCR) on 18. July 2024

Beyond Standardization: A Comprehensive Review of Topic Modeling Validation Methods for Computational Social Science Research

Submitted to Political Science Research Method (PSRM) on 24. November 2023
Accepted for Revise & Resubmit on 10. March 2024
Accepted for Second Round of Revise & Resubmit on 13. September 2024

Topic Model Validation Methods and their Impact on Model Selection and Evaluation

Bernhard, J., Teuffenbach, M., & Boomgaarden, H. G. (2023). Topic Model validation methods and their impact on Model selection and evaluation. Computational Communication Research, 5(1), 1. https://doi.org/10.5117/CCR2023.1.13.BERN

Contents

Acknowledgements i

Abstract iii

Kurzfassung v

Information on Publication Status of the Studies included in this Dissertation vii

1 Introduction 1

2 Theoretical Foundation 5
- 2.1 Science and Methodology . . . 5
- 2.2 (Automated) Text Analysis and Text as Data . . . 7
- 2.3 Quality Criteria in Science . . . 12
- 2.4 Validity . . . 13

3 Research Questions 17

4 Studies 19
- 4.1 Comprehensive Validation of Word Embeddings for Social Science Research 19
- 4.2 Beyond Standardization: A Comprehensive Review of Topic Modeling Validation Methods for Computational Social Science Research . . . 52
- 4.3 Study 3: Topic Model Validation Methods and their Impact on Model Selection and Evaluation . . . 83

5 Central Results 111
- 5.1 Study 1: Comprehensive Validation of Word Embeddings for Social Science Research . . . 111
- 5.2 Study 2: Beyond Standardization: A Comprehensive Review of Topic Modeling Validation Methods for Computational Social Science Research . 112
- 5.3 Study 3: Topic Model Validation Methods and their Impact on Model Selection and Evaluation . . . 113

6 Critical Assessment and Limitations 115

7 Discussion 119

Bibliography 123

1 Introduction

The ongoing digitization of society has resulted in a notable increase in digital and computer-readable text availability. This increased availability of digitized text data presents a significant opportunity for the social sciences, as it has become considerably more common to utilize text as a data source (Grimmer et al., 2022). Consequently, this has resulted in the introduction of a range of automated text analysis techniques, which have been adopted from adjacent disciplines such as computer science or computational linguistics (Boumans & Trilling, 2018). Researchers can now analyze vast amounts of written content and explore different research questions or social science variables with newfound depth (Domahidi et al., 2019; Van Atteveldt & Peng, 2018). These techniques have shown remarkable versatility, being widely employed over the past decade to address different research questions in the social sciences.

Automated text analysis methods, while increasingly popular, do not come without challenges. Originating from the computer sciences, these methods often carry assumptions that may not fit well in the social sciences (see for example Bonikowski & Nelson, 2022). This misalignment requires that researchers understand the nuances of the methods they are applying to their datasets so that they can make informed decisions when using a particular method (e.g. Baden, Pipal et al., 2022). This has sparked a discussion within the computational social sciences community about effectively using automated text analysis for sound social science research. Many of the criticisms leveled at these methods concern issues of validity - that is, whether they actually measure what they purport to measure. This issue is further illustrated by the fact that a researcher's decisions can significantly impact study results (see for example Antoniak & Mimno, 2018; Denny & Spirling, 2018; Tolochko et al., 2024). As a result, the topic of validation remains a prominent and ongoing debate within the field (e.g. Baden, Pipal et al., 2022; Birkenmaier et al., 2023; Lind et al., 2023).

In more traditional methods, such as manual content analysis, various validation tasks have been developed to address questions of validity. However, the issue is more complex for computational science approaches to text analysis. The most appropriate method must be selected based on several factors, including whether the approach is supervised or unsupervised, the availability of a gold standard, and the specific nature of the study (Baden, Pipal et al., 2022; Birkenmaier et al., 2023; Song et al., 2020). Validation depends on many factors and is further complicated by the choice of model. Many automated text analysis methods involve a model selection step, which includes tasks such as setting hyperparameters or choosing the algorithm that converts text into numerical data. This complexity means that validation needs to be considered early in the process and before the final model choice is made and results are finalized, which is a significant departure from traditional validation methods. Consequently, established validation strategies need

to be re-evaluated and adapted when working with automated text analysis methods.

This *methodological dissertation* aims to contribute to the ongoing discussion on the validation of automated text analysis methods, with a particular focus on their application in computational communication science. The overarching research question that guides this work is: *how can unsupervised computational text analysis methods be evaluated to produce more valid results?* By addressing this question, the dissertation seeks to offer novel insights and practical solutions to some of the most pressing challenges in the field. It criticizes current practices, identifies potential areas for improvement, and suggests first steps towards rethinking validation, intending to provide a basis for developing more useful validation strategies in the field of computational social science. In more detail, this dissertation focuses on two widely used computational methods: word embeddings in the first study and topic modeling in the second and third studies. These methods are chosen based on their approaches to representation. Word Embeddings capture linguistic context by representing words in continuous vector spaces, thus capturing semantics and cultural nuances of terms. Topic Modeling, on the other hand, represents the thematic structure of text, revealing hidden patterns within the data. Unlike other methods, such as cluster analysis or dictionary approaches, primarily serving as "measurements," word embeddings and topic models provide deeper "representations" of the data, making them particularly suitable for examining complex social and cultural phenomena.

The first study, "Comprehensive Validation of Word Embeddings for Social Science Research" emphasizes the importance of validating embeddings to ensure their accuracy and interpretability in capturing semantic meanings within large text corpora. The study systematically applies various intrinsic and extrinsic validation techniques to assess the impact of hyperparameter settings on model performance. The study shows that different validation methods favor different parameter settings, and a practical example further highlights that the different settings leads to different substantive interpretations. Thus, the study shows that the choice of validation method impacts the word embedding model choice and thus, how words are vectorized and *understood*.

The second study, "Beyond Standardization: A Comprehensive Review of Topic Modeling Validation Methods for Computational Social Science Research", presents a comprehensive literature review of the validation methods used for topic modeling. The review synthesizes findings from 792 studies over two decades, and offers a detailed account of how topic models are validated. It contributes to the field by showcasing a profound lack of convergence toward specific validation methods. The review attributes this to an inherent mismatch between the inductive and qualitative nature of topic modeling and the deductive, quantitative research tradition that seeks standardized validation practices. As a result, it advocates for better considering qualitative validation understandings by building on transparency and detailed reporting to enhance the credibility of the use of topic models.

The third study, "Topic Model Validation Methods and their Impact on Model Selection and Evaluation" focuses on the impact of different validation methods on topic model selection. By applying two topic modeling algorithms, LDA and Top2Vec, to the same text corpus, the study assesses how various validation strategies influence model choice and,

thus, also the results. The findings reveal significant discrepancies in model performance based on the chosen validation methods, underscoring the need for a transparent and thorough validation approach. This study proposes a four-step recommendation plan to guide researchers in selecting and validating topic models effectively, aiming to improve the application of topic models in the social sciences.

This methodological dissertation is organized as follows: First, the theoretical foundation is established. It begins by exploring the definition of science and the role of methodology in scientific inquiry. The discussion then turns to what distinguishes social science from other scientific fields, before introducing the concept of "text as data," detailing both manual and automated approaches to text analysis, and situating these methods within the broader landscape of social science research. Next, it addresses the quality criteria that are essential for conducting scientific research. Finally, it delves into the concept of validity in social science research, offering a detailed examination of how validity is defined, applied, and evaluated, and discussing the specific challenges of ensuring validity in both research methods and findings. The core of the dissertation presents the three empirical studies. Additionally, each study is given its own section where the central findings are highlighted again. Afterwards, the dissertation discusses the limitations of these research efforts. This section critically evaluates the approaches taken in the studies, acknowledging the limitations and potential weaknesses of the methodologies used. The dissertation concludes with a comprehensive discussion synthesizing the three studies' findings. This final section highlights the practical recommendations that can be drawn from the research and provides guidance for future work in the field. It also outlines the theoretical contributions to ongoing discussions about validation in social science research. Finally, the dissertation reflects on what these findings mean for the broader use of automated text analysis methods in studying social science phenomena, emphasizing the importance of rigorous validation practices in producing valid and valuable research results.

The novelty of this dissertation lies in problematizing and reframing the conversation around the validation of automated text analysis methods. Validation is not as simple as 'just doing it', especially in computational social science, where the diversity of research questions and operationalization of latent constructs, introduces considerable complexity. Rather than prescribing a one-size-fits-all approach, this work serves as a call to action, urging the field to critically engage with the inherent challenges of validating automated text analysis methods. By examining the problem from multiple perspectives, this dissertation seeks to initiate a critical thought process within the research community that encourages a more nuanced and reflective discussion of the difficulties involved. While I do not claim to offer definitive solutions to these problems, this dissertation aims to highlight the need for innovative approaches and promotes a more conscious and transparent approach to validation, thus contributing to the foundations for valid methodologies in the field.

2 Theoretical Foundation

Science plays a central role in society by shedding light on pressing societal problems and helping to identify, test, and measure effective solutions (Scharrer & Ramasubramanian, 2021). This principle applies to all fields of science, provided that the methods of inquiry are appropriate and scientific. Without rigorous methodology, we cannot be sure that our findings reflect what is real (Babbie, 2020; Godfrey-Smith, 2009; Scharrer & Ramasubramanian, 2021), which could potentially hinder the scientific process and lead to a growing public distrust of scientists (European Commission & Directorate-General for Communication, 2021). This mistrust is particularly worrying as science is essential for tackling major crises that societies are facing and necessitates research about how credible results can be obtained.

There is a rich history of scientific methodology and a long-standing debate about what constitutes scientific knowledge. However, technology and society are changing, and so are our approaches to science. Therefore, it is crucial to continue to advance methodological research, especially by considering new methods from recent advances in computation. These methods often involve complex steps humans can no longer fully understand, such as those found in unsupervised automated analysis. Nevertheless, they are also essential to keep up with the pace of modern society and scientific research. This dissertation thus wants to contribute to the discussion of how validity can and should be included in automated content analysis.

This chapter begins by exploring the fundamental question, "What is science?" and delves into the importance of the scientific method, also emphasizing its importance for the quality of scientific findings. Following this, the discussion turns to the social sciences, examining what sets them apart from other fields of inquiry, including a brief exploration of the epistemological differences between quantitative and qualitative approaches. Next, the chapter narrows down to text analysis methods and discusses manual and computational techniques. Finally, the chapter introduces different quality criteria in science, focusing on the social sciences, before engaging in a more extended discussion on the concept of validity. This discussion highlights various approaches to validity in the social sciences and the inherent challenges of measurement in these fields.

2.1 Science and Methodology

At its core, science is about seeking to know more. It is derived from the Latin word "scientia", which means "the results of logical demonstration, revealing general and necessary truths" or *knowledge* (Godfrey-Smith, 2009, p. 5). The field of philosophy of science has long been concerned with the question of what constitutes science, which areas

are scientific, and where the boundary between scientific and non-scientific disciplines lies. This is done to find a possible explanation for how humans can know things about the world around them (Okasha, 2002). This discussion is of great importance because it reflects the evolving boundaries of science as it adapts over time to incorporate new methodologies and areas of inquiry. For example, there was a long debate questioning whether the social sciences should be considered true sciences (Burawoy, 2008). However, they are widely accepted as valid scientific disciplines today.

What distinguishes the social sciences from other scientific disciplines is their focus on understanding human behavior, social structures, and cultural phenomena. Unlike the natural sciences, which often seek to uncover universal laws governing physical and biological processes, the social sciences grapple with the complexity of human experience, which is usually contextual and shaped by historical, cultural, and societal factors (Godfrey-Smith, 2009; Scharrer & Ramasubramanian, 2021). This means that social science research must consider the diversity of human perspectives and the fluidity of social constructs, which often complicates the standardization found in other fields. Moreover, the subjects of social science research —human beings and their societies— are inherently reflexive; they can interpret and respond to the research process, introducing a level of difficulty that is less common in the study of the natural sciences. As a result, social scientists must be particularly attuned to issues of bias, representation, and ethical considerations while balancing the need for methodological rigor with the challenges of capturing the richness and variability of social life.

Godfrey-Smith (2009) emphasizes three theories of how science works: a) empiricism, which asserts that the only source of knowledge about the world is organized and systematic experience; b) mathematics, which posits that science excels at understanding the world through mathematical concepts and tools; and c) social structures, which argues that what distinguishes science from other types of inquiry is its unique social structure of trust and collaboration among scientists. While the first point is closely connected to the methods of observation, the last point raises further questions about determining whom to trust, what experiences are relevant, and who is a reliable source. Thus, this definition opens up questions of power and inequality in scientific processes. The historic Eurocentric dominance in science has often led to the marginalization of non-European methodologies and perspectives, sometimes dismissing them as unscientific, which has perpetuated a narrow and exclusionary view of what constitutes legitimate scientific knowledge (Scharrer & Ramasubramanian, 2021).

The question of how science works is closely related to the question of scientific methodology and how scientific research is conducted. Different scientific fields use different methods to advance their knowledge. While, medicine relies heavily on experimentation; physics relies on both experimentation and mathematical proof. In contrast, the social sciences rely on experimentation and surveys, content analysis, or observation to gather insights into their participants. There are many different versions of these archetypal methods. The choice of which often depends on several factors, including the specific research questions being addressed, the resources available, the training and expertise of the researchers. Consequently, the decision of which approach is taken on a specific

methodology (i.e. quantitative or qualitative) will significantly shape the research design and can influence everything from data collection to analysis, the type of data needed, and sometimes also the theoretical framework guiding the study. This highlights the diverse and complex landscape of scientific inquiry withing the social sciences. For example, qualitative researchers may conduct focus group interviews to explore participants' in-depth experiences, while quantitative researchers may use panel studies to collect numerical data about these experiences over time. Similarly, in content analysis, qualitative researchers may interpret a small set of texts in detail, while quantitative researchers may count the frequency of specific terms or actors in a large corpus of texts.

Both approaches aim to understand phenomena, but use different versions of the same methods to achieve their goals (Babbie, 2020; Scharrer & Ramasubramanian, 2021). Nevertheless, many scholars have also advocated for mixed methods, which bridge the qualitative-quantitative divide by combining both approaches (Baškarada & Koronios, 2018; Bryman, 2006), as each has its strengths and weaknesses: qualitative research provides depth and context, while quantitative research offers generalizability and statistical power. Technological advances have also led to the rise of computational methods, which build on traditional core methods but offer new possibilities and are often not easily placed in this divide. These methods allow researchers to handle vast amounts of data and uncover previously undetectable patterns (Grimmer et al., 2022). Moreover, this evolution is not just about data generation but also data analysis. How statistical analysis is carried out has changed, incorporating more sophisticated techniques and tools (Coenen & Smits, 2022; Kruschke & Liddell, 2018). Despite these ongoing developments and consistent differences, all research methods share the same goal: discovering insights about the real world.

Thus, one way to define science is by examining how it is conducted, which is inherently tied to the scientific methods applied. This is understood as empiricism, which suggests that science excels at finding answers due to its organized and systematic nature. These systems of inquiry – or methods – , rooted in either qualitative or quantitative traditions, are essential tools guiding research in the social sciences. The continuous debates on the best methodologies and their advancement are crucial to upholding the core principles of science: seeking sound and comprehensive knowledge about the world and ensuring the validity and robustness of our findings. Methodological research not only refines scientific practices but also ensures that science remains a dynamic and self-correcting process.

2.2 (Automated) Text Analysis and Text as Data

In today's society, many of our activities are text-based, including messaging, writing reports, or consuming news (Gentzkow et al., 2019). In addition, spoken words can be transcribed into text, allowing conversations to be recorded and analyzed. Visuals can also be described in text through captions and detailed descriptions. As such, text is a critical medium for conveying information in various contexts. Text, either segmented into sentences, claims, or paragraphs or even as full documents, has long been an essential source of data in the social sciences, especially in communication science and linguistics

(Chen et al., 2023; Grimmer & Stewart, 2013; Lehmann, 2004). The inherent complexity of language adds to the richness of text as a data source, offering researchers a substantial repository of information to analyze and interpret (Tolochko & Boomgaarden, 2019). Text production is a highly social process that adheres to linguistic rules like grammar, as well as social norms, trends, and specific formalities, such as character limits on social media. This social nature of text creation means that it reflects cultural, societal, and contextual nuances that are invaluable for social science to study. Moreover, the meaning of words is often context-dependent, with connotations adding further layers of meaning, making textual analysis a nuanced and intricate task (Grimmer et al., 2022). These layers of meaning can reveal underlying social dynamics, power relations, and cultural norms, providing a richer understanding of human communication. Furthermore, the ever-changing nature of language and the introduction of new terminology and slang contribute to the increasing complexity and richness of text as a data source. In text, researchers can analyze both manifest variables, such as the frequency with which a particular actor is mentioned (Balluff et al., 2024), and latent variables, such as the sentiment expressed towards that actor (Balluff et al., 2023). This dual capability facilitates a nuanced understanding of social phenomena.

One notable advantage of using text as data is the ability to analyze text retrospectively, allowing researchers to study historical data if it is accessible. This retrospective analysis can provide insights into past social trends and behaviors, enabling scholars to track changes over time and understand the evolution of societal norms and issues (Babbie, 2020; Scharrer & Ramasubramanian, 2021). In addition, using textual data often eliminates the need for direct human interaction, thus simplifying some ethical considerations compared to experiments involving human subjects. Textual analysis allows for examining naturally occurring communication, thus providing a more authentic glimpse into human interactions and societal narratives (McKee, 2003). However, there are also notable drawbacks to using text as data. Coding text is labor- and resource-intensive, requiring knowledge of the study topic and the type of text and a significant time investment. (Grimmer et al., 2022; Hovy & Prabhumoye, 2021). Furthermore, researchers face the challenge of accurately interpreting the original meaning intended by the authors of the text. Contextual nuances, cultural references, and idiomatic expressions complicate the task, potentially leading to misinterpretations and biases in the analysis (Montgomery & Crittenden, 1977; Scharrer & Ramasubramanian, 2021).

Manual text analysis is a widely employed methodology in the social sciences, both qualitative and quantitative. It can be used independently or in conjunction with other methods, such as analyzing written survey responses or creating stimuli for experiments (Babbie, 2020). As many automated text analysis methods are discussed from the point of view of a quantitative research tradition, this section focuses more on quantitative manual text analysis. Manual analysis can be conducted directly by researchers, research assistants, or student coders. In recent years, online platforms like MTurk or Prolific have also been increasingly used for text analysis (Lind et al., 2017; Peter & Lauf, 2002). The quality of the coding procedure is significantly influenced by who performs the coding, as all coders must possess knowledge about the study's topic, the language of the

texts, and the context and content of the texts (Hovy & Prabhumoye, 2021), and thus coder training is essential to ensure that everyone understands what the study aims to measure. Despite its benefits, manual text analysis faces criticism, especially concerning the consistency and accuracy of coding complex concepts like sentiment (van Atteveldt et al., 2021) or latent variables such as racism (Kathirgamalingam, Lind, Bernhard-Harrer & Boomgaarden, 2024). These challenges have led to the practice of involving multiple coders per text, aiming to mitigate inconsistencies by averaging the coding results from multiple perspectives or using a majority vote on the final coding. This approach is supposed to enhance the reliability and validity of the analysis. However, others have also highlighted the risks behind this aggregation (Plank, 2022; Scharrer & Ramasubramanian, 2021) and argued for taking into account this variation in coding and analyzing it as well (Baden et al., 2023; Kathirgamalingam, Lind & Boomgaarden, 2024). The increasing demand for efficient and reproducible content analysis has led researchers to explore automation through technological advances. This shift has led to applying computational science methods, which promise significant improvements, especially on time and data restrictions (Grimmer et al., 2022). As a result, the last decade has seen the rise of computational social science, a burgeoning field with unique implications, methods, and traditions. While the computational social sciences also include methods beyond automated text analysis, such as network analysis (see for example Heft & Buehling, 2022), or simulations, such as agent-based models (Waldherr, 2014), this dissertation focuses only on text analysis methods. This interdisciplinary field continues to expand, reflecting its growing importance in academic research (Bonikowski & Nelson, 2022; Domahidi et al., 2019; Lazer et al., 2020; Van Atteveldt & Peng, 2018; Wallach, 2018).

Automated text analysis encompasses a variety of different methodologies, each rapidly evolving due to technological advances and changing trends in academic training and application (Grimmer et al., 2022; Lazer et al., 2020; Margolin, 2019; van Atteveldt et al., 2022). Among the pioneers in the field, Grimmer and colleagues (Grimmer & Stewart, 2013; Grimmer et al., 2022) proposed an initial classification of these methods based on their objectives: discovery, measurement, or inference. This framework highlights the versatile nature of computational methods, which often defy traditional categorizations as either qualitative or quantitative approaches. Instead, they can serve both paradigms, depending on the research questions. Another classification, looks specifically at the realm of automated text analysis, and distinguishes three primary approaches: supervised, semi-supervised, and unsupervised learning. Supervised learning involves training models on labeled data, with input texts paired to known outputs. This method relies heavily on human-annotated datasets to teach the model how to accurately categorize or predict future data. It is highly effective when ample labeled data is available, allowing for precise and targeted analysis (Boumans & Trilling, 2018; Song et al., 2020), thus often used for the goals of measurement or inference. Semi-supervised learning bridges the gap between supervised and unsupervised methods by utilizing a small amount of labeled data alongside a larger pool of unlabeled data. This approach leverages the labeled data to inform and guide the analysis of the unlabeled data, improving the model's performance without requiring every piece of data to be labeled (Grimmer et al., 2022). Unsupervised

learning, on the other hand, does not require labeled data. Instead, it seeks to uncover hidden patterns and structures within the text. Techniques such as clustering and topic modeling are standard in unsupervised learning, allowing researchers to identify natural groupings and themes in the data. This method is particularly useful when the goal is to explore the underlying structure of large, unannotated text corpora. This versatility of automated text analysis opens up many possibilities for researchers, allowing them to explore and analyze vast amounts of textual data with unprecedented efficiency and depth (Guo et al., 2020; Welbers et al., 2017). This adaptability additionally, underscores the potential of computational methods to transform content analysis across disciplines, fostering innovative approaches and insights (Grimmer et al., 2022; Lazer et al., 2020).

As described above, this dissertation focuses on unsupervised learning in computational text analysis. Specifically, it focuses on two methods that have been frequently applied in the computational social sciences: word embeddings and topic modeling. Both are commonly employed in the social sciences to analyze and comprehend textual data (Grimmer et al., 2022; Hilbert et al., 2019). Word embeddings represent words as vectors in a high-dimensional space, capturing semantic relationships. For instance, in political communication, they can reveal how terms such as "democracy" and "freedom" relate together in the communication of different parties over time (Antoniak & Mimno, 2018). Word embeddings can be used solely as an input for other computational methods (Bowman et al., 2017; Grootendorst, 2022), as a way in which computers can understand and read text, or as a research tool themselves (see for example, Garg et al., 2018; Kroon et al., 2020), which makes them an essential contribution to the computational methods. This possible two-fold use, however, complicates the validation process, which needs to be adapted to how it is applied, and if applied in combination with other methods, requires a two-step validation (first, the embedding model itself; second, the application of the model). Topic modeling, in contrast, identifies underlying latent themes in large text corpora (Angelov, 2020; Blei et al., 2003; Doogan, 2022). In political communication, topics related to elections, external events, or specific policy outcomes might be uncovered. Topic modeling is a computational method that includes strong qualitative inputs (Isoaho et al., 2021), as the topics need to be labeled and interpreted to be useful for analysis. This makes it especially useful for studies focusing on discoveries (Chen et al., 2023; Grimmer et al., 2022). Both methods are model-based and inherently simplify language, which means they are neither absolutely true nor false (Grimmer & Stewart, 2013; Grimmer et al., 2022), or as Box and Draper Box and Draper (1987) put it, "wrong". However, they serve as valuable points of comparison. This simplification, rather than rendering them useless, provides significant value by distilling complex language data into more manageable and interpretable forms, thereby aiding scientific research (Godfrey-Smith, 2009).

While automated text analysis offers significant advantages, it is not a panacea and comes with several pitfalls (Baden, Pipal et al., 2022; Törnberg & Uitermark, 2021). One major concern involves the ethical implications of data access. Researchers sometimes exploit readily available text data, such as scraping tweets without obtaining consent or assuming that the data's sheer volume ensures anonymity. This practice overlooks basic

ethical considerations and also potentially harms individuals' privacy (Williams et al., 2017). Additionally, copyright laws pose legal challenges when scraping content from news websites or blogs, necessitating careful adherence to legal standards (Van Atteveldt et al., 2019). Data collection is also complicated by the sheer amount of possible data sources (e.g. the same actors on different platforms, linking to news media outlets), which are difficult to link (see for example Heft et al., 2024). Moreover, many research methods, especially in the field of unsupervised approaches, are quite unstable or rely strongly on researchers' decisions, such as data included (Maier et al., 2020), data cleaning (Denny & Spirling, 2018; Tolochko et al., 2024) or hyperparameter setting (Antoniak & Mimno, 2018), which can create ethical issues around reproducability. Another issue is the tendency to analyze all available data, driven by the belief that more data equates to better results. This approach often ignores established research principles on sampling, leading to potential biases, as the absence of specific texts can also introduce systematic biases, skewing the analysis and leading to incomplete or misleading conclusions (Grimmer et al., 2022; Hovy & Prabhumoye, 2021). Resource availability is another significant challenge. Automated text analysis requires substantial computational power, which can be costly and limits accessibility. Scholars at well-funded institutions have a distinct advantage, with greater means to employ advanced methods (Bender et al., 2021). Moreover, a disproportionate amount of research focuses on English texts, given the prevalence of tools developed for the English language. This, in turn, creates a barrier to analyzing texts in other languages at an equivalent level (Baden, Dolinsky et al., 2022; Licht & Lind, 2023; Lucas et al., 2015). The rapid evolution of technology further complicates the landscape. Many automated methods originate from computational science, prompting resistance from social scientists who emphasize the need for a socio-scientific perspective in evaluating and adapting these methods (Ignatow, 2016; Nelson, 2020; Wallach, 2018), before they are used in the social sciences. This resistance stems from concerns that these methods are often applied without a thorough understanding of their mathematical underpinnings or the assumptions they make about the data (Baden, Pipal et al., 2022; Bonikowski & Nelson, 2022). This lack of understanding is particularly problematic with complex methods like neural networks or large language models, which function as black boxes. Researchers cannot fully comprehend the inner workings of these models, raising questions about the transparency and interpretability of the results. This opacity challenges the scientific rigor and accountability of research, necessitating careful consideration of the implications for the scientific process and social science in particular (Pääkkönen, 2021).

Despite significant research into the pitfalls and methods of automated text analysis, particularly from a social scientific viewpoint, the rapid pace of technological advancement often outstrips the slower, more meticulous nature of academic research. This discrepancy creates tension, underscoring the need to critically examine how we apply computational methods, interpret our findings, and communicate our results. Maintaining scientific quality requires transparency in the research process, reproducibility of analyses, and a discussion of the reliability and validity of measurements. Computational social scientists must not lose sight of the fact that their work is based on scientific principles, and that rigorous methodology is a fundamental part of this. The evaluation of computational

methods, especially those involving unsupervised steps, is essential for established scientific criteria. Only then can we realize the full potential of automated text analysis to advance our field, while maintaining the integrity and reliability of the results and outcomes.

2.3 Quality Criteria in Science

To ensure the integrity and quality of findings in the sciences, adhering to the quality criteria established in the history of science is essential. As this dissertation is rooted in the social sciences, we will focus on the related quality criteria. However, many of these criteria are also important in the natural sciences. Reliability and validity are cornerstone criteria in social science research and essential for all types of analysis, although there are differences between qualitative and quantitative approaches (Babbie, 2020).

Reliability refers to the consistency of results, which, in text analysis, can be measured by the agreement between different coders (intercoder reliability) or the same coder over time (intracoder reliability) and is often quantified using overlap measures or Krippendorff's Alpha (Scharrer & Ramasubramanian, 2021). Reliability is a prerequisite for validity; without stable and consistent results, accurate or valid measurement is impossible. Validity assesses whether the study measures what it intends to measure (Scharrer & Ramasubramanian, 2021). Various authors propose different types of validity, each representing different perspectives. Some focus on face validity (whether the measure appears logical) or construct validity (whether it aligns with established theoretical concepts). While some studies may prioritise either validity or reliability over the other, both are essential, whether traditional or computational methods are used, or qualitative or quantitative traditions are followed.

Additionally, advancing scientific knowledge relies on reproducibility, transparency, and open science principles (Romney et al., 2015). It is imperative to emphasize the significance of replication in scientific research (Wiggins & Christopherson, 2019). Without replicability, scientific progress is jeopardized, as the veracity and potential for further investigation of findings cannot be verified. This equally applies to manual and automated analyses (Dienlin et al., 2021; Humphreys et al., 2021). An essential aspect of replication is the transparency of the information provided about each study. There has been an increasing demand for the standardization and transparency of reporting on methodologies, particularly in the context of automated methods (Hoyle et al., 2021; Reiss et al., 2022). Transparent reporting enables readers to assess the methodology's rigor and evaluate the results' validity. In recent years, there has been a growing emphasis on the importance of transparency throughout the entire research process, with scholars advocating for adopting open science practices (Lewis, 2020). Open science practices encompass a range of activities, including the publication of materials, data, and code, as well as abiding by the principles of FAIR Data (Wilkinson et al., 2016), where possible; the preregistration of studies and the submission of registered reports; the conduct of replications; the fostering of collaboration; the development of open science skills; the implementation of transparency and openness promotion guidelines; and the incentive for open science practices (Dienlin et al., 2021).

Another crucial aspect closely related to reliability is stability or robustness, which refers to the consistency of (automated) measurements across repeated applications. Stability is essential to ensure that the content analysis results are reliable and reproducible. However, some automated methods are non-deterministic or use black-box algorithms that researchers can neither influence nor fully understand. This lack of transparency and control can result in instability, where running the same analysis multiple times yields different results (Reiss, 2023). To address these issues, researchers have introduced multiverse analysis (Pipal et al., 2023), a method that runs multiple analyses with varying assumptions, parameters, or models. By exploring a range of plausible analytical paths, multiverse analysis increases the credibility of research results and helps researchers understand the potential uncertainty associated with relying on a single model.

All of these quality criteria are of significant importance. They have been developed over time by the scientific community, and numerous studies have highlighted the problems that arise when they are not adhered to (for the topic of this dissertation, see, for example, Baden, Pipal et al. (2022) and Grimmer and Stewart (2013)). While no single criterion is more important than the others, this dissertation focuses on validity. Validity is central to building scientific knowledge and remains less standardized compared to other criteria. Currently, no single test or method is universally predominant across various fields and furthermore, the validation process also varies depending on the method used (Birkenmaier et al., 2023).

2.4 Validity

Validity is a fundamental concept in social scientific research, which refers to the extent to which researchers measure what they intend to accurately. Scharrer and Ramasubramanian (2021) explain this concept by emphasizing its role in accurately capturing the meaning of complex and multifaceted concepts, such as social justice. They compare understanding a concept like 'democracy' to visualizing a vast, ambiguous cloud where no single measure can capture its full complexity. Each researcher's task is to capture as much of this "cloud" as possible in their study. In this context, validity refers to whether readers can reasonably connect their understanding of the concept to how it is measured in the research. Similarly, Babbie (2020) defines validity as the extent to which a measure accurately reflects the true meaning of the concept being studied. However, he highlights a critical issue widely discussed in the social sciences: social scientific concepts do not possess inherent, "real" meanings. As a result, it is widely recognized that determining whether a measure perfectly captures a concept's meaning is inherently challenging, if not impossible.

In the field of philosophy of science, this issue is closely related to the broader question of what constitutes the truth. Baškarada and Koronios (2018, p.2) writes that "Without a clear understanding of the philosophical underpinnings, logically deriving applicable validity criteria becomes difficult (if not impossible)." The philosophy of science distinguishes between two views on science and the truth. Correspondence views assess the alignment between theory and fact, while coherence views evaluate the consistency among

beliefs and their connection to social structures. Many scholars in the social sciences agree that truth is relative to perspective, thus honoring both views of what truth in science can be, and also raising questions about objectivity and the identity of those conducting science (Godfrey-Smith, 2009; Scharrer & Ramasubramanian, 2021). The social sciences have long debated the importance and/or (im-)possibility of objectivity. Philosophical discussions also consider the theory-ladenness of observation, which emphasizes how theories shape research processes and findings, thereby again influencing subsequent theories (Godfrey-Smith, 2009). Consequently, determining whether scientists accurately measure their intended concepts and uncover their true meanings is both complex and crucial. As a result, scientists have adopted various perspectives on how to approach validity.

Moreover, there has been some discussion on the different understandings of validity in qualitative versus quantitative studies (Baškarada & Koronios, 2018; Winter, 2000), which showcases that validity is not the same for each approach, nor for each methodology, and it also depends on the kind of truth researchers aim to measure. Quantitative research is rooted in the positivist tradition, emphasizing empirical conceptions. Validity in this context is closely tied to systematic theories and the generalizability of findings. In contrast, qualitative research comes from post-positivist traditions, focusing more on individuals' or groups' meanings and personal experiences. It views reality as subjective and constructed through interactions and personal accounts. In quantitative research, researchers often strive to disassociate themselves from the research process to maintain objectivity, as involvement is seen to reduce validity. Conversely, qualitative researchers embrace their involvement, considering it essential for understanding the phenomena. In qualitative research, denying this involvement is seen as a threat to validity. Approaches to validity differ: quantitative research emphasizes internal and external validity, while qualitative research focuses on the depth and richness of data, where validity is often reframed as trustworthiness, credibility, or plausibility. Evaluation and interpretation highlight further differences. Quantitative research tries to apply standardized tests and methods. In contrast, qualitative research recognizes the inevitability of interpretation and the subjective nature of data collection, viewing researchers' perspectives and interactions as an integral part of the research process. Despite these differences, there are also notable similarities between the two methodologies. Both are concerned about the appropriateness of the research process and its relation to the phenomena under investigation. Additionally, both acknowledge the need for some form of validity or credibility in their findings (Humphreys et al., 2021; Kirk & Miller, 1986; Winter, 2000).

Given these differing conceptions and understandings, it may be essential to delve deeper into the discussion of validity in the field of content analysis, and computational text analysis in particular. Krippendorff (2013) highlights the unique challenges of content analysis regarding validation, emphasizing the importance of context sensitivity. Next to face validity, which he defines similarly to others before him, he also focuses on social validity, which assesses whether a study addresses a social scientific problem. Most importantly, however, he outlines three main types of evidence-based validity for content analysis. First, evidence that justifies the treatment of texts includes sampling

validity, which ensures the sample represents the population accurately, and semantic validity, which ensures the analysis categories align with the text's contextual meanings. Second, evidence for abductive inferences includes structural validity, which checks the correspondence between data and analytical constructs, and functional validity, which verifies that the analysis works as intended, similar to successful past analyses. Lastly, evidence justifying the results involves correlative validity, which assesses whether findings correlate with those from other valid methods, and predictive validity, which assesses the accuracy of predictions based on the analysis.

In one of the first landmark articles on using computational analyses in social science, Grimmer and Stewart (2013) clearly positioned the importance of validation—understood as the practice of achieving validity—as one of their four main principles of quantitative text analysis. Nevertheless, computational methods in the social sciences have had a difficult history with validation. Baden, Pipal et al. (2022, p.1) attest, among others, a validation gap when using computational text analysis for social sciences. Birkenmaier et al. (2023, p.1) further support this notion and note that although "researchers apply a great variety of validation steps, [...], [they] however, are rarely selected based on a unified understanding of validity." As described above, computational methods are diverse, and while validation for supervised or semi-supervised methods seems more straightforward (e.g., comparing the results of the automated analysis to a manually created gold standard, see, for example, Song et al. (2020)), the validation for unsupervised methods is more complicated and this problem also extends beyond using text as data (James et al., 2013). The importance of validation has sparked a considerable interest in the field, leading to many publications on validation, either in general (Baden, Pipal et al., 2022; Birkenmaier et al., 2023; Lind et al., 2023; Song et al., 2020) or on specific methods, such as topic modeling (Chen et al., 2023; DiMaggio, 2015; Maier et al., 2018; Ying et al., 2022) or word embeddings (Antoniak & Mimno, 2018; Faruqui et al., 2016; Gladkova & Drozd, 2016; Rodriguez & Spirling, 2022).

In summary, the fundamental understanding of science is closely linked to the methodologies used by researchers, be they qualitative or quantitative approaches. This chapter places particular emphasis on text analysis methods, both manual and automated, and highlights the importance of maintaining methodological rigor and ensuring validity, particularly in the developing field of computational analysis. Key criteria are examined to demonstrate their essential role in maintaining the credibility of scientific findings, particularly in social science research. In the context of unsupervised text analysis, robust validation procedures become even more important. Unlike supervised methods, which can be evaluated against known benchmarks, unsupervised methods operate in a more exploratory space, making it difficult to assess their accuracy and validity. This highlights the need for continued development and careful application of validation techniques to support progress in this emerging field. As computational methods in the social sciences continue to evolve, the refinement of validation practices will remain critical to ensuring the integrity and value of research.

3 Research Questions

It is crucial to highlight the difficulty in assessing validity, particularly in unsupervised settings, as researchers cannot fully observe some aspects of the research process. This makes validation even more crucial to ensure the accuracy and credibility of findings. Without rigorous validation, the risk of generating unreliable or misleading results increases, undermining the overall research objectives. It is important to note that the three studies of this dissertation focus specifically on the two methods explained above: word embeddings and topic modeling. By focusing on these methods, the studies aim to address specific validation challenges and contribute to developing best practices in computational text analysis. These studies are unified by a central research question: how can unsupervised computational text analysis methods be evaluated to produce more valid results? They investigate the role of hyperparameter settings in word embeddings, review current validation practices in topic modeling, and evaluate how different validation strategies influence model selection. Furthermore, a common goal of these studies is to examine and critically reflect on how validation is currently performed in the field. By identifying existing gaps and shortcomings, the objective is to formulate best practices that ensure the rigor and reproducibility of computational text analysis. Collectively, these efforts aim to enhance the credibility and accuracy of research findings by providing precise and actionable guidelines for social science researchers.

The first study (Bernhard-Harrer et al., 2024) investigates the validation of word embedding models by assessing the influence of hyperparameter settings on the performance and stability of word embedding models trained on a large Austrian news media corpus. It seeks to answer the research question: *How do different hyperparameters affect model performance across various evaluation tasks?* The study evaluates models on intrinsic and extrinsic tasks, including semantic and syntactic tasks, author classification, topic classification, and sentiment analysis, to determine how different hyperparameters affect model results. The study highlights the importance of thorough validation and transparency in model selection and hyperparameter tuning to ensure reliable and meaningful results. It aims to improve the understanding of best practices in word embedding validation in the social sciences.

The second study (Bernhard et al., 2024) presents a systematic review of validation methods used in topic modeling for social scientific research questions. It seeks to answer the research question: *Is there a convergence towards a gold standard of validation methods for topic modeling?* By analyzing 792 studies, it aims to determine how validation practices have been developed and applied over time. This comprehensive review maps the existing landscape of validation methods, highlighting the diversity and prevalence of different approaches over time. Overall, it does not find signs pointing towards a

convergence over the past decade. The study aims to provide an overview for researchers and encourage adopting more qualitative validation practices to enhance the credibility of topic modeling studies.

The third study (Bernhard et al., 2023) aims to evaluate the impact of different validation methods on topic model selection. It seeks to answer the research question: *How does the choice of validation method affect the outcomes of topic modeling?* The study applies two different topic modeling algorithms to the same text corpus and uses four different validation strategies to understand how these methods influence model selection and evaluation. It shows that the choice of validation method can influence, which model is seen as performing "best", and showcases that this has a large impact on the topics found in a text corpus. Thus, the study highlights the need for objective and transparent validation practices to minimize researcher bias and ensure valid results in topic modeling, ultimately contributing to more robust theory development and practical applications in the field.

4 Studies

4.1 Comprehensive Validation of Word Embeddings for Social Science Research

Comprehensive Validation of Word Embedding Models for Social Science Research

Jana Bernhard-Harrer, Paul Balluff, and Ahrabhi Kathirgamalingam
University of Vienna

Abstract

In computational communication science, accurate representation of text as data is critical. Traditional off-the-shelf models often face challenges such as contextual misalignment, insufficient validation documentation, and ethical concerns regarding training data. To address these issues, we create specialized word embeddings tailored to our dataset of over 5 million articles from nine media outlets in Austria spanning the years 2010-2022. Using fastText for embedding training, we incorporate sub-word level information and perform extensive hyperparameter searches to assess their impact on performance. Each hyperparameter combination is evaluated ten times to ensure model stability and robustness. We rigorously validate our models using both intrinsic (semantic and syntactic tasks) and extrinsic (author classification, topic classification, sentiment analysis) methods. We illustrate the impact of model choice through a use case focused on gender bias, showing how different models can yield different results when analyzing media coverage of sensitive social issues. This research addresses the question: How do different hyperparameters affect model performance across various evaluation tasks? It highlights the need for thorough validation processes in the development of word embeddings, thereby improving the accuracy and reliability of computational text analysis in the social sciences.

Keywords: Word Embedding, Text-As-Data, Validation

Content Warning: The following study contains a section related to femicide towards the end of the results, where it is used as a substantive example to showcase the impact of model selection. This content may be triggering or upsetting to some readers. Please proceed with caution.

Introduction

When computational communication science uses text as data, one of the main steps of any research process is to represent words as numbers so mathematical calculations are possible. This transformation of words in the corpus into vectors can be done in a variety of ways, from simply counting word frequencies in documents to more complex ways of embedding words in an n-dimensional vector space. Throughout the last decade of com-

putational research in communication science, how we turn text into numbers has gone through several stages, with each new technological advance promising *better* models for scholarly research (Grimmer et al., 2022). After initial attempts at counting words, including various weighting strategies, many scholars shifted to vector representations (Mikolov et al., 2013). For the latter, scientists often employ off-the-shelf language models, such as BERT, which have shown to be extremely useful for various problems in the social sciences (Rodriguez & Spirling, 2022; Rudkowsky et al., 2018; van der Veen, 2023).

However, in pursuit of linguistic precision and domain-specific relevance, the goal of developing a custom word embedding model is justified for social scientific research questions. Generic off-the-shelf models often fall short of accommodating the intricacies of a domain-specific language, whether it involves the specialized terminology of a specific industry or the nuanced characteristics inherent in a given dataset. Additionally, another significant drawback of using pre-trained models is the lack of transparency into their training processes, which means researchers have limited influence over crucial parameters, such as whether the model understands casing. Constructing embeddings independently tailored to the peculiarities of the data not only facilitates a more accurate representation but also allows for the adjustment of training parameters to align with the task at hand. This research aims to develop such a domain-specific model for text analysis in communication science, and discuss which problems might arise with validation and application of these models.

Word embeddings are regularly used in communication science, particularly in political communication, including in Austria (Haselmayer et al., 2022; Rheault & Cochrane, 2020; Rudkowsky et al., 2018). Word embeddings include both language and context. Language uses the words from the training corpus, and context reflects how these words are used together. While written German in Austria is similar to the one in Germany, the context differs between the two countries. In the context of Austrian newspaper articles, a custom word embedding model could effectively capture the domain-specific language, providing a nuanced understanding of regional topics, sentiments, and relationships within the Austrian media landscape, as compared to off-the-shelf models, which are trained on German text, including German, Austrian, and Swiss contexts. This is significant for communication science, where precise analysis of communication patterns is essential. We, thus, propose to develop an Austria-specific word embedding model that adheres rigorously to scientific standards at every stage, ensuring that all processes are transparent and replicable. Additionally, to promote open science, we will make the model and all related outcomes freely available to other researchers. Furthermore, to support reputability and collaborative advancement, all the associated code is openly available on GitHub [1]. Our study is guided by the following research questions: How do different hyperparameters affect model performance across various evaluation tasks? Our findings are discussed within the broader discussion of validation in automated approaches to content analysis. This is crucial as it helps us understand the broader implications of model selection and the influence of hyperparameter settings on the outcomes of computational analyses in communication science.

We show that model choice impacts substantive results, emphasizing the impor-

[1] https://osf.io/gxzvs/?view_only=6df9df425b6c42828539ac5193a98098

tance of validating word embeddings. We compare our custom model's performance across different hyperparameter combinations and against off-the-shelf models, using a use case focused on gender bias to explore differences in semantic representation of related concepts. Overall, there is no clear road map to validating word embeddings, and how to choose the "best" model. As different evaluation methods favor different "best" models, validation is a complex task and needs to be approached closely tailored to each individual research question.

Literature Review and Related Work

Word Embedding Models in Social Science

Word embeddings are created to numerically represent text as data. For this, words are translated into a continuous vector space, where the distance between these word vectors corresponds to the semantic similarity of words. During the training of these embeddings, different parameters can be chosen, such as the word window, the minimum count of words in the training data, or the consideration of casing. These changes in training parameters or the training corpus can lead to different models that understand words in slightly different ways, for instance, impacting the nearest neighbors of a specific target word. Word embeddings are used in research in two different ways. First, they can be used solely as an instrument to embed the text data numerically, for further downstream tasks (Antoniak & Mimno, 2018). Second, they can be treated as a distinct data source to infer, for example, how words relate to each other to study semantic meanings, which offers enormous potential to study implicit associations and biases in our language (e.g., Garg et al., 2018; Kroon et al., 2020). In both applications, it is crucial for researchers to ensure that their models not only accurately capture the semantic meanings of words, but that they do so specifically within the particular context of their study, thereby ensuring that the findings and analyses are truly relevant and valid.

Researchers can either train their own word embedding model, or use large pre-trained models. The use of off-the-shelf language models can be problematic for different reasons: First, the amount of text needed for training these embeddings makes it difficult to rely solely on contextually related texts for training and testing. For example, the German BERT is trained on Wikipedia, which does not differentiate between the regional differences in language and culture between Germany, Austria, and Switzerland (B. Chan et al., 2020). Second, these models are often developed and published by major industry players, such as Google or Meta. Thus, researchers have to rely on the provision of these models and do not have influence over which information on model training and specifics is provided, which training data has been used, and how reliably the model will be provided for replication. Thus, for example, it is not possible to decide whether the model takes into account casing or not. Third, as often with the use of off-the-shelf tools, it is difficult to assess the quality of the word embedding for one owns study in particular or implement other hyperparameter settings that are more suitable for the downstream task in question (Antoniak & Mimno, 2018). Nevertheless, the use of off-the-shelf models does have advantages; on the one hand, it reduces the resources needed for the creation of word embeddings, both in the case of data availability and computational power, and on the other hand the sheer size of these models can lead to potentially broad application cases.

Quality Criteria for Word Embedding Models

While there are many quality criteria in science, in the specific case of word embedding models, we want to emphasize the importance of stability, validity, and transparency.

A prerequisite of a valid model is to have a *stable* model, that leads to robust results. Rodriguez and Spirling (2022) assess the stability of word embedding models over different corpora sizes and languages and find that the off-the-shelf models they compare are quite robust over these variations. However, Antoniak and Mimno (2018) find that the individual word vectors are quite sensitive to hyperparameter changes during the training process, thus arguing that researchers should not rely on a single model for their analyses. Before starting the validation process of the word embedding model, researchers should make sure that their model specifics are fairly stable. Both studies detail approaches to assessing the stability of models. While Antonaik and Mimno (2018) rely on calculating distance metrics between word vectors and their nearest neighbors, Rodriguez and Spirling (Rodriguez and Spirling, 2022) assess the Pearson correlation between different cue words. Both approaches allow for a highly contextualized approach, that can be tailored to the substantive goal of a study.

Validation is an important step in science in general, but especially when researchers apply automated tools that leave much of the data unseen by them (Grimmer et al., 2022). It has been argued elsewhere (Baden et al., 2022) that some computational studies show a lack of validation practices and thus lack an important cornerstone of scientific knowledge production. Validation can be defined as being about the "accuracy and scientific veracity of measures and, by that, of research results and downstream conclusions and recommendations" (Bernhard et al., 2022). Validity is essential to the accuracy and soundness of the analytical outcomes, ensuring that the tool measures what it purports to measure. All steps regarding the validation process must be reported, otherwise, it is not possible to judge the scientific rigor of the study or know how the researchers came to their conclusions.

Thus, transparency also emerges as a crucial concept in the development and utilization of any text analysis tool. Transparency for using word embeddings can be achieved by being open about which model is used, either off-the-shelf or self-trained. If an off-the-shelf model is used, researchers can describe how they arrived at their choice, where the model can be found, and how they adjusted the model to fit into their study context. In the case that researchers decide to train their own models, transparency should start with the question as to why a new model is needed (Bender et al., 2021), with which data it was trained, which hyperparameters were chosen, and how the researchers arrived at a final model choice, in case they performed a grid search over multiple hyperparameter settings, as well as how they validated their embedding model. Romney and colleagues (2015, p.32) emphasize the importance of transparency in replication, but also that it "helps to communicate the essential procedures of new methods to the broader research community. Thus, transparency also plays a didactic role and makes results more interpretable."

Together, stability, validity, and transparency serve as cornerstones for establishing the credibility and utility of text analysis tools, promoting responsible and effective research practices in the field of communication science. This study will specifically focus on vali-

dation methods for word embeddings, exploring how different hyperparameter settings for embedding models perform on different evaluation tasks, and discussing what this means for model validation.

Validating Word Embeddings

The aforementioned advantages of off-the-shelf models should not lead researchers to bypass sound validation practices. Especially as word embedding models are increasingly adopted in the social sciences for substantive research, social scientists run the risk—as common for transferring methods from one field to another—of adoption outpacing understanding, which can threaten, among others, validity (Baden et al., 2022; Rodriguez & Spirling, 2022). While off-the-shelf embeddings seem to work rather well (Rodriguez & Spirling, 2022), researchers still risk applying models for which they do not know what data was used and what training decisions were made. Such off-the-shelf embeddings are often offered by large, profit-oriented companies, thus further jeopardizing the researchers' autonomy and posing ethical questions. Moreover, when using off-the-shelf models, researchers assume that these off-the-shelf models will also work for them, without incorporating the needed validation checks to be sure of this claim. However, this general effectiveness and ease should not lead researchers to bypass the essential steps of validation and contextual verification.

Validating word embedding models is not trivial, and researchers in the computational sciences have been thinking about this for quite some time. In 2016, the *1st Workshop on Evaluating Vector-Space Representations for NLP* started discussions on how researchers can evaluate vector spaces that are trained for natural language processing (NLP) tasks. Although the idea behind word embeddings is, that they are trained as general tools, each study that applied a word embedding model, does have a specific task the model needs to perform. Thus, it is imperative for researchers to validate to what extent their models are accurately capturing what they intend to measure. The most notable distinction of validation methods is between *intrinsic* and *extrinsic* validation methods. The former checks whether the model can capture language, which is often measured as similarities between words. The latter assesses how well the embedding works for specific downstream tasks.

Intrinsic Validation

The goal of intrinsic evaluation methods is to measure in how far the model can *understand* syntactic or semantic meanings of a language. This is measured by applying different tasks asking the model to find similar or dissimilar words. While there are some other types of intrinsic tasks to evaluate word embeddings (Chiu et al., 2016), the following four are among the most commonly used:

1. Similarity scores based on manually annotated datasets: This involves querying the model to calculate the distance between two word vectors and comparing the result to a manually coded dataset of similar word pairs to see if the model matches human assessments. Despite its intuitive nature, this task faces criticism regarding the quality of gold standards, the definition of "similarity" versus relatedness, polysemous words, and frequency effects of central words in the model (Avraham & Goldberg, 2016;

Batchkarov et al., 2016; Faruqui et al., 2016).

2. Finding intruder words in manually created tasks: This task takes a similar idea but presents it inversely, by querying the model to identify which word on a list of words does not belong to the others. This eases the burden of creating a gold standard that finds similarity pairs, as only dissimilarity has to be agreed on, which is easier for humans to do (Camacho-Collados & Navigli, 2016).

3. Finding analogies instead of words: Some scholars have opted away from predicting another word vector and are looking more closely at the line connecting two word vectors. This is often used to compare grammatical use, as the line leading from "is" to "been" should be close to the line from "have" to "had". However, it can also predict relationships outside of grammar such as connecting capitals or currencies with countries (Che et al., 2017; Liza & Grześ, 2016; Schnabel et al., 2015).

4. Human Evaluation: In an effort to come up with more sophisticated intrinsic validation methods, some scholars have suggested querying the model to find words that are close to each other and then asking humans for their judgment afterward, as to how related these two words are (Gladkova & Drozd, 2016; Rodriguez & Spirling, 2022).

Researchers have criticized these tasks on various grounds, such as low manual agreement, difficulties in defining similarity or relatedndess, or not taking into account polysemy (Faruqui et al., 2016) and Schnabel and colleagues (2015) have shown that the performance of a model in intrinsic tasks does not predict how well it will perform in downstream tasks of interest. This goes so far that Faruqui et al. (2016, p. 33) write that "until a better solution is found for intrinsic evaluation of word vectors, we suggest task-specific evaluation: word vector models should be compared on how well they can perform on a downstream NLP task." Nevertheless, these methods remain frequently used (Che et al., 2017) and Gladkova and Drozd (2016, p. 40) add the following two thoughts for everyone using intrinsic validation methods: "None of these above-mentioned characteristics of word embeddings provides a one-number answer about how 'good' a model is. But we can take a more exploratory approach, identifying the properties of a model rather than aiming to establish its superiority to others. [...] Lastly, when evaluating word embeddings, we should not forget that the result of any evaluation is down to not only the embedding itself, but also the test, the corpus, and the method of identifying particular relations." However, these intrinsic tasks are highly formalized and can be implemented rather easily, and thus serve as good points of comparison between models. Additionally, the question arises, in how far social scientists should apply word embedding models that fail at "easy" tasks such as identifying an outlier and what it means for models to not be able to solve analogies.

Extrinsic Validation

The second part of possible validation methods is subsumed as *extrinsic* evaluation, which describes practices of validating embedding models by applying them to downstream NLP tasks and measuring how well they fare under specific circumstances. This "in vivo" (Nayak et al., 2016, p. 19) approach is argued to be an advancement as compared to intrinsic validation on two ends: a) the improvement on multiple downstream tasks is a

better predictor of model performance than intrinsic tests and b) higher fidelity in the assessment of strengths and weaknesses of the model. With this, it becomes possible to compare models to benchmarks and look at their relative gains.

However, this kind of validation is not without its own drawbacks. First, implementing various downstream tasks can be time-consuming and again raises the question of which gold standards we use to do so. Gold standards are notoriously difficult to obtain, and off-the-shelf datasets rarely fit the exact context that social scientists have in their study, especially when the language of the model is not English. Additionally, researchers need to be aware that evaluating a model on the task it is supposed to perform can lead to overfitting. Lastly, it is again difficult to generalize from a model's score in one downstream task to other tasks, other languages, etc. (Schnabel et al., 2015; Seyed, 2016).

Embedding validation in the social sciences

As mentioned above, this discussion on evaluating and validating models has not yet been translated widely into the social sciences. While some authors (Grimmer et al., 2022; Rodriguez & Spirling, 2022) discuss word embedding validation for social science, in substantive studies that employ word embeddings, many researchers do not include much information about the embedding they used.[2] In general, many authors validate their models extrinsically by applying them to the specific tasks in question and then comparing the results with a manual gold standard (**kroon_beyond_2022**; Viehmann et al., 2023). Others look at specific relevant words that are central to the study (Liang et al., 2023) to gauge face validity or use off-the-shelf models while specifically highlighting how it was validated in the first place (Buehling, 2023; van Atteveldt et al., 2021). This gap, whether in reporting or in validating itself, is problematic, as only valid text analysis tools provide researchers with confidence in the precision and relevance of their findings.

We, thus, want to contribute to bridging the discussion from computer sciences to the computational social sciences by applying a rigorous validation scheme to a set of self-trained Austrian word embeddings. This allows us to show how hyperparameter settings impact model performance and how the choice of validation method can impact the final model selection and, thus, the decision on which tool is being applied in substantive research.

Method & Data

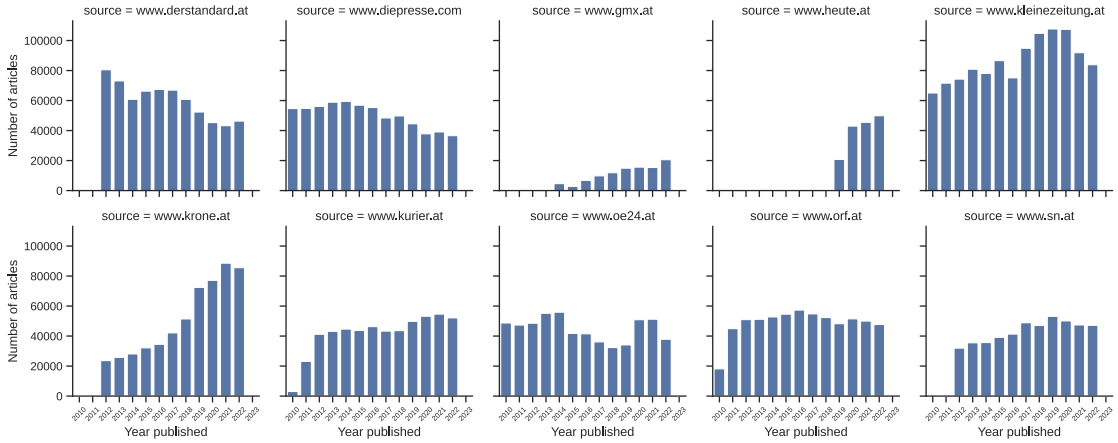
The training data consists of online news media articles from nine different Austrian media outlets. The full text of all accessible outlets was scraped retrospectively from the outlet's websites and in line with Austrian Copyright Laws (§ 42h öUrhG). This resulted in a dataset of 5,495,185 articles from 2010 to 2022. For a more detailed documentation of the data composition, see Table 5 and Figure 1.

All articles were pre-processed at the paragraph level. In line with previous work (Bojanowski et al., 2017), we first split every article into paragraphs, and then only kept unique paragraphs. Second, we discarded all non-text characters (i.e. punctuation, emojis, numbers, and HTML fragments) as well as hyperlinks and email addresses, and normalized

[2]For notable exceptions, please see (C. H. Chan et al., 2020; Kroon et al., 2024; Schuld et al., 2023; Viehmann et al., 2023)

Figure 1

Overview of Sources included in the Training Data



different spelling practices regarding the gendered aspect of the German language, resulting in a total training corpus comprised of 1.7 billion words.

We use the *fastText* software library to train our embeddings (Bojanowski et al., 2017), instead of the often-used *word2vec* (Mikolov et al., 2013) or *GloVe* (Pennington et al., 2014), as it learns each word on a sub-word-level, thus being especially useful for embedding compound words, which are extremely prominent in the German language. To find the most useful hyperparameter combination, we vary the following three hyperparameters:[3]

- cased vs. uncased: German is a language in which the casing of words can impact the meaning. However, including the casing increases the size of the model considerably. Thus, we train the models with and without casing.

- minimum word count: the number of times a word must be present in the training data to be included in the model. Very rare words (such as spelling mistakes) are discarded so that the model is a) smaller and b) not over-fitted towards rare words. We train our models with minimum occurrences of 5 (default), 10, 50, and 100.

- window size: The window size denotes how many words before and after the target word are taken into account for embedding the target word. Rodriguez and Spirling (2022) found that the model's performances for English do not increase substantially when using window sizes larger than six. As German sentences tend to be longer than English sentences, we will train the models with window sizes of 5 (default), 6, 12, and 24.

[3]Strictly speaking, only "window size" is a hyperparameter, the other to factors (casing and minimum word count) can also be understood as pre-processing steps. However, as the discussion in the field often denotes all of these operations as hyperparameter settings, we will continue with this terminology.

Stability

As introduced before, Antoniak and Mimno (2018) showed that word embeddings are highly sensitive to the training corpus as well as to random variations due to the nature of the embedding process. We, thus, estimate each hyperparameter combination ten times, which allows us to measure the stability of the word vectors between model variations and to assess the robustness of each hyperparameter setting as well as the impact of the corpus on the model stability (Lai et al., 2015). In the end, we trained 320 models (32 different hyperparameter combinations, which were estimated ten times), which we then evaluated both intrinsically and extrinsically.

Intrinsic Validation: Semantic & Syntactic Tasks

Our tasks on *intrinsic* validity are based on semantic and syntactic pairs as presented by Müller (2015). This evaluation is based on the notion that "the geometric relationship between these vectors captures meaningful semantic relationships between the corresponding words." (Garg et al., 2018, p. 3635). Regarding semantics, we query the model to find opposite words (e.g., big–small), to provide the best matches in a series of relationships (such as "France is related to Paris, as Italy is related to..."), as well as to identify intrusion words (e.g., "apple, orange, banana, rabbit"). Syntactic, on the other hand, builds on grammatical similarities, and we thus query the model to give us the plural form of a noun or the superlative of an adjective. This first evaluation step aims at measuring how well the model "understands" the semantic relationship as well as the grammatical rules of the training corpus.

Extrinsic Validation: Downstream Tasks

As stated above, word embeddings are often only one step in a computational method, and the value of a particular model can only ever be as good as its performance on the task it is supposed to perform. To assess *extrinsic* validity we apply the models to the following three different downstream tasks for which gold-standard data are available:

- Author Classification (single label classification task): As suggested by Grimmer et al. (2022), computational methods can be validated by employing a fictitious prediction problem by having the method solve a problem to which we already know the answer. In our case, the models are tasked with predicting the author of different political texts (Facebook posts, tweets, press releases, and parliamentary speeches) from five Austrian parties.
- Topic Classification (multi-label classification task): Another widely applied downstream task is the classification of texts into different topics. We use the manual text classification put forward as part of the *Austrian Election Study* 2017 and 2019 (AUTNES) as training data for a topic classifier (Litvyak et al., 2022).
- Sentiment Analysis (single label classification task): The third downstream task is the annotation of text with sentiment scores. We use the gold standard created as part of a research project by the first two authors (masked for review, 2023) to evaluate how well the model can assess the sentiment (positive-negative) of a given sentence.

Table 1

Reference models for comparison

Model	Author	Training Corpus	Model Type	Dimensions	Vocabulary Size
Facebook Embeddings	Grave et al., 2018	Common Crawl & Wikipedia (German)	Fasttext (window size: 5)	300	2,000,000
DistilBERT-Multilingual Cased	Yuan, 2023	Wikipedia	BERT	768	119,547
GBERT-Base	B. Chan et al., 2020	OSCAR (CC)	BERT	768	31,102
GottBERT-Base	Scheible et al., 2020	OSCAR (CC)	BERT	768	52,009
XLM-RoBERTa-Base	Conneau et al., 2020	Common Crawl	BERT	768	250,002

In addition to comparing the models (hyperparameter combinations) to each other on both intrinsic and extrinsic tasks, we also compare how well our models perform versus off-the-shelf models (see Table 1). We do this by running multiple multilevel regressions to distill the impact of hyperparameter settings on the models' individual performance on all validation tasks.

Practical Example

To further highlight the substantive impact that model choice can have on the results of a substantive study, we use the approach that Antoniak and Mimno have used (2018) in assessing the differences in nearest neighbors for the different models. The nearest neighbor approach in word embedding models involves identifying n words that are closest in vector space to a given target word. By measuring the distance between vectors, we can determine which words are most similar to the target word. We chose an application case, which has been studied with word embeddings before: gender bias (Garg et al., 2018; Kroon et al., 2020). We identify the nearest neighbors for two terms, one term that is more generic, and used in many contexts—"Frau" (Women) and another, less frequent term "Femizid"[4] to understand how the model represents related concepts such as gender, violence, crime, and societal response. The selection of the second words is particularly motivated, by its pressing relevance in Austria (Europäisches Institut für Gleichstellungsfragen, 2022), and as understanding media narratives on such sensitive issues is crucial to highlighting societal problems and promoting gender equality (Aldrete et al., 2024). Comparing a more frequent, and widely used word with a word less frequent word, denoting a more complex construct, helps us assess the semantic accuracy and relevance of the word embeddings generated by the model, when being applied in a social scientific context.

Results

Our analysis of the self-trained models provides a detailed performance evaluation across different tasks, highlighting the impact of different hyperparameter settings. An overview of all results for the self-trained models can be found in Table 2. For comparison,

[4]Femicides, or the gender-based killing of females (The official statistics of Austria, do not count the murder of trans, intergender, and non-binary people, as well as underaged females as femicides. Thus underreporting the scope of gender-based violence.) represent a major social problem with significant implications for society. In recent years, Austria has faced a troubling increase in femicides, as compared to other EU countries, with a significant number of women being murdered, and like many other countries, has struggled to address and reduce the number of femicides.

the performance data for the off-the-shelf models is summarized in Table 3 in the Appendix. This section first addresses how different hyperparameters affect model performance across various evaluation tasks and in the end focuses on how the choice of validation task impacts model choice.

Stability

Analogous to (Rodriguez & Spirling, 2022) we assess the stability of our models, by looking at correlations *within* "model families", that is models from the same run that share the same hyperparameters. For this, we use different pre-defined sets of words (from the area of politics, culture, economy, environment, migration, social groups, and sports),[5] and one random set of 100 words to assess the stability of our models across a variation of contexts. We calculated the cosine distance for each word embedding against every other word embedding in the model. We take the distance measures between the cue words in two models and compare them pairwise by calculating Pearson's Rho.

Figure 2 shows a rather consistent picture of lower correlations with growing window sizes.[6] This relationship is not found by (Rodriguez & Spirling, 2022) for English models. We attribute this different finding to the grammatical structure in German, where verbs can be at the end of sentences, which are also on average longer, and window size changeing in which contexts words are understood. We find that larger window sizes as well as lowercased models are less stable, as they include more outliers. Overall, all models seem rather robust over the different cues and contexts, with larger margins for the random cues, as these contained a larger set of words.

The second measure for stability is the correlation across model families, which we calculate in a similar fashion to the within correlations. However, in this case, we take all cosine distances from models with the same model family and calculate their mean. This means that we get one representation for the entire model family. We then use this mean representation to pairwise compare it to other model families (Rodriguez & Spirling, 2022). Note that we do not run this comparison across lower-cased and cased models because they do not share the same vocabulary. The results are shown in Figure 3, where we see that the correlations across model families are rather high and stable. The largest impact is due to the window size parameter (see Figures 3 and 11). However, when we compare the correlations with the Facebook model, we can see that the values are much lower and the variations are higher depending on the set of cues.

Intrinsic validation

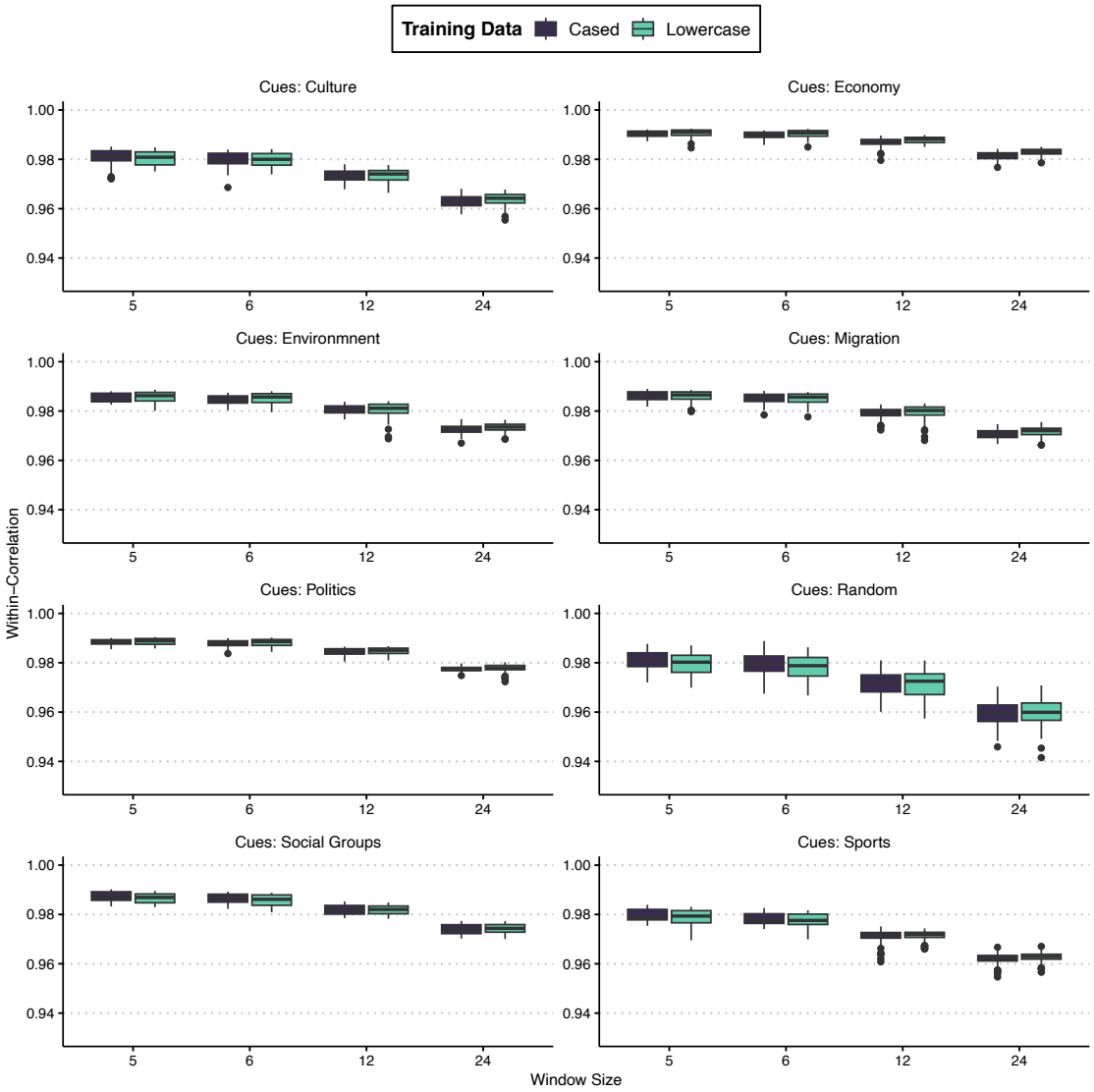
To evaluate the intrinsic validation of our models we have posed three semantic and one syntactic task, which were proposed by Müller (2015). All our models achieved full coverage on all semantic tasks, which indicates that they have a sufficiently large vocabulary. However, some models only achieved coverage of 60–70% on syntactic tasks. Figure 4 shows an overview of all task results. The cased models tend to perform better on the syntactic

[5]Please see Table 4 in the Appendix for a full list of words that were included, as well as graphical representations for all correlations

[6]The impact of minimum count can be seen in Figure 11 in the Appendix

Figure 2

Within correlations of all pairwise combinations which is our measure for the stability within model families. Illustrating the impact of lowercasing as well as window size on the robustness of models.



tasks. The Facebook embeddings generally perform best across all tasks. BERT-like models, however, have the worst performance.

best match

The different model families vary greatly in how well they perform the best match task (see Figure 5), while the lower bound revolves around correctly identifying 42% of the relationships, the better models fare around 62% of correct answers, which is a stark difference.

We find, that lowercase models tend to perform better, however, this difference diminishes as the minimum word count in the model increases. In general, the models with a higher minimum word count perform much better, while on the opposite side, larger window sizes hurt the model performance in identifying semantic relationships. A multilevel linear regression shows that all three hyperparameters have a significant impact on performance (see Table 2 and Figure 10 for details).

By comparing to off-the-shelf models, we find that BERT-based off-the-shelf models perform rather poorly on this task, ranging from only getting 5% to 15% of the questions right, which is well below all our model families scores (see Figure 10 and 3). The Facebook model trained on the common crawl model outperforms all of our 320 models, with a score of 76%.[7]

opposite

We can see in Figure 6 that all models perform rather poorly in this task, with performance ranging from correctly identifying 5% to 18% of the questions. As explained above, there is valid criticism against these kinds of intrinsic methods, which assume a clear gold standard of opposite words, which is something that is, in many cases, not as clear as it might seem. Nevertheless, the differences between the models are independent from this criticism and we find that cased models perform slightly better than their lowercase counterparts. Regarding the minimum count, we find that pruning the vocabulary leads to better performance, especially when cased words are included. Increasing the window size hurts performance again, especially again for the cased models. The Bayesian multilevel regression also showed that lowercasing is the only significant predictor of the models' performance in finding the correct opposite word (see Table 2 and Figure 10 in the Appendix).

We find that except for the multilingual BERT, all off-the-shelf models easily outperform our models in this task, correctly identifying more than a third of the questions, and the GBERT base model even half of the questions correctly.

word intrusion

All of our models perform rather well in this task, with averages above 85% for almost all model families. Overall, Figure 7 shows that the case does not impact performance

[7]For the off-the-shelf models correctly identified tasks were calculated by dividing the number of correct answers by the number of words that were covered. However, as some queries include words that are not part of these models, these questions were not taken into account. Thus, the correctly identified answers are measured only against the questions the models could fully *understand*.

Figure 3

Correlations across model families. Common Crawl (Facebook) correlations with our Cased models

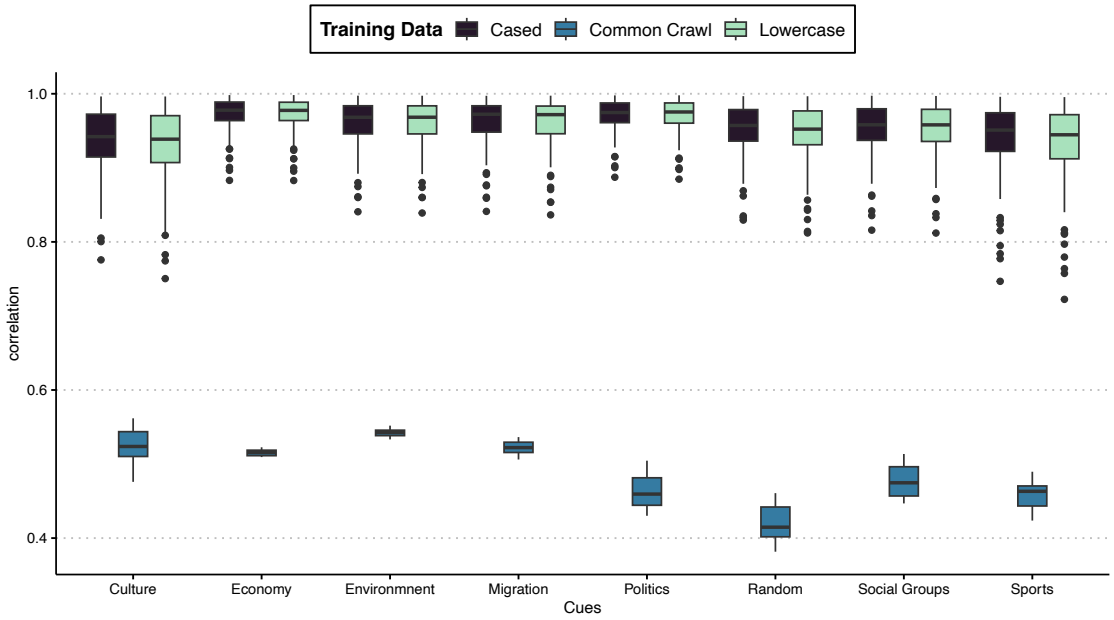


Figure 4

Results of semantic and syntactic tasks.

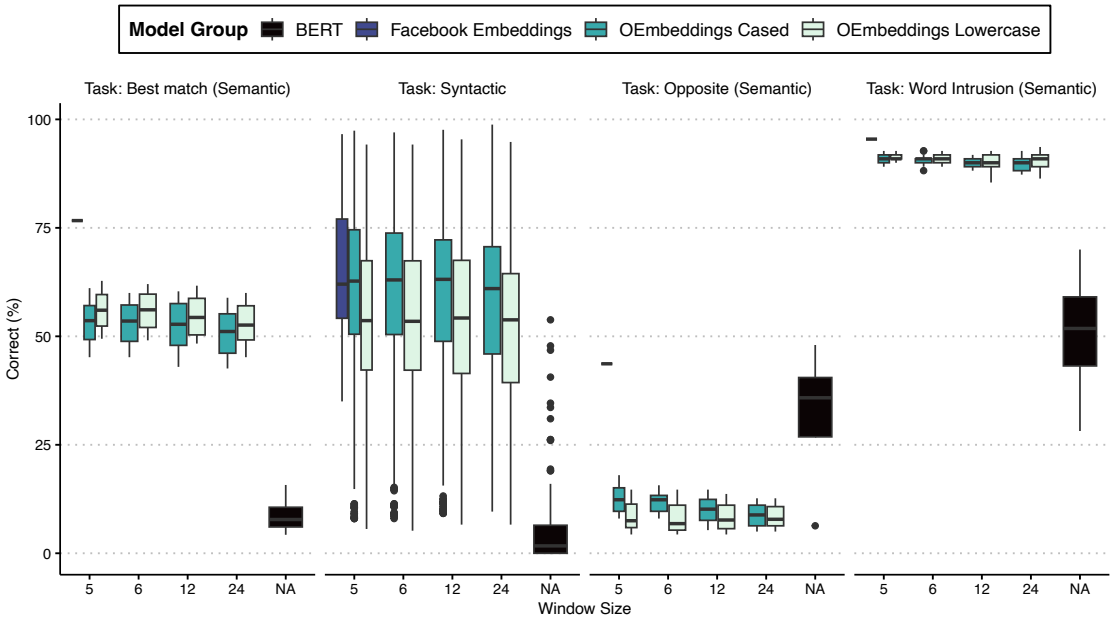


Figure 5

Intrinsic Task Semantic: Best Match

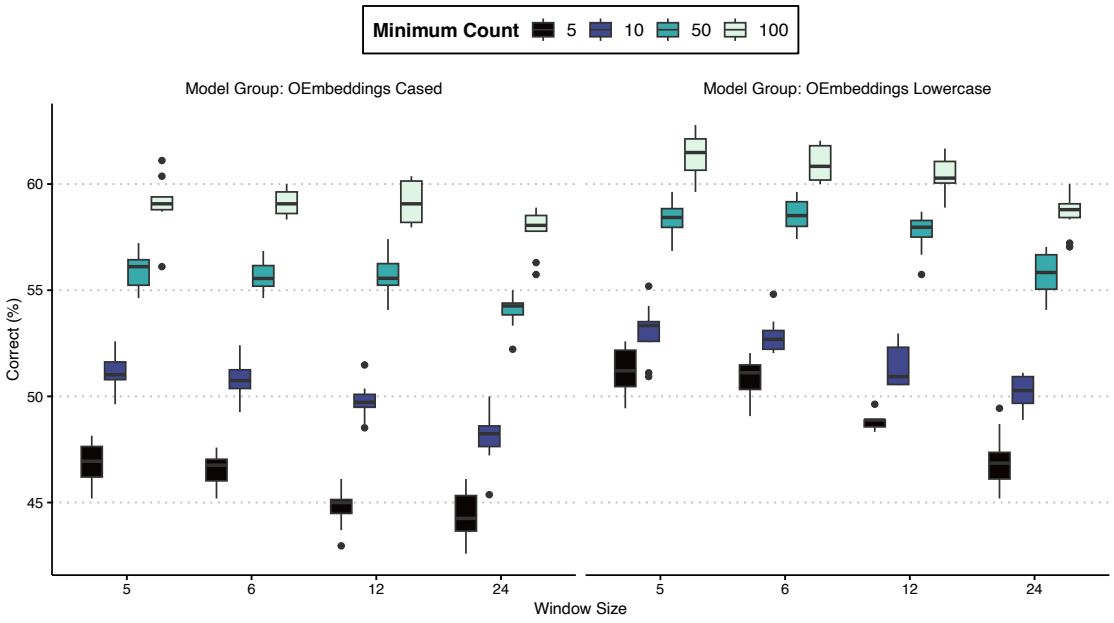
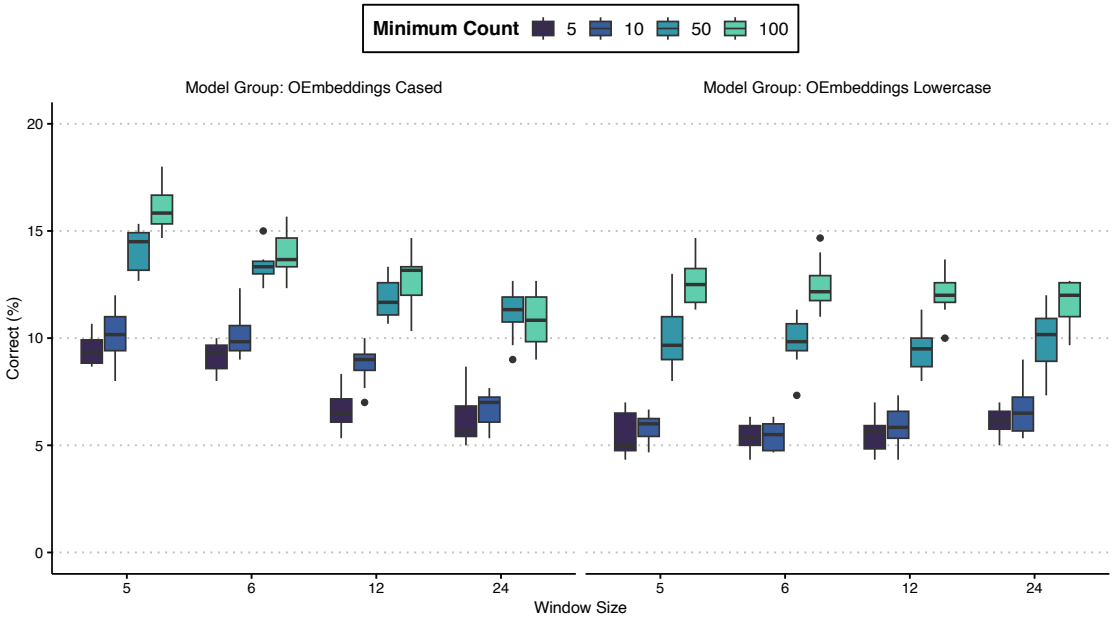


Figure 6

Intrinsic Task Semantic: Opposite



as clearly as it does in the other two tasks. It seems that a larger minimum count in this case hurts performance, but only when combined with a smaller window size. While the performances of the models are rather consistent, all three hyperparameters have a small but significant impact on the performance (see Table 2 in the Appendix).

The performance of the third-party models is quite sparse, as the XLM-RoBERTa performs very poorly, but also the other BERT models perform worse than the models we trained on the Austrian News Corpus. The Facebook model on the other hand outperforms our models, as it correctly identifies the intruder word in 95% of the cases.

Syntactic Task: Grammar

The last intrinsic task queries the syntactic understanding of our models. While performance within families tends to be fairly consistent, we find that performance varies widely between families. Especially casing has a large and significant impact on performance scores. This is not surprising, since nouns are capitalized in German, and thus especially when looking at grammar, it is important to take into account casing to differentiate between two words that are written the same way, but the casing marks the difference between the word being understood as a noun or verb. Moreover, we find that a higher minimum count of words also improves the performance of our models significantly. The window size does have a significant impact, and Figure 8 also shows that performance is quite consistent, except for the largest window size (24) which continuously seems to underperform.

Only the Facebook model can compete with the syntactic task performance of our models, with a performance of 0.66. All other BERT off-the-shelf models perform rather poorly with less than 15% of correctly answered tasks.

Taking all different intrinsic validation tasks together, we find some patterns. Except for the second task, lowercasing the training data hurts the performance of German word embedding models significantly. Window Size is also negatively related to intrinsic model performance, however, its impact is only significant for the the best match and word intrusion task. A higher minimum count is generally related to an increase in model performance, however, the hyperparameter is not significant throughout the different dependent variables. Among the off-the-shelf models, the Facebook embeddings show the highest performance, with an average of 70%. The BERT models tend to perform correctly on average in 20%–30%. The best of our self-trained models perform correctly in 58% of the tasks.

Extrinsic validation

The extrinsic validation of our model was assessed with single-label and multi-label classification tasks. The two tasks: author prediction (5 classes) and sentiment classification (3 classes) are single-label tasks, as there is only one correct answer per training unit. The topic classification is more difficult because there are multiple labels per unit and the number of classes is comparatively high at 13. Here, too, we shed light on the differences between the model families to assess the impact of the different hyperparameter settings (See Figure 9).

Figure 7

Intrinsic Task Semantic: Word Intrusion

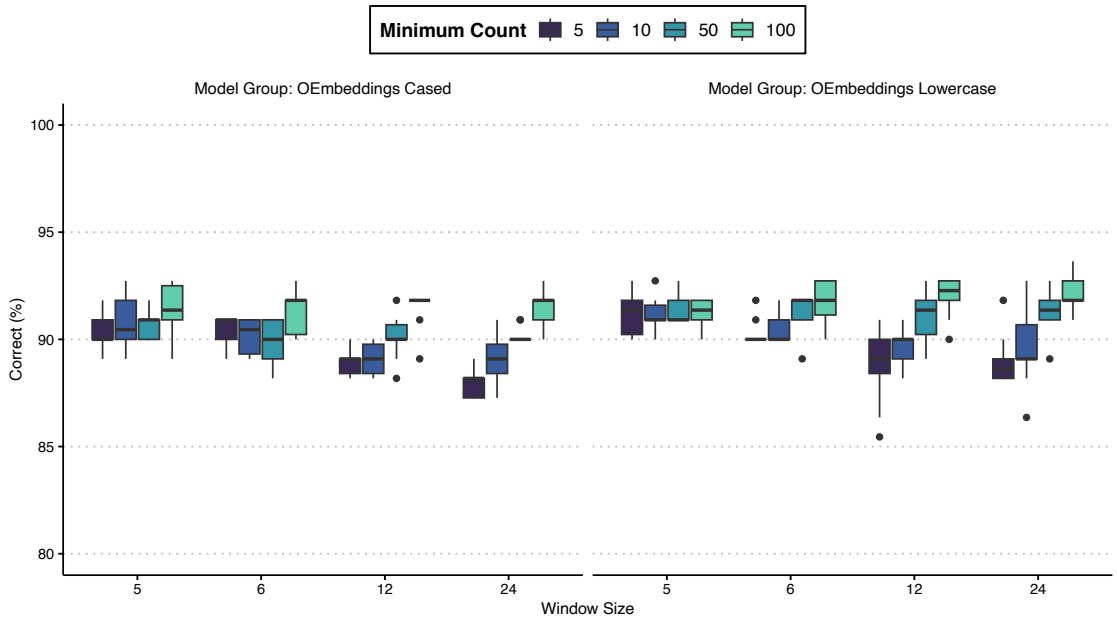
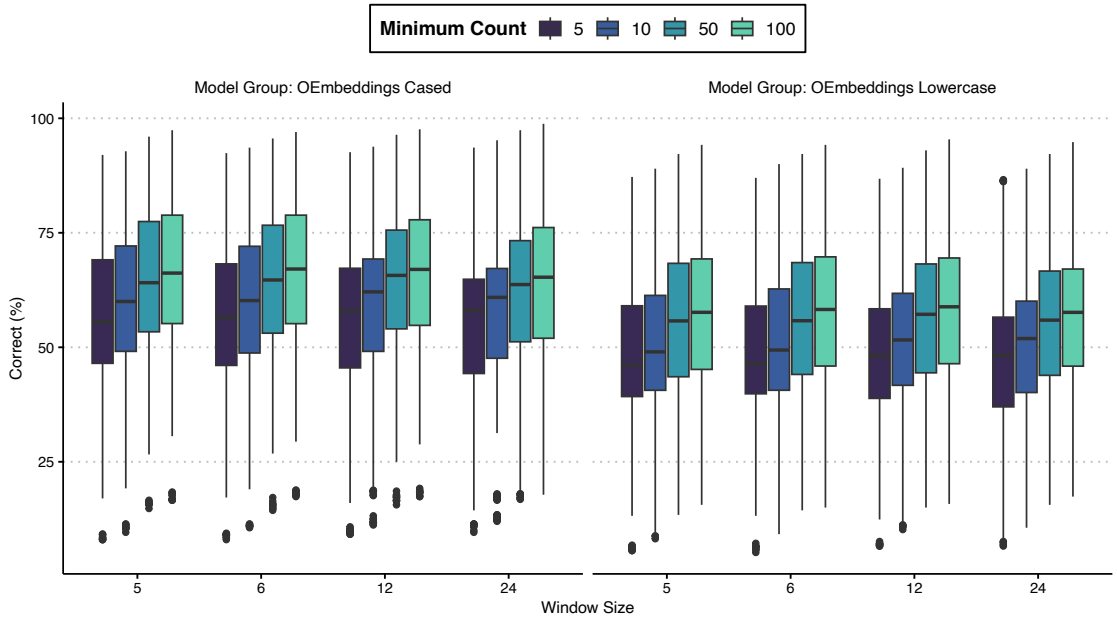


Figure 8

Intrinsic Task Syntactic: Grammar



Author Prediction

As can be seen from Figure 9, the models' performances are quite stable over all three hyperparameters and also within model families. The only significant hyperparameter is casing (see Figure 10), which has a positive impact. Thus, models built on lowercased data tend to show better performance for author prediction.

When comparing the off-the-shelf models, we find that all BERT models outperform our self-trained models by far, as all of them correctly predict the authors in at least three-quarters of the tasks. However, the Facebook model performs rather similarly to our models, being correct in 52% of the tasks.

Topic Classification

For the second extrinsic task, topic classification, we find larger differences in model performance, also within model families, especially those with smaller window sizes. Similar to the first extrinsic task, casing is the only significant predictor of performance, with uppercased models outperforming their lowercased counterparts. We note that while Figure 9 visually shows an increase in performance for models with a smaller window size, statistical tests (see Figure 10) do not confirm significant effects. Regarding the minimum word count, we also find no impact on the performance of the models in classifying topics.

The topic classification was a difficult task for the off-the-shelf models. GottBERT and XLM-RoBERTa are the only ones that perform on a similar level as our self-trained models (Table 3), while the other models are clearly performing worse, only correctly classifying the topic in around 17% (Facebook).

Sentiment Classification

The third and final extrinsic task is sentiment classification, which asks the model to sort paragraphs of news articles based on their sentiment (positive, negative, and neutral). Figure 9 shows a consistent performance of the models, and the regression analysis (Table 2) confirms that neither minimum word count nor window size have a significant influence on the model's performance. As with the first two extrinsic tasks, lowercasing the models does have a significant impact. In this application, however, the effect is negative, indicating that cased models perform worse than their lower-cased counterparts. Nevertheless, casing only has a small effect on the model's performance overall.

Again, the BERT models outperform our models on the sentiment classification as all of them score above 0.59. The Facebook model correctly infers the sentiment in 45% of the tasks, which is below the average of our models.

Overall, with respect to extrinsic validation, we find that the window size does not have a significant impact on the performance of the models, nor does the minimum count of words in the training corpus. The only hyperparameter that has an impact on the performance is lowercasing, but the direction of the hyperparameter is inconsistent. Regarding the third-party models, we find that the BERT models are better equipped to solve the extrinsic tasks as compared to the Facebook model, as well as all of our self-trained models.

Our evaluation indicates that different tasks favor different model families, and the performance differences, depending on the task, can be quite substantial. Consequently, from this general but comprehensive validation strategy, it is not clear which hyperparameter combination should be favored or how to weigh the performance of the models within a family. Therefore, to gain a deeper understanding, we further investigate how these models differ when applied to substantive cases.

Practical Example

From the analysis of the different validation tasks, it was not possible to distill a combination of hyperparameters that performed "best" across the different validation methods (i.e., intrinsic or extrinsic). On average, however, the models performed quite similarly. This could lead to the conclusion that it does not matter which model we choose for our analysis. To see whether the models would lead to similar substantive results, we include a practical application to further investigate the performance of the embedding models. We run the analysis on one model per family that performs best on average across all seven validation tasks; since the model families were quite stable, we believe this is an appropriate reduction. In this study, we specifically evaluate the performance of word embedding models by examining the nearest neighbors of the words "Frau" (woman) and "Femizid" (femicide) in the context of Austrian news articles. This approach helps us evaluate the model's ability to understand and contextualize important social issues, and the stability of our results when we apply word embeddings to analyze social science phenomena.

Word embedding models have different vectors for the different grammatical versions of a word, however, they are often very close to each other. Thus, the nearest neighbors often include different versions of the same word. As one would do in a substantive analysis, we reduced all words to their base form. The lemmatizing of the words was done manually.[8] In order to compare both cased and lowercased models, we also lowercased all nearest neighbor words.

Figure 13 in the Appendix shows the relative overlap in the 100 nearest neighbors of the word "Frau" between our 32 models, which is on average 50.26%. Overall, there were 339 different words within the nearest neighbors, but their frequency varies drastically. There are only twelve words which can be found in every model,[9] nevertheless only 26 words in total are present in more than three-quarters of the models, and 53 words (15%) can only be found in one single model. The majority of neighbors could be connected to the word "women" however there were also some words, which seemed to be artifacts of the text pre-processing. Different media outlets use different ways of gender-sensitive language and while a great deal of work was put into a solid pre-processing and normalization, there are still some artifacts that stayed.

Figure 12 in the Appendix shows the relative overlap in the 100 nearest neighbors

[8]We first tried both the spaCy as well as NLTK lemmatizer, however, were not satisfied with their performance and, thus, corrected the words manually.

[9]partner, ex wife, life partner, rival, ex-girlfriend, neighbor, wife, housekeeper, spouse, sister-in-law, companion, babysitter

of the word "Femizid" between our 32 models, the average being 51,33%. We find 311 different words, although their frequency is quite diverse. There are eight words, which are, predictably and reassuringly, present in all models, such as grammatical variations of the word "femicide" or synonyms such as "women murder". Additionally, "suicide", "sexual murder", "hate towards women", "women murderer", "male violence" as well as "women hater" are present in every model. Nevertheless, there are only 17 words which are present in more than three-quarters of the models and almost a fifth of all words (61) are only present in one single model. While all 311 words can be related to femicides, the differences between the models still convey different nuances.

From a substantive point of view this is quite concerning, as this lack of overlap between the models does show that even though the validation strategy tuning did not clearly indicate a "best" model, the model choice does clearly impact the results of our investigations. The variability in results across different models underscores the volatility and dependency of substantive study outcomes on the specific word embedding model chosen by researchers, highlighting the critical need for transparent communication regarding model selection and its potential impact on research conclusions.

Discussion

Our study evaluates the impact of three different hyperparameter settings—case, minimum word count, and window size—when computing word embedding models for a large Austrian news media corpus, on stability as well as performance on intrinsic and extrinsic validation tasks. Furthermore, we compare these customized models with off-the-shelf models in order to contribute to the ongoing debate on the validation of domain-specific word embedding models. Our results do not present a clear-cut picture of word embedding validation, highlighting the complexity and challenges inherent in this process. However, they do reveal critical insights into current practices and potential areas for improvement.

First, our study highlights several aspects regarding the stability and performance of word embedding models. This shows, in line with previous research from computational linguistics, that a single validation method or task is not sufficient to evaluate the efficacy of word embeddings comprehensively (Antoniak & Mimno, 2018) and that intrinsic evaluation methods fail to be good predictors of how well an embedding performs on extrinsic tasks (Chiu et al., 2016). The proposed model families themselves seem rather stable based on cue words from different contexts, which we attribute to the considerable size of the training corpus. However, we found noticeable differences in performance between the families, indicating that hyperparameter settings significantly impact model performance. Despite this internal stability, the performance of the model families varies significantly across different tasks, and importantly, performance does not correlate across these tasks. In line with the lack of overlap in the practical application, this raises concerns about the robustness and validity of findings derived from word embedding models in general, especially if findings are based on singular models, or if validation is not reported transparently.

Second, our analysis suggests that not all validation tasks are equally important for every model, leading to misleading conclusions about model performance. Researchers must, therefore, rethink what validity means for word embeddings in their specific context,

but beware of overfitting their model. This observation calls for a more nuanced approach to validation, to better capture the complexities of language use and meaning. It is crucial to decide beforehand which tasks are most pertinent to their study and why these tasks are chosen. This proactive approach ensures that the chosen validation methods align closely with the research objectives. We can improve the reliability and applicability of word embeddings in social science research by developing a deeper understanding of these nuances.

Third, to ensure the robustness of findings, it is advisable to run multiple models whenever possible, at least for parts of the analysis. Transparency about the use of different models and the variation in results is essential. This approach of a multiverse analysis not only enhances the credibility of the research but also helps in understanding the potential uncertainty associated with relying on a single model (Pipal et al., 2023). Accepting and communicating the inherent uncertainty in results due to model reliance is also important. Researchers should acknowledge that while their findings are based on specific models, there is always some level of uncertainty, which needs to be transparently discussed in their work. This holds true for self-trained models, as well as off-the-shelf models.

Despite the comprehensive nature of our study, several limitations warrant consideration. First, while our analysis included multiple well-known and commonly used validation strategies, there are alternative approaches that could be explored. For instance, assessing the distance between answers to questions or the distance between cue words might provide additional insights into the usefulness of the different word embedding models. Second, using an even larger training corpus might mitigate some impacts of hyperparameter tuning; however, it also introduces challenges related to computational resources and time constraints. This large dataset, while valuable, may overshadow the need for a more refined tuning process that smaller datasets might necessitate. Thirdly, we did not implement a human-based validation method, such as the Turing test suggested by Rodriguez and Spirling (2022). Such a test could offer an in-depth evaluation of the human-like quality of our word embeddings. However, the implementation for 32 model families was outside of the scope of this study. Finally, the practical application focused on differences between the model families and not within the model families, which would have added more insights into the stability of word embedding models.

The resource-intensive and time-consuming nature of validation processes may explain their infrequent use in social science contexts. However, neglecting thorough validation exposes research to the risk of misinterpretation and invalid conclusions. As a field, we should strive to establish validation, thoughtful consideration of model selection, and hyperparameter settings as standard practice. This shift will require a concerted effort by authors, reviewers, and editors alike to ensure that these critical aspects are not overlooked. In summary, our study underscores that different models excel at different tasks and that there is no one-size-fits-all solution, either for downstream tasks or for validation. Researchers need to determine which model best suits their specific needs (i.e., is casing needed?), taking into account the unique characteristics of their data and the particular requirements of their research questions. When researchers train their own word embedding models, the focus should not be disproportionately on tuning hyperparameters alone. Instead, it may be more

productive to run the same settings multiple times and explore different methods to check the validity and robustness of the results. Hyperparameter tuning, while important, can often yield diminishing returns compared to the benefits of rigorous validation practices. By prioritizing comprehensive validation strategies and contextual evaluations, researchers can significantly improve the quality of their results. This approach fosters a more robust application of word embeddings in computational communication science, ultimately leading to more trustworthy and impactful research results.

Moreover, we have trained a set of models on a rather unique dataset, pertinent to the Austrian context. Researchers working on Austrian-related projects now have these models at their disposal. These models perform differently depending on various training factors (e.g., lowercased vs. uncased), allowing researchers to select the models based on their specific needs and tasks. They can determine the most suitable model by reviewing the validation results provided in our paper and comparing them to their specific tasks. Alternatively, researchers can use several of these models simultaneously to enhance their work. We provide all the models we trained for public access and encourage further experimentation and application in diverse contexts. Researchers can access the models here: (link will provided upon publication).

Acknowledgements

Masked for review

Conflicts of interest

There are no conflicts to declare.

Data availability statement

The code for replicating this study is available at https://osf.io/gxzvs/?view_only=6df9df425b6c42828539ac5193a98098. Due to copyright restrictions the data for training the models cannot be publicly shared.

References

Aldrete, M., Taşkale, N., Rivera Ramirez, E., & Gil Vera, V. D. (2024). Media representations of femicide. A systematic review of literature in English and Spanish [Publisher: Routledge]. *Annals of the International Communication Association*, 1–18. https://doi.org/10.1080/23808985.2024.2336924

Antoniak, M., & Mimno, D. (2018). Evaluating the Stability of Embedding-based Word Similarities. *Transactions of the Association for Computational Linguistics*, *6*, 107–119. https://doi.org/10.1162/tacl_a_00008

Avraham, O., & Goldberg, Y. (2016). Improving Reliability of Word Similarity Evaluation by Redesigning Annotation Task and Performance Measure. *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, 106–110. https://doi.org/10.18653/v1/W16-2519

Baden, C., Pipal, C., Schoonvelde, M., & van der Velden, M. A. G. (2022). Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda. *Communication Methods and Measures*, *16*(1), 1–18. https://doi.org/10.1080/19312458.2021.2015574

Batchkarov, M., Kober, T., Reffin, J., Weeds, J., & Weir, D. (2016). A critique of word similarity as a method for evaluating distributional semantic models. *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, 7–12. https://doi.org/10.18653/v1/W16-2502

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Bernhard, J., Ashour, R., & Boomgaarden, H. G. (2022, June). Towards Validity Standards of Topic Models in Computational Social Science [Jahrestagung der Fachgruppe Methoden der DGPuK 2022].

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. https://doi.org/10.1162/tacl_a_00051

Buehling, K. (2023). Message Deletion on Telegram: Affected Data Types and Implications for Computational Analysis [Publisher: Routledge _eprint: https://doi.org/10.1080/19312458.2023.2183188]. *Communication Methods and Measures*, *0*(0), 1–23. https://doi.org/10.1080/19312458.2023.2183188

Camacho-Collados, J., & Navigli, R. (2016). Find the word that does not belong: A Framework for an Intrinsic Evaluation of Word Vector Representations. *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, 43–50. https://doi.org/10.18653/v1/W16-2508

Chan, B., Schweter, S., & Möller, T. (2020). German's Next Language Model. *Proceedings of the 28th International Conference on Computational Linguistics*, 6788–6796. https://doi.org/10.18653/v1/2020.coling-main.598

Chan, C. H., Zeng, J., Wessler, H., Jungblut, M., Welbers, K., Bajjalieh, J. W., van Atteveldt, W., & Althaus, S. L. (2020). Reproducible Extraction of Cross-lingual Topics (rectr) [Publisher: Routledge _eprint: https://doi.org/10.1080/19312458.2020.1812555]. *Communication Methods and Measures*, *14*(4), 285–305. https://doi.org/10.1080/19312458.2020.1812555

Che, X., Ring, N., Raschkowski, W., Yang, H., & Meinel, C. (2017, September). Traversal-Free Word Vector Evaluation in Analogy Space. In S. Bowman, Y. Goldberg, F. Hill, A. Lazaridou, O. Levy, R. Reichart, & A. Søgaard (Eds.), *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP* (pp. 11–15). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-5302

Chiu, B., Korhonen, A., & Pyysalo, S. (2016). Intrinsic Evaluation of Word Vectors Fails to Predict Extrinsic Performance. *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, 1–6. https://doi.org/10.18653/v1/W16-2501

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020, April). Unsupervised Cross-lingual Representation Learning at Scale [arXiv:1911.02116 [cs]]. https://doi.org/10.48550/arXiv.1911.02116

Europäisches Institut für Gleichstellungsfragen. (2022). *Zahlenmäßige Erfassung von Femizid in österreich* (tech. rep.) (doi:10.2839/22070). https://eige.europa.eu/sites/default/files/documents/20211560_pdf_mh0121100den_002.pdf

Faruqui, M., Tsvetkov, Y., Rastogi, P., & Dyer, C. (2016). Problems With Evaluation of Word Embeddings Using Word Similarity Tasks. *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, 30–35. https://doi.org/10.18653/v1/W16-2506

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644. https://doi.org/10.1073/pnas.1720347115

Gladkova, A., & Drozd, A. (2016). Intrinsic Evaluations of Word Embeddings: What Can We Do Better? *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, 36–42. https://doi.org/10.18653/v1/W16-2507

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.

Haselmayer, M., Dingler, S. C., & Jenny, M. (2022). How Women Shape Negativity in Parliamentary Speeches—A Sentiment Analysis of Debates in the Austrian Parliament [Publisher: Oxford University Press UK]. *Parliamentary Affairs*, *75*(4), 867–886.

Kroon, A., Trilling, D., & Raats, T. (2020). Guilty by Association: Using Word Embeddings to Measure Ethnic Stereotypes in News Coverage - Anne C. Kroon, Damian Trilling, Tamara Raats, 2021. *Jouralism & Mass Communication Quarterly*, *89*(2). https://doi.org/10.1177/1077699020932304

Kroon, A., Welbers, K., Trilling, D., & van Atteveldt, W. (2024). Advancing Automated Content Analysis for a New Era of Media Effects Research: The Key Role of Transfer Learning [Publisher: Routledge]. *Communication Methods and Measures*, *18*, 1–21. https://doi.org/10.1080/19312458.2023.2261372

Lai, S., Liu, K., Xu, L., & Zhao, J. (2015). How to generate a good word embedding?

Liang, H., Ng, Y. M. M., & Tsang, N. L. T. (2023). Word Embedding Enrichment for Dictionary Construction: An Example of Incivility in Cantonese [Number: 1]. *Computational Communication Research*, *5*(1). https://doi.org/10.5117/CCR2023.1.10.LIAN

Litvyak, O., Balluff, P., Müller, W. C., Kritzinger, S., & Boomgaarden, H. G. (2022, August). AUTNES Manual Content Analysis of the Media Coverage 2019 (SUF edition). https://doi.org/10.11587/FVTDXG

Liza, F. F., & Grześ, M. (2016). An Improved Crowdsourcing Based Evaluation Technique for Word Embedding Methods. *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, 55–61. https://doi.org/10.18653/v1/W16-2510

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013, September). Efficient Estimation of Word Representations in Vector Space. https://doi.org/10.48550/arXiv.1301.3781

Müller, A. (2015). German Word Embeddings. Retrieved March 7, 2023, from https://devmount.github.io/GermanWordEmbeddings/

Nayak, N., Angeli, G., & Manning, C. D. (2016). Evaluating Word Embeddings Using a Representative Suite of Practical Tasks. *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, 19–23. https://doi.org/10.18653/v1/W16-2504

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

Pipal, C., Song, H., & Boomgaarden, H. G. (2023). If You Have Choices, Why Not Choose (and Share) All of Them? A Multiverse Approach to Understanding News Engagement on Social Media [Publisher: Routledge]. *Digital Journalism*, *11*(2), 255–275. https://doi.org/10.1080/21670811.2022.2036623

Rheault, L., & Cochrane, C. (2020). Word embeddings for the analysis of ideological placement in parliamentary corpora. *Political Analysis*, *28*(1), 112–133.

Rodriguez, P. L., & Spirling, A. (2022). Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research. *The Journal of Politics*, *84*(1), 101–115. https://doi.org/10.1086/715162

Romney, D., Stewart, B. M., & Tingley, D. (2015). Plain Text: Transparency in the Acquisition, Analysis, and Access Stages of the Computer-assisted Analysis of Texts. *Qualitative and Multi-Method Research*, *13*(1), 32–37.

Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š., & Sedlmair, M. (2018). More than bags of words: Sentiment analysis with word embeddings. *Communication Methods and Measures*, *12*(2-3), 140–157.

Scheible, R., Thomczyk, F., Tippmann, P., Jaravine, V., & Boeker, M. (2020, December). GottBERT: A pure German Language Model [arXiv:2012.02110 [cs]]. https://doi.org/10.48550/arXiv.2012.02110

Schnabel, T., Labutov, I., Mimno, D., & Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. *Proceedings of the 2015 conference on empirical methods in natural language processing*, 298–307.

Schuld, M., Durrheim, K., & Mafunda, M. (2023). Speaker landscapes: Machine learning opens a window on the everyday language of opinion [Publisher: Routledge _eprint: https://doi.org/10.1080/19312458.2023.2277958]. *Communication Methods and Measures*, *0*(0), 1–17. https://doi.org/10.1080/19312458.2023.2277958

Seyed, A. (2016). Subsumption Preservation as a Comparative Measure for Evaluating Sense-Directed Embeddings. *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, 90–93. https://doi.org/10.18653/v1/W16-2516

van Atteveldt, W., van der Velden, M. A. C. G., & Boukes, M. (2021). The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dic-

tionary Approaches, and Machine Learning Algorithms [Publisher: Routledge _eprint: https://doi.org/10.1080/19312458.2020.1869198]. *Communication Methods and Measures*, *15*(2), 121–140. https://doi.org/10.1080/19312458.2020.1869198

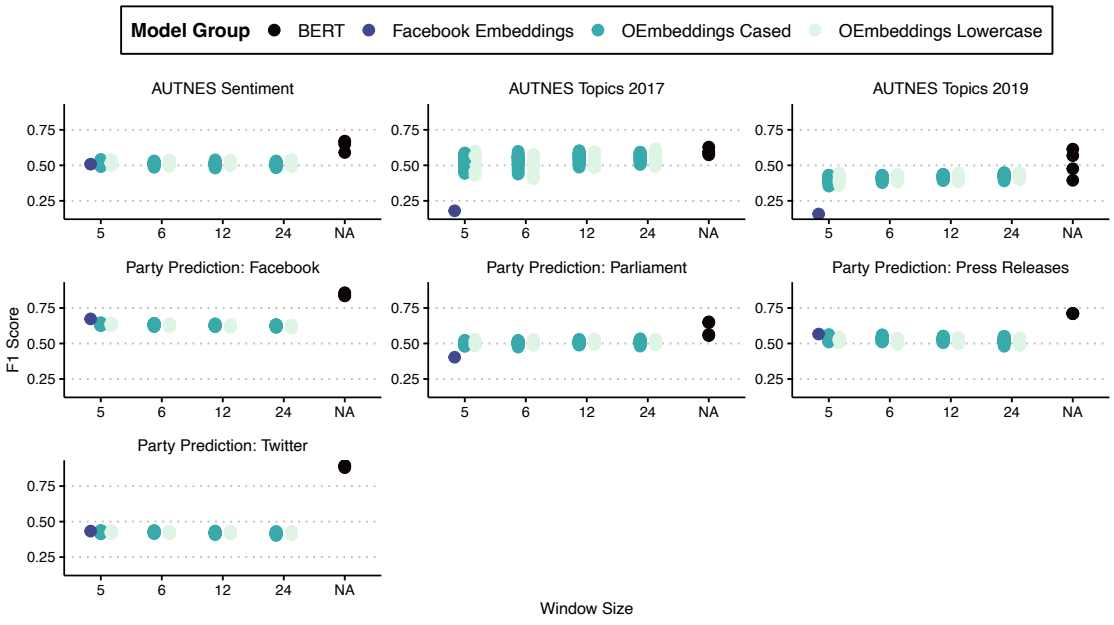
van der Veen, A. M. (2023). Word-level machine translation for bag-of-words text analysis: Cheap, fast, and surprisingly good [Publisher: Amsterdam University Press Amsterdam]. *Computational Communication Research*, *5*(2), 1.

Viehmann, C., Beck, T., Maurer, M., Quiring, O., & Gurevych, I. (2023). Investigating Opinions on Public Policies in Digital Media: Setting up a Supervised Machine Learning Tool for Stance Classification [Publisher: Routledge _eprint: https://doi.org/10.1080/19312458.2022.2151579]. *Communication Methods and Measures*, *17*(2), 150–184. https://doi.org/10.1080/19312458.2022.2151579

Yuan, L. X. (2023). Distilbert-base-multilingual-cased-sentiments-student (Revision 2e33845). https://doi.org/10.57967/hf/1422

Figure 9

Extrinsic Task: Author, Topic and Sentiment Classification



Appendix

Figure 10

Posterior Distributions of Model Parameters per Validation Task

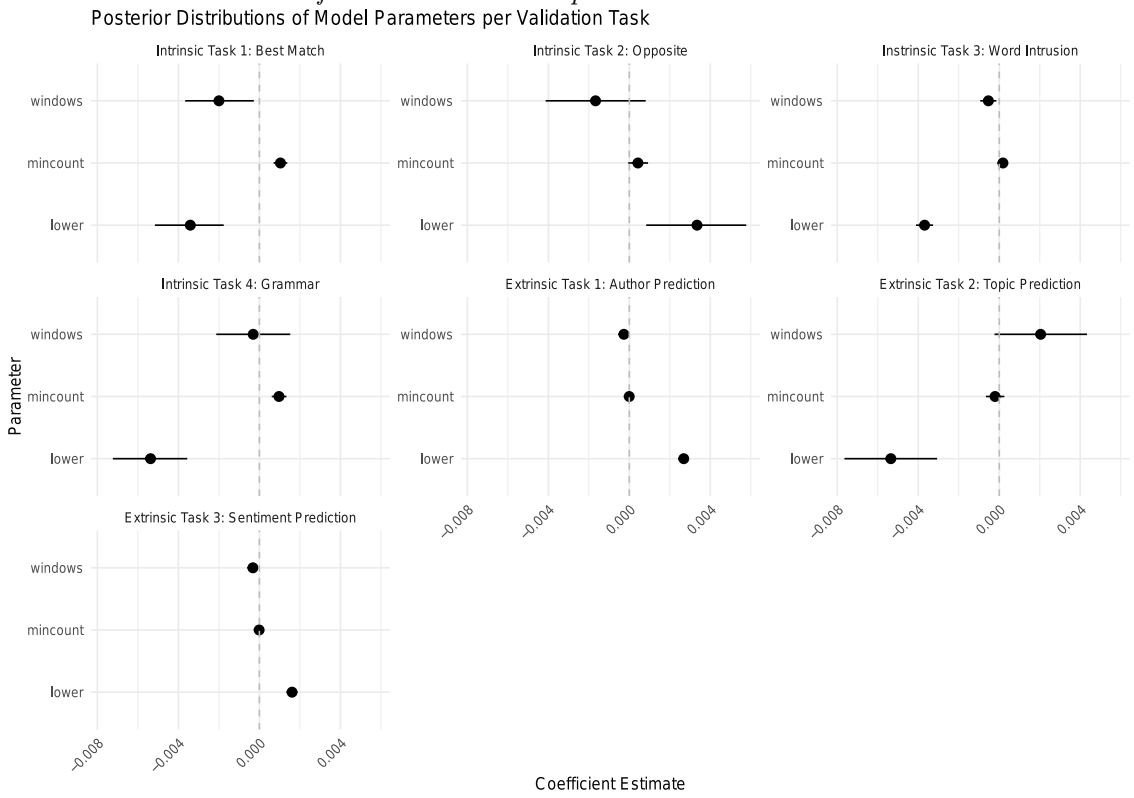


Table 2

Posterior Distributions of Model Parameters per Validation Task

	Intrinsic Task 1: Best Match	Intrinsic Task 2: Opposite	Intrinsic Task 3: Word Intrusion	Intrinsic Task 4: Grammar	Extrinsic Task 1: Author Prediction	Extrinsic Task 2: Topic Classification	Extrinsic Task 3: Sentiment Classification
Intercept	0.52275	0.10409	0.90618	0.53520	0.52316	0.45919	0.51866
	(0.49622-0.54907)	(0.06593-0.14225)	(0.90000-0.91261)	(0.50713-0.56355)	(0.52121-0.52511)	(0.42363-0.49469)	(0.51467-0.52255)
Lower Casing	-0.00341	0.00334	-0.00369	-0.00538	0.00269	-0.00535	0.00161
	(-0.00516 - -0.00176)	(0.00084-0.00577)	(-0.00411 - -0.00327)	(-0.00724 - -0.00356)	(0.00256-0.00281)	(-0.00764 - -0.00307)	(0.00135-0.00188)
Minimum Count	0.00104	0.00042	0.00018	0.00097	-0.00001	-0.00021	-0.00001
	(0.00070-0.00138)	(-0.00005-0.00092)	(-0.00005-0.00026)	(0.00061-0.00134)	(-0.00003-0.00002)	(-0.00066-0.00025)	(-0.00006-0.00004)
Window Size	-0.00200	-0.00167	-0.00054	-0.00081	-0.00027	0.00204	-0.00032
	(-0.00367 - -0.00028)	(-0.00413 - -0.00081)	(0.00010 - -0.00014)	(-0.00213 - 0.00152)	(-0.00039 - -0.00014)	(-0.00223-0.00433)	(-0.00057 - -0.00007)

Table 3

Overview Validation Results for Off-the-shelf Models

Task	Facebook Embeddings	GBERT	DistilBERT	GottBERT	XLM-RoBERTa
Intrinsic (% correct)					
Syntactic	**0.657**	0.145	0.020	0.024	0.108
Best match (Semantic)	**0.767**	0.157	0.043	0.089	0.067
Opposite (Semantic)	0.437	**0.480**	0.063	0.380	0.337
Word Intrusion (Semantic)	**0.955**	0.482	0.700	0.555	0.282
Extrinsic (F1 Score)					
Party Prediction: Twitter	0.433	0.891	0.888	0.880	**0.895**
Party Prediction: Press Releases	0.566	0.712	0.710	0.709	**0.713**
Party Prediction: Facebook	0.673	0.855	0.836	0.836	**0.857**
Party Prediction: Parliament	0.403	**0.652**	0.563	0.555	0.647
AUTNES Sentiment	0.508	0.667	0.591	0.650	**0.670**
AUTNES Topics 2017	0.180	0.596	0.589	0.574	**0.629**
AUTNES Topics 2019	0.157	**0.614**	0.476	0.395	0.568

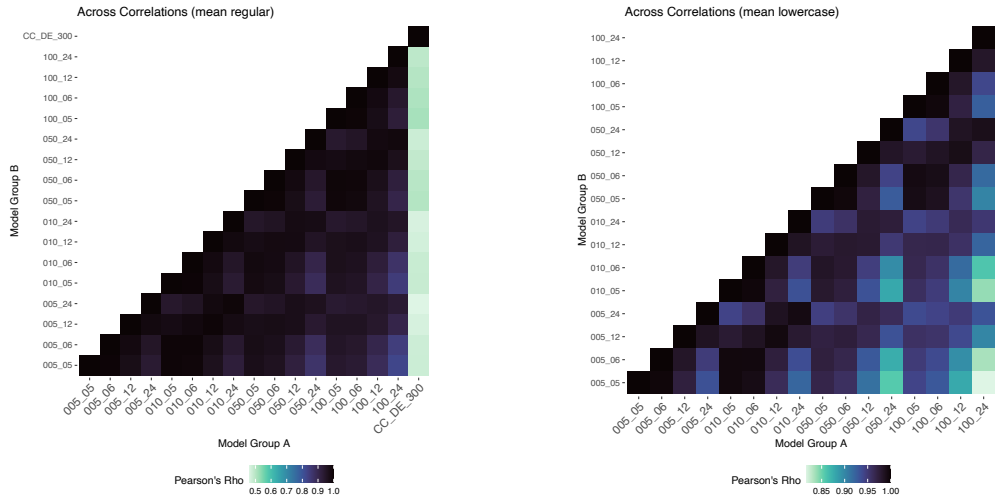
Table 4

Overview of Cue Words used for Stability Assessments

Category	Keywords
Politics	Demokratie, Gleichheit, Gerechtigkeit, Einwanderung, Pension, Sozialstaat, Bildung, Steuern, ÖVP, SPÖ, FPÖ, Grüne, NEOS
Economy	Wirtschaft, Steuer, Teuerung, Euro, Banken, Budget
Culture	Kunst, Kultur, Theater, Zensur, Festspiel, Oper, Konzert
Sports	Sport, Fußball, Skifahren, Medaille, Podest, Sieg, Niederlage
Environment	Klima, Umwelt, Schutz, Nachhaltigkeit, Energie, Nationalpark
Migration	Einwanderung, Asyl, Migration, Flucht, Schlepperei, Fremdenhass
Social Groups	Wissenschaftler, Wissenschaftlerin, Unternehmer, Unternehmerin, Verbrecher, Verbrecherin, Polizist, Polizistin, Politiker, Politikerin, Stadtbewohner, Stadtbewohnerin, Lehrer, Lehrerin, Lehrling, Künstler, Künstlerin, Landbewohner, Landbewohnerin, Landwirt, Landwirtin, Alleinerziehende, Millionär, Millionärin, Geringverdiener, Geringverdienerin, Student, Studentin, Studierende, Raucher, Raucherin, Pensionist, Pensionistin, Homosexueller, Homosexuelle, Flüchtling, Frau, Mann, Behinderter, Behinderte, Beeinträchtigter, Beeinträchtigte, Beamter, Beamtin, Arbeiter, Arbeiterin, Autofahrer, Autofahrerin, Arbeitsloser, Arbeitslose

Figure 11

Across correlations. The model families are abbreviated with their minimum word count and window size. E.g., "050_24" points at the model family that was trained with a minimum count to 50 and a window size of 24.



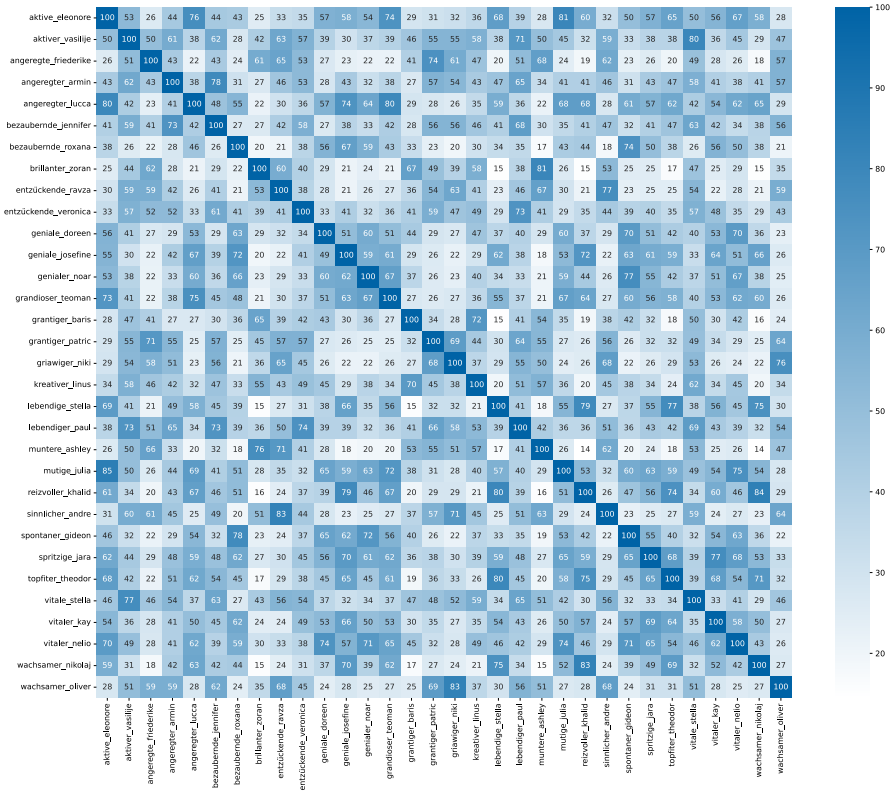


Figure 12

Overlap for the word femicide between the 100 nearest neighbors in percent

Table 5

Number of Articles per Year and News Outlet in the Training Data

News Outlet	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
www.derstandard.at	0	0	80,501	73,076	60,819	66,268	67,408	66,939	60,793	52,344	45,287	43,214	46,321
www.diepresse.com	54,642	54,728	56,109	58,897	59,438	56,904	55,353	48,397	49,740	44,502	37,829	39,064	36,577
www.gmx.at	0	27	18	12	4,608	2,743	6,801	9,803	11,873	14,918	15,627	15,381	20,562
www.heute.at	0	0	1	0	2	2	1	73	225	20,754	42,971	45,484	49,875
www.kleinezeitung.at	65,038	71,598	74,287	80,890	78,039	86,606	75,104	94,853	104,817	107,699	107,451	91,881	83,869
www.krone.at	1	0	23,621	25,824	28,081	32,140	34,496	42,106	51,396	72,433	77,127	88,570	85,580
www.kurier.at	3,034	23,094	41,181	43,151	44,646	43,713	46,280	43,298	43,594	49,835	53,213	54,624	52,089
www.oe24.at	48,708	47,347	48,540	55,186	55,903	41,772	41,506	36,117	32,311	34,091	50,903	51,223	37,846
www.orf.at	18,139	44,921	50,945	51,167	52,737	54,578	57,312	54,773	52,333	48,219	51,451	49,989	47,695
www.sn.at	0	25	31,955	35,486	35,655	39,162	41,311	48,851	47,027	53,129	50,091	47,403	47,153

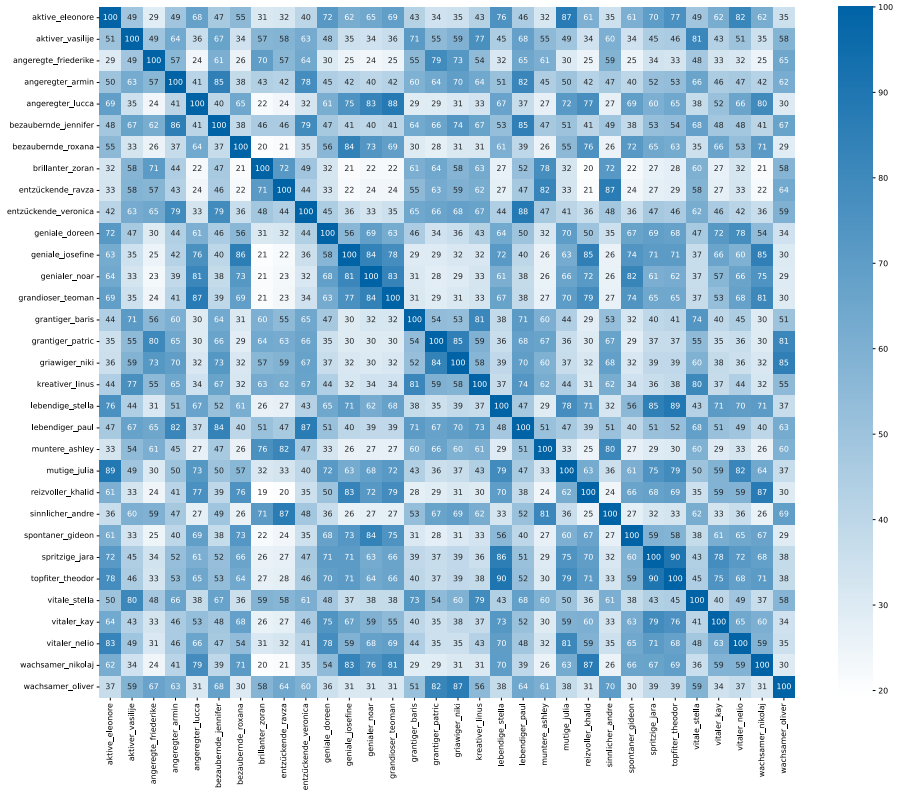


Figure 13

Overlap for the word women between the 100 nearest neighbors in percent

4.2 Beyond Standardization: A Comprehensive Review of Topic Modeling Validation Methods for Computational Social Science Research

Beyond Standardization: A Comprehensive Review of Topic Modeling Validation Methods for Computational Social Science Research

Jana Bernhard[1], Randa Ashour[1], Petro Tolochko[1], Jakob-Moritz Eberl[1], and Hajo G. Boomgaarden[1]
[1]Department of Communication, University of Vienna

Abstract

With the proliferation of computational text analysis in the social sciences, topic modeling has gained substantial popularity for uncovering latent themes in textual data. However, concerns have arisen regarding the validity of its outcomes, given its largely automated and inductive nature and scholars have noted the absence of clear patterns in validating topic models. In response to this, we present a comprehensive systematic review of 792 studies employing topic models. Our study seeks to address the question of whether the field has started to converge towards a common validation framework of topic models. In our systematic review we identified a notable absence of standardized validation practices and a lack of convergence towards specific validation methods. This may be attributed to the inherent mismatch between the inductive and qualitative nature of topic modeling and the deductive, quantitative research tradition that seeks standardized validation practices. As a result, we advocate for better taking into account qualitative validation understandings, building on transparency and detailed reporting to enhance the credibility of findings in the computational social sciences.

Keywords: Topic Modeling, Validation, Text-as-Data, Standardization, Convergence

Introduction

With the maturation of computational text analysis within the social sciences (Bonikowski & Nelson, 2022), topic modeling has become a particularly popular method. Topic modeling is a process for identifying the underlying themes or topics present in a collection of text documents, making it a versatile method (Boyd-Graber et al., 2017; Grimmer et al., 2022), that has been applied to various areas, such as journalism studies, political communication, international relations, political science, or migration studies (Heidenreich et al., 2019; Jacobi et al., 2016; Lucas et al., 2015; Roberts et al., 2014; Watanabe & Zhou, 2022). A recent literature review by Chen and colleagues (2023) emphasized the importance of topic modeling for communication science, arguing that it is "an effective and innovative

tool for many communication researchers" (Chen et al., 2023, p.1), due to the abundance of digitized text data. The popularity of topic models can be linked to their vast applicability and cost-effectiveness. As a bottom-up approach, it can help researchers identify latent structures within large volumes of text (DiMaggio et al., 2013), detect new categories or concepts (Nelson, 2020), and overall infer meanings from text (Chen et al., 2023; Grimmer et al., 2022).

Topic modeling was first developed in the computer sciences but was quickly adopted by scholars in the core social sciences. Nevertheless, there are considerable differences between the computational and the social sciences. As Wallach (2018, p. 4) puts it: "[C]omputer scientists may be interested in finding the needle in the haystack [...], but social scientists are more commonly interested in characterizing the haystack." This fast adoption of a new methodology to a different field required novel methods of validation that would be suitable for social scientific problems. While some could be borrowed from computer science, they would not always fit social scientific applications (Baden, Pipal, et al., 2022). Furthermore, topic modeling methods are being further developed. For example, the introduction of neural networks into topic modeling (Zhao et al., 2021) creates the necessity to continue developing validation methods to account for new modeling approaches and their uses in the social sciences. Often, however, the validation approaches have lagged behind the widespread use that topic models have enjoyed (e.g., Baden, Pipal, et al., 2022).

Needless to say, given the largely automated and inductive nature of the process, it is particularly crucial to validate the outcomes and interpretations of topic models to ensure their accuracy and scientific veracity. First and foremost, a lack of validation practices is problematic from a scientific point of view, as missing validation signifies a lack of scientific rigor (Scharrer & Ramasubramanian, 2021). Second, a lack of validation practices complicates the use of topic modeling for theory building (Grimmer et al., 2022; Ying et al., 2022) as well as giving policy recommendations (Baden, Dolinsky, et al., 2022). Third, neglecting validation gives way to criticism and skepticism around the use of computational methods in the social sciences more generally (Bonikowski & Nelson, 2022). Some have already voiced such critique (Margolin, 2019) due to a lack of transparency and uncertainty around applications and outcomes (DiMaggio et al., 2013).

During the past years multiple scholars have raised their concerns about a lack of clear patterns toward validating topic models (see for example Baden, Pipal, et al., 2022; Hoyle et al., 2021). The computational social science community has started to respond to these claims of lacking standardization, with studies providing first roadmaps to using topic modeling in the social sciences, more generally, (Chen et al., 2023; Maier et al., 2018) as well as first studies discussing topic model evaluation, specifically (Bernhard et al., 2023; Harrando et al., 2021; Ying et al., 2022). However, these studies look only at specific subfields or specific evaluation tasks. In light of this, we argue that a systematic overview of validation methods applied to topic modeling is still lacking. We thus propose a thorough and systematic review of research applying topic models to assess the alleged lack of standardization of validation methods in the field. This study inductively analyses 792 substantive and methodological studies applying topic modeling pertinent to the social sciences. We shed light on the following research question: Is there a convergence towards a gold standard of validation methods for topic modeling? To do so, we will first consider which methods are applied regularly and whether there are changes over time, as well as

whether there are combinations of validation methods that are applied more often than others. Additionally, we take into account potential methodological differences between research published in the different fields of interest (i.e., core social sciences and peripheral social sciences), and analyze them separately as well. Such an overview will make visible the breadth of potential validation methods that exist for topic modeling, thus serving as a benchmark against which researchers can compare their work. Additionally, tracking the frequency with which different validation methods are used over time can help identify emerging trends in research applying topic models. Awareness of which validation methods are widely accepted can foster consensus within the research community. This could eventually lead to more standardized practices and more efficient resource allocation. Last but not least understanding the prevalent validation methods can guide the education and training of students and researchers interested in applying topic models to social scienctific research questions. In sum, this paper should provide researchers with the information needed to navigate the – rather complex – landscape of topic modeling validation techniques.

On Validation

Validity in the social sciences, is concerned with the accuracy and scientific veracity of measures and by that, as well, of research results and downstream conclusions and recommendations. In simple terms, validity very generally refers to the question of whether measures actually measure what they are designed to measure. Therefore, the quest for validity underpins the very essence of scientific progress, also serving as a cornerstone for the construction of reliable knowledge upon which impactful and effective policies and interventions can be built. Various types of validity are considered in social science, including *face validity* (aligning with common understanding), *criterion-related validity* (logical connection with external variables), and *content validity* (representing the full concept's meanings) (Scharrer & Ramasubramanian, 2021). Since validation is important for all methods in social science, many differing terminologies and sub-dimensions have been developed. In the following, we will specifically discuss those concepts central to content analysis overall and topic modeling in particular.

When it comes to manual content analysis, Krippendorff (2013) uses a threefold classification of validity into *face*, *social*, and *empirical* validity. Face validity is understood as a result being plausible. Social validity, here, refers to a meta-perspective addressing the question of whether a scientific inquiry and measurements connected to it have societal relevance. Of more central concern for the current study, empirical validity is further differentiated into three sub-dimensions. First, there is the sub-dimension *content validity* (see also above), which, here, also includes questions relating to the appropriateness of sampling strategies. Second, he discusses the sub-dimension that he defines as *relations to other variables*, which is similar to the aforementioned criterion-related validity. The third sub-dimension, *construction and use*, relates to the internal structure of measures, which includes taking a look at the structural correspondence between available data or established theory and the modeled relationships, and demonstrating functional correspondence between what a content analysis does and what successful analyses have done. We argue that this detailed taxonomy of different types of validity – although developed for

manual content analysis – can also function as a guide when thinking about how to classify validation methods in topic modeling.

Literature on validation in automated content analysis is particularly concerned with quality assurances regarding human annotation as the 'gold standard' or 'ground truth' that is used in dictionary (lexicon-based) approaches and supervised machine learning approaches (Birkenmaier et al., 2023; Grimmer et al., 2022; Song et al., 2020). As an inductive approach, topic modeling, however, cannot rely on such 'ground truth' measurements to be compared against. When DiMaggio and colleagues (2013) write about different perspectives on topic model validation, specifically, they refer to what they call *semantic or internal validity* – defined as whether the model meaningfully discriminates between different meanings of the same or similar terms (i.e., similar to content validity and validity on internal structure) as well as *external validity*, which is similar to previous ideas of criterion-related validation. With topic modeling, furthermore, being a statistical approach to content extraction (see also Laver et al., 2003; Lowe, 2008), importantly DiMaggio and colleagues put a novel emphasis on the critical role of *statistical validity*, which assesses if the model specification inherent to the specific topic modeling approach is appropriate for the data at hand.

On Validation of Topic Models

As described above, topic modeling approaches are inductive, and most are unsupervised, which means that the data generation process and, with that, model outputs cannot be well assumed prior to analysis. This makes their validation less straightforward than that for supervised methods in computational content analysis (Grimmer et al., 2022). To make matters even more complicated, previous studies have shown that specific decisions relating to pre-processing (Denny & Spirling, 2018; Tolochko et al., 2024), vocabulary choice (Maier et al., 2020), as well as model selection (Bernhard et al., 2023), can lead to tremendous changes in the model results. Validation is thus important both ex-ante (i.e. to decide which topic modeling algorithm should be applied) and ex-post (i.e. to evaluate the model's performance in relation to its designated task) (Gentzkow et al., 2019). However, the degrees of freedom in pre-processing and hyperparameter settings that researchers tend to have, combined with the fact that topic models learn and assign documents in one step, place particularly high importance on the post-hoc validation of the models in connection to their results.

In a first hands-on user guide, Maier and colleagues (2018) provide an overview of how to use and evaluate LDA topic models in communication research. Importantly, they also discuss a wide range of validation approaches, including coherence metrics, qualitative expert judgment in the first step of model selection, as well as statistical validation, interpretability checks, document-topic relationships, and hierarchical clustering for mergeable topics on model validation post hoc. Notwithstanding these first efforts, Baden and colleagues (2022) have recently criticized the sustained emphasis on technological advancements over validation concerns in computational text analysis methods, including topic modeling, when it comes to the field of computational social science as a whole. While first user guides to topic model validation may, therefore, exist, it is unclear whether or to

what extent researchers follow them. Moreover, concerning the increasing use of computational methods, and in particular topic modeling, in theory-driven research, researchers have criticized that computational social science studies suffer from a lack of social scientific contextualization (Baden, Pipal, et al., 2022; Bonikowski & Nelson, 2022). Neglecting validation in the face of technological advancements, therefore, makes it difficult to evaluate the methodological soundness of topic modeling studies, build theories, or make policy recommendations based on their model outputs (Baden, Dolinsky, et al., 2022). Thus, especially studies posing substantive research questions, which measure social constructs, are at risk of misinterpreting the results they get from topic models. However, methodological research building on and further developing topic modeling approaches, as well, is in need of proper validation, in order to ensure (among other things) generalisability and comparability of the methods.

Due to the inherent abstraction in computational analyses, the call by researchers for the establishment of well-defined and universally accepted validation standards becomes even more self-evident when studying more latent social science concepts (Jacobs & Wallach, 2021). In addition to the lamented insufficient emphasis and reporting on validation practices full stop, researchers also criticized the absence of agreed-upon methods or benchmarks applied across studies utilizing topic modeling more generally. The "validation gap" (Baden, Pipal, et al., 2022) for topic models is thus argued to be accompanied by a "standardization gap" (Hoyle et al., 2021). In fact, there are also no standardized forms for the reporting of the method or its validation (Reiss et al., 2022), thus introducing a "reporting gap" as well.

In the past decade, many validation methods and metrics have been proposed and put to use. At first glance, however, persistent criticisms would suggest that no convergence was achieved and that this abundance of possibilities has made it even more difficult to develop points of comparison between studies and to judge the quality of the models and subsequent model outcomes. Given the widespread use of topic models and the plethora of proposed validation methods, it is high time for a systematic overview of applied validation strategies.

Of general interest to the current investigation would be to understand if, in the two decades of topic modeling application, researchers have started to come up with a set recipe for validating topic models, or at least whether certain validation techniques or combinations of validation techniques tend to be more favored over others when using topic modeling depending on the designated task. In other words, has the field started to converge to a common validation framework of topic models, and are there first signs that universally accepted standards of topic modeling validation are starting to develop?

Methodology

To address our research question, we conducted a comprehensive systematic literature review encompassing all studies using topic modeling pertinent to the social sciences. We recognize that the computational social sciences are still a young field and that topic modeling has only recently been adopted by the core social sciences (e.g., communication science, political science, and sociology), while it has a rich history also in other disciplines (e.g., computer sciences and linguistics), who are also applying topic modeling to research

questions relevant for studying human-generated text in the periphery of social sciences. Thus, we deliberately broadened our systematic review scope beyond the confines of social science journals. We argue that this inclusion of studies beyond the field of the core social sciences adds interesting information to the study of topic model validation, as we will take into account possible differences between methodological and substantive papers.

Sampling

To create our sample, we searched for relevant studies in four scientific databases (Web of Science, Mass Media Complete, ACM Digital Library, and EBSCOhost (Communication & Mass Media Complete, Humanities Source Ultimate, SocINDEX)) using a search string, which looked for different spellings of "Topic Model" as well as "Topic Modelling" in the title, Abstract or Keywords of a text. Additionally, we specified that "valid*" needs to be found at least once in the full text. We did not limit our search regarding the publication date so that we could give an overview since the introduction of topic models in social research. Initially, this search yielded 1,556 studies. In the first step, we coded the entire sample to assess which studies would be relevant for our review.

Regarding formal characteristics, studies were excluded if they were (I) not accessible to the coders, (II) duplicates of already selected studies, and (III) texts like extended abstracts, posters, presentation slides, panel descriptions, or studies without empirical analyses. Furthermore, studies were excluded if (IV) not at least one of the text corpora used in the studies was based on human-generated speech or (V) if the term "valid*" was mentioned in the paper but did not actually refer to the topic model or its output. After reviewing, 792 studies met our criteria and were included in the sample for further analysis. The earliest study was published in January 2004, and the most recent one was published in March 2022, which coincides with the month of data collection [1].

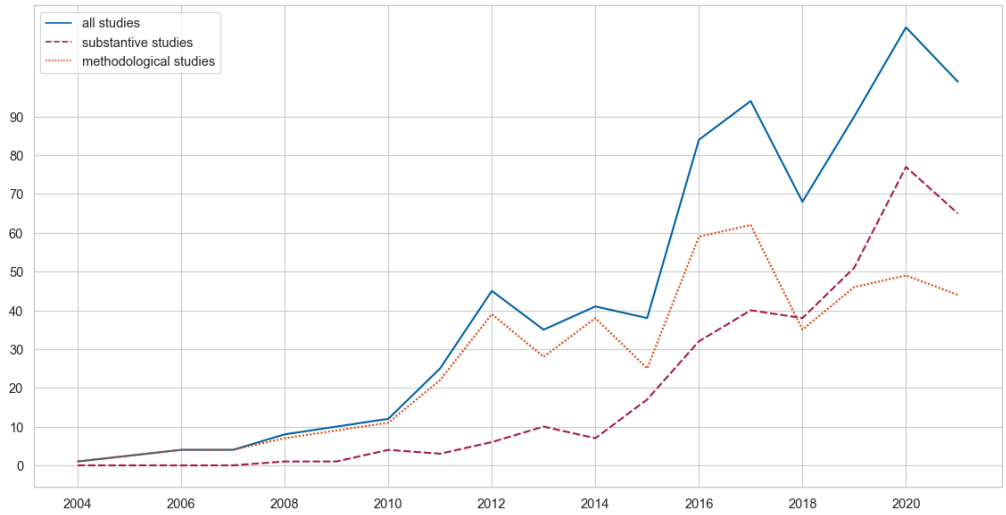
Description of the Dataset

Figure 1 shows how the number of topic modeling studies pertinent to the domain of the social sciences has increased over the past two decades. We find that the number of studies with substantive research questions has been more or less steadily increasing since 2011. The number of studies focusing on methodological advancement, however, peaked in 2017 before dropping and plateauing for the upcoming years. Overall, more studies were published in conference proceedings (54%), than in journals (46%). Top publication outlets, suggest that a considerable number of studies pertinent to this investigation are (still) published in computational science journals and conferences (see Table 3 in Appendix C).

[1] Please see Appendix C and osf.io for our sample and replication data

Figure 1

Number of Studies with a substantive or methodological goal in our sample over time.



Note: As only a quarter of the year 2022 is included in the sample we did not include it in the graph, as it would have resulted in a misleading trend.

Codebook and Manual Classification

The identified studies were then coded by two of the authors. We first annotated whether the studies had substantive research questions from the social sciences and/or focused on methodological advancements in topic modeling. Note that these two categories are not mutually exclusive. Our main variables of interest, however, were the validation methods mentioned and applied by the original study authors. Here we employed an inductive coding scheme where we added columns each time a method was mentioned for the first time. Given the absence of a comprehensive systematic review on topic modeling validation methods to date, the lack in standardization in how to report validation methods, and the ongoing debates and developments in this area, we believe that an open and inductive approach to coding validation methods is needed. Rather than relying on a potentially incomplete pre-defined list of validation methods, our methodology, therefore, involved coding any approaches mentioned by study authors and deemed relevant to the validation of their topic models and topic model outputs [2].

This approach allowed us to capture the breadth of practices and strategies in this dynamic and evolving field but has two key implications. Firstly, we might have incorporated methods that some researchers might not consider adequate for topic model validation, yet were communicated by the study authors as a means to validate their approach. Secondly, our analysis focused solely on the validation methods explicitly articulated in the

[2]The studies obtained from the EBSCOhost database have been added at a later stage in response to a reviewer comment during peer review. Thus, these 62 studies were coded deductively based on the categories identified during the inductive coding of 730 studies sampled through the other databases that were used for initial submission of the manuscript.

final manuscripts, potentially overlooking crucial, albeit unreported, steps.

In the initial phase of inductive coding, we identified a total of 445 distinct methods that were mentioned as pertaining to topic model validation. Subsequently, a meticulous refinement process was implemented, addressing instances where different terminologies denoted identical procedures. We, furthermore, amalgamated closely related approaches[3]. This consolidation led to a streamlined set of 138 validation methods. Recognizing the need for a more manageable framework, we further pruned the list by excluding validation methods that accounted for less than 1% of our sample, resulting in a more practical and focused compilation of methods. The remaining 32 validation methods were then grouped into eight overarching categories based on their shared methodological perspective: *Model Comparison*, *Internal Qualitative Inspection*, *External Qualitative Inspection*, *Error Rate Analysis*, *Distinctiveness of Top Words*, *Information Theory Metrics*, *Similarity and Distance Metrics*, and *Downstream Tasks*. These eight categories can also be understood in relation to the theoretical perspectives on validation introduced earlier in this study, while *Model Comparison* can be connected to Krippendorff's (2013) understanding of internal validity, *Internal Qualitative Inspection* is understood in the sense of Krippendorff's understanding of face and context (semantic) validity or DiMaggio and colleagues (2013) internal validity, while *External Qualitative Inspection* can be related to internal and relational validity (for Krippendorff) and external validity (for DiMaggio et al). The four groups comprising statistical measures *Error Rate Analysis, Distinctiveness of Top Words, Information Theory Metrics as well as Similarity and Distance Metrics* are difficult to characterize in Krippendorff's assessment, as he was focusing on manual analysis, however they can be connected to what Dimaggio and colleague's (2013) understand as statistical validity. The last category *Downstream Tasks* is again related to Krippendorff's relational validity.

These eight overarching categories synthesize the numerous, often highly specific approaches commonly associated with validation. They comprehensively encompass previously suggested types of validation, providing a broader and overarching classification of the diverse methods for validating topic models that are pertinent to social research and are widely used in practice. Employing our inductive approach, we successfully grouped all the methods used into these eight overarching categories. It should be noted that intercoder, as well as intracoder reliability, was assessed using Krippendorff's alpha, revealing satisfactory levels of agreement. Please refer to Table 1 in Appendix A for detailed information. For an overview of the mentioned validation methods corresponding to each of the categories above, please see Table 2 in the section below.

Results

Our comprehensive review uncovered a diverse array of approaches to topic model validation. Subsequently, we will delve into these approaches in greater detail, emphasizing their development over time. We will present our findings according to the frequency with which each category was mentioned in our sample (see Figure 2 for a visualization [4] and Table 2 in Appendix B, for an overview).

[3]For example, we aggregated methods that are variations of each other, such as micro- and macro Precision, or different k-fold splits.

[4]For a visualization of these categories, which presents them separately for substantive and methodological papers, see Figure 7 in Appendix B.

Most studies (61.4%) in our sample mention at least one validation method pertaining to the overarching validation category of *Comparing Models*. This category encompasses instances where authors emphasize having executed various types of topic models or specified their topic models differently to determine the most suitable approach for the specific task. The decision to either use a single model or compare multiple models should be regarded as a precursory step preceding all subsequent methods of validating topic modeling. Of course, the basis for comparing topic model outputs must derive from one of the other seven overarching categories. As methodological advances in topic modeling tend to require comparisons to pre-existing modeling techniques, this category is mentioned more often in methodological studies (77.9%) as compared to substantive studies (37.5%) [5].

In second place, we find *Internal Qualitative Inspection* (54.3%), aiming at internal, face, and content validity. This category entails the application of qualitative methods to evaluate the quality and relevance of topics generated by a topic modeling algorithm, relying solely on the model's output. Common practices within internal qualitative inspection include assessing the plausibility of topics, which heavily relies on the concept of face validity. Substantive studies rely much more on these methods (76.9%) than studies with a methodological focus (38.3%).

A bit under half of the studies (44.4%) in our sample use different kinds of *Error Rate Analysis*, which is based on the assessment and quantification of the performance of the model by comparing its predictions against a ground truth (oftentimes manual annotations based on a deductive coding schema). These metrics help in understanding the model's effectiveness and its ability to make accurate predictions. Metrics of *Error Rate Analysis* include well-known statistics such as calculating Recall, Precision, or the F-Score. While only 22.6% of substantive studies employ some kind of error rate analysis, 60.3% of methodological studies do so.

Still, 40.4% of the studies argue that they evaluate the validity of their model and its output by applying it to a specific *Downstream Task*. *Downstream Tasks* in the context of topic model validation refers to the assessment of the effectiveness and utility of topic models by evaluating their performance in tasks that depend on the output of these models. Instead of concentrating on the internal characteristics or outputs of the topic models, this approach assesses the contribution of the generated topics to the success of subsequent tasks, like, for example, serving as a covariate in a regression analysis. This method is more widespread for substantive research (49%), as compared to methodological research (36.2%).

A quarter of studies (24.4%) in our sample rely on validation methods building on statistical validity, using *Information Theory Metrics*. These are statistical measures used to assess the quality, uncertainty, and information content of topic models. These metrics help quantify the differences between probability distributions and evaluate the efficiency and accuracy of the models. Included in this category are the Jensen Shannon as well as Kullback-Leibler Divergence, Perplexity, and Entropy. 28.7% of methodological studies are validated with methods from this group, while only 19.3% of substantive studies apply the same methods.

[5] Studies can have both a substantive and a methodological focus. Thus, we give the prevalence of each category in percentage and not frequency, as the numbers, in this case, would not add up to the number of studies in our sample.

The category of *External Qualitative Inspection* is applied by 22.3% of studies. *External Qualitative Inspection* involves qualitatively evaluating the meaningfulness and relevance of topics, explicitly leveraging model-external information, such as theoretical assumptions or real-world contexts, including events or dynamics. Some studies also compare topic modeling outputs with inductively human-annotated subsets of the text corpus. Similar to *Internal Qualitative Inspection*, substantive studies apply these more frequently (32.2%) than methodological studies (15.2%).

Metrics relating to *Distinctiveness of Top Words* are referred to in 22.3% of all studies analyzed. This category comprises methods that evaluate statistical validity based on the uniqueness and quality of high-probability words (i.e., Top Words) within the topics generated by a topic model. Such measures aim to evaluate the meaningfulness and relevance of the top words within each topic. This category is used by substantive and methodological studies at a similar rate (substantive: 22.6% and methodological 22.1%).

Finally, *Similarity and Distance Metrics* are the smallest overarching category, present in 8.3% of studies overall. In the context of topic models these metrics quantify the degree of (dis)similarity between topics, documents, or words, enabling the evaluation of the relationships, overlaps, and distinctiveness within the generated topics. Related metrics are, for example, the Jaccard Coefficient, or Silhouette. Again, this category is distributed similarly across all studies, 7.7% in substantive studies and 9.5% in methodological studies.

Next, we examine how the salience of different categories of validation methods changed over time (see Figure 3). Given the limited number of cases in the initial years of our dataset and to ensure meaningful comparisons, we narrowed our focus to a subset of the data, commencing from 2011 (n = 753). Three discernible trends emerge: categories that exhibited consistent use over time, those that experienced a decline in usage, and those that observed an increase. Among those that have been used consistently over time are *Downstream Tasks* as well as *Similarity and Distance Metrics*. Categories such as *Error Rate Analysis*, *Information Theory Metrics*, and *Model Comparison* have lost in popularity over time. For instance, *Model Comparison* was referenced in 80% of studies in 2011, but only in 60% of the studies in 2022. Similarly, *Error Rate Analysis* decreased by nearly 30 percentage points, from being utilized in 60% of studies to only a third in 2022 Lastly, *Information Theory Metrics* appeared in almost half of all studies in 2011, but only in every fifth study in 2022. Conversely, validation methods such as *Internal Qualitative Inspection*, *External Qualitative Inspection*, and *Distinctiveness of Top Words* approaches are experiencing growing prominence. The *Internal Qualitative Inspection* has surged from under 40% to almost 70%. The growth in studies mentioning *External Qualitative Inspection* is even more striking, starting at than 10% in 2011 and surging to nearly 25% within a decade, possibly speaking to the increasing adoption of topic modeling particularly by researchers trained in social sciences rather than computer sciences. The validation methods pertaining to the *Distinctiveness of Top Words*, as well, have witnessed a large increase, initially being employed in fewer than 10% of studies and now featuring in over a third of the studies.

Figure 2

Percentage of Studies employing validation methods

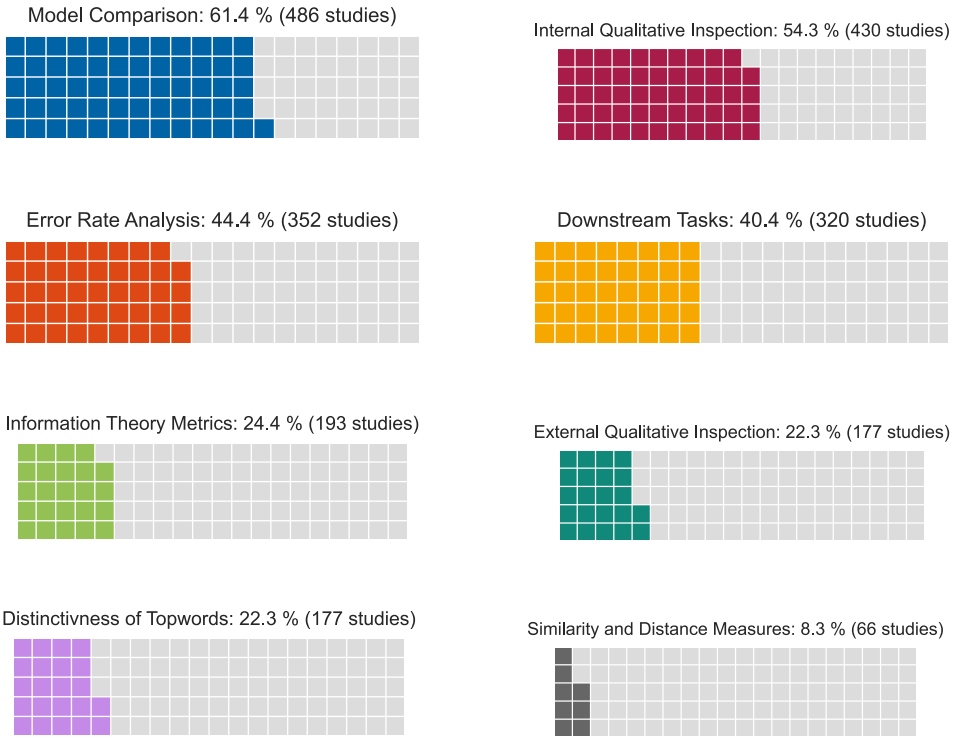
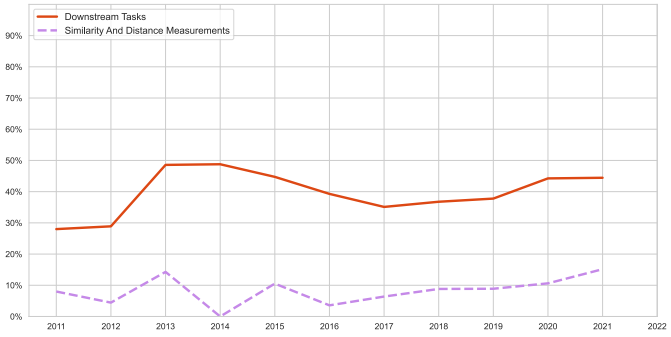


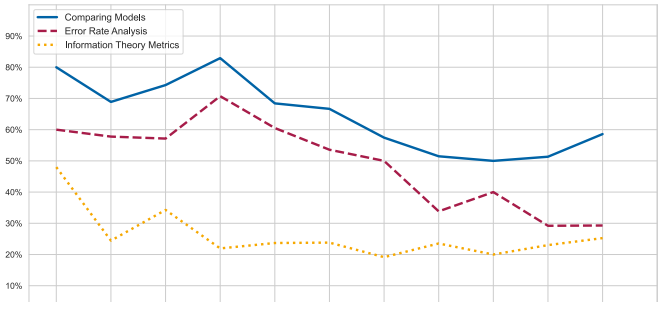
Figure 3

Changes in the Application of Validation Categories over Time

(a) *Validation Categories that are consistent over time.*



(b) *Validation Categories that decrease over time.*



Numerous scholars have underscored the importance of accommodating diverse perspectives validation as a prerequisite for appropriate topic modeling validation (Chen et al., 2023; DiMaggio et al., 2013; Krippendorff, 2013; Maier et al., 2018), thereby emphasizing the necessity of incorporating a synthesis of various validation categories in research studies. Aiming to shed light on the extent to which researchers acknowledge and integrate diverse validation perspectives within their studies, we first look at the average number of validation categories within each study: On average, each study incorporates approximately three overarching validation categories, as illustrated in Figure 4). The average number of individual validation methods used in a single study is four, with no significant disparities observed between substantive and methodological studies [6]. Overall, we find no discernible trends suggesting a growing inclination toward the integration of a higher number of validation perspectives in conjunction within single studies.

Figure 4

Average number of validation categories used per study over time.

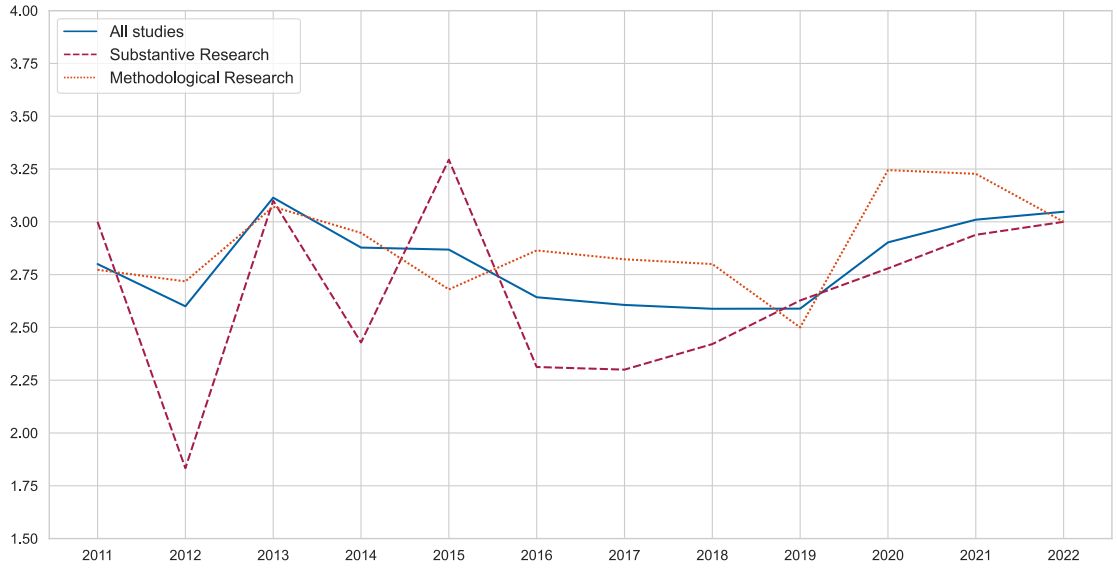


Figure 5 illustrates the extent to which different validation categories co-occur in individual studies within our dataset. We acknowledge that we limit our examination to dyadic combinations, given the complexity introduced by the increasing number of possible combinations (e.g., triads). Despite this limitation, the comparison reveals intriguing insights. For instance, 81.3% of the studies mentioning validation methods related to *Error Rate Analysis* compare these metrics by running different *Model Comparisons*. Note that the diagonal represents the number of studies that pertain to a method within the specific category mentioned in the column without any reference to also having used validation methods from other overarching categories.

As expected, there are notable overlaps between methods relying on quantitative

[6]Please note that the average number of validation methods should be interpreted carefully, as this is strongly dependent on how we summarized them.

metrics, and studies using these categories often employ approaches related to *Model Comparison*. Conversely, studies relying on *External Qualitative Inspection* heavily leverage on *Internal Qualitative Inspection* (76.8%), as well. At the same time, *Internal Qualitative Inspection* emerges as the overarching category of validation methods most frequently used in isolation, without reference to methods from other categories (9.1%).

Figure 5

Dual Co-Occurrences of two Validation Categories in percent.

	Model Comparison	Internal Qualitative Inspection	Error Rate Analysis	Downstream Task	Information Theory Metrics	External Qualitative Inspection	Distinctivness of Topwords	Similarity & Distance Measures
Model Comparison	3,7%	48,6%	81,3%	55,9%	74,1%	41,2%	66,7%	59,1%
Internal Qualitative Inspection	43,0%	9,1%	34,9%	64,7%	52,8%	76,8%	62,7%	54,5%
Error Rate Analysis	58,8%	28,6%	4,3%	37,8%	44,0%	29,4%	39,0%	45,5%
Downstream Task	36,8%	48,1%	34,4%	3,8%	34,7%	49,2%	41,2%	43,9%
Information Theory Metrics	29,4%	23,7%	24,1%	20,9%	4,7%	20,9%	26,0%	37,9%
External Qualitative Inspection	15,0%	31,6%	14,8%	27,2%	19,2%	3,4%	27,1%	25,8%
Distinctivness of Topwords	24,3%	25,8%	19,6%	22,8%	23,8%	27,1%	0,6%	30,3%
Similaity & Distance Measures	8,0%	8,4%	8,5%	9,1%	13,0%	9,6%	11,3%	1,5%

Note: The diagonal marks the share of studies that include only validation methods from one validation category.

While we have demonstrated the evolution of the importance of certain overarching categories of validation methods over time and the frequent combination of various perspectives in topic modeling validation, it is intriguing to measure the possibility of convergence in these combinations. Specifically, we seek to understand whether, as time progresses, there is a tendency within the field to increasingly agree on specific combinations of validation categories rather than others. In order to assess this possibility, we calculate *Information Entropy* [7], a metric from information theory that estimates the diversity within a community, and plot the values over time (see Figure 6). Information entropy is a measure of "surprise" of seeing another data point. In other words, when reading a random topic modeling paper, we would expect that the most "agreed upon" dyadic combination of validation methods is most likely used. If researchers using topic models would converge on some set(s) of validation approaches that are more "standard" than others, we would see

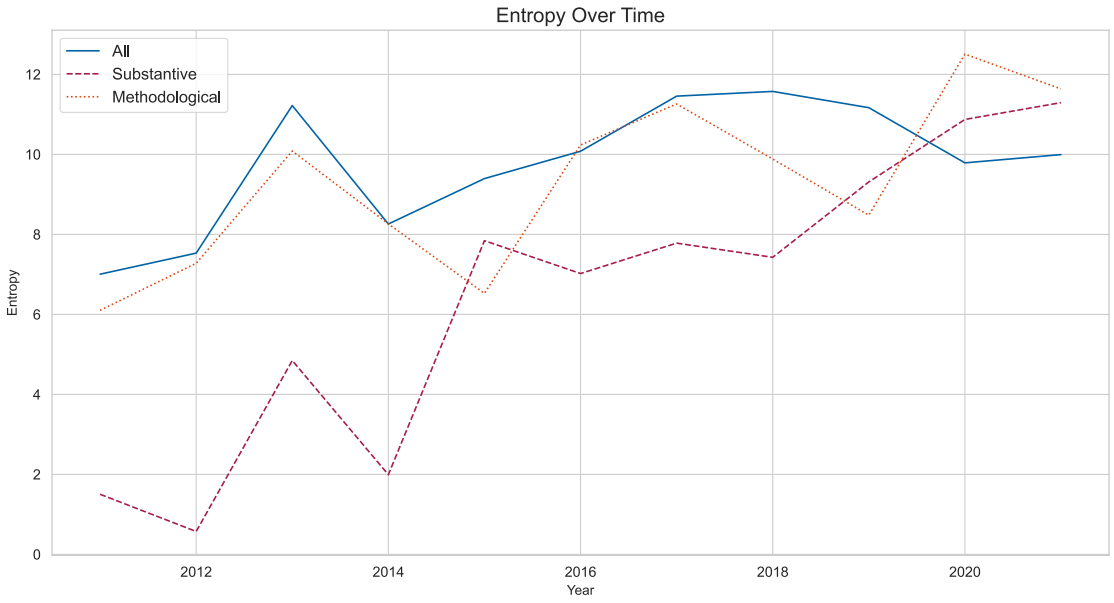
[7] In instances where one of the categories or methods was not present, Laplace smoothing was applied to ensure the calculation of Shannon's entropy. A pseudocount value of 0.0001 was imputed for missing categories or methods to avoid division by zero.

the value of *Information Entropy* decreasing over time.

However, the observed values remain relatively consistent over a decade, suggesting that the diversity of dyadic combinations of validation perspectives employed in studies has not significantly decreased over time [8]. Consequently, no observable convergence of the field towards a specific combination of validation categories is evident. A similar pattern emerges when considering all 32 validation methods instead of the overarching categories (see Figure 8 in Appendix B). However, it is noteworthy that there was actually a trend towards divergence in validation methods, particularly for substantive studies, until 2016 [9].

Figure 6

Information Entropy for binary validation categories over time



Despite the widespread adoption of various validation techniques, often spanning distinct categories and examining different facets of model validity, our investigation uncovers a distinct absence of dominant or widely accepted combinations. Curiously, no overarching trends or consensus practices regarding the amalgamation of these diverse validation methods become apparent, highlighting the current lack of standardization of topic modeling validation approaches and no clear signs of convergence in topic model validation more generally.

[8] For an additional robustness check we further categorized all relevant studies into studies from the field of core social sciences and the social science periphery to see whether convergence is happening in either of the fields, however, again we could not find any indication of convergence, please see Appendix E for details

[9] The dip regarding the substantive studies in 2012 might stem partly from the low number of studies in this year.

Discussion

Validation is an important part of scientific research. Thus, the critique by many scholars that computational methods, especially topic modeling, lack validation has set into motion a number of projects aimed at enhancing our understanding of topic model validation. While some scholars have focused on presenting a road map (Maier et al., 2018), others have highlighted the process in general (Chen et al., 2023) to see what has been done in the field. However, the notion of missing standards has been a point of discussion at many times (e.g., Baden, Pipal, et al., 2022; Maier et al., 2018). What was missing so far from the literature was a systematic review that approaches validation in a long-term and inductive perspective, which allows us to account for the plethora of possible ways to validate topic models. This is especially important as it allows us to test claims of missing standards while taking into account how topic model validation has changed in the computational social sciences and may have converged in standard routines.

We find that in 20 years of applying topic modeling as a method in the computational social sciences, there have neither been clear signals pointing towards standardization of validation practices nor first signs of convergence towards specific validation methods over time. These findings hold true when looking at overarching validation categories as well as when focusing on specific validation methods. Moreover, while some validation approaches more strongly rely on a combined use with other approaches, we could not find any emerging convergence in terms of dyadic combinations of validation perspectives.

However, as evident from this literature review, this lack of convergence is not necessarily indicative of a lack of trying. Perhaps, this inability to converge on a strict set of quantifiable validation criteria comes from an inapt approach to the problem. Unlike classical statistical methods (e.g., a regression analysis), topic models, like many noted before, are an inherently inductive approach to data analysis. Instead of imposing preconceived notions or hypotheses onto the data, topic modeling attempts to uncover hidden thematic structures within texts. This inductive nature means that topic modeling is by nature exploratory, aiming to reveal latent patterns and topics that might or might not be apparent through deductive analysis alone.

Classical statistical methods are validated based on how well they fit a predefined hypothesis. The core idea of topic models, by contrast, lies in their ability to uncover hidden structures and emerging themes in the data - a process that, almost by definition, is unsuitable for a deductive validation paradigm that strongly rests on standardization for the sake of comparability.

This distinction becomes even more apparent when thinking about the assumptions one has to make when dealing with classical methods versus topic models (or any other unsupervised machine learning algorithm). In classical deductive analysis, there is a foundational assumption that a singular "real" data-generating process does exist that researchers aim to approximate as closely as possible. Under this assumption, validation techniques can be designed to measure the model's ability to capture this singular underlying truth. Classical validation methods in regression, for example, such as cross-validation and goodness-of-fit tests, are tailored to assess how well the model aligns with this assumed reality. Under this assumption, however, topic models are an objectively wrong way to analyze the data. Different topic modeling validation solutions can be valuable for distinct

purposes. For instance, one solution might be optimal for summarizing documents, while the other for analyzing latent themes over time. Also, depending on the corpus at hand and, importantly, *independent of the hypothesis*, a different number of topics may be appropriate for the data. The usefulness of a topic model is deeply dependent on the specific research question, dataset, and expected outcomes of the analysis. This diversity in utility renders the notion of a single, universal "real" data-generating process inappropriate.

Thus, we argue that traditional validation methods rooted in the assumption of a singular "truth" cannot accommodate the multifaceted nature of topic modeling solutions. The appropriateness of a topic model is contingent upon its application context, making it impractical (and, potentially, impossible) to devise a one-size-fits-all validation framework. We here suggest that it is this mismatch, that makes it so incredibly difficult to find standard approaches, or at least a convergence towards a few, select methods or even categories of validation methods.

This argument is not to say, that we should forgo topic modeling validation altogether. It does mean, however, that we need to change our way of thinking about it. Especially, when utilizing topic models without "ground-truth" data, instead of understanding validation as hitting a specific cut-off point for F1-Scores, we might focus our attention towards more qualitative interpretations of validation. Humphreys and colleagues (2021), for example, define validity for qualitative research after Kirk and Miller (1986, p.20) as “the degree to which the finding is interpreted in the correct way". In their opinion, this could mean heightened transparency throughout the whole process (Dienlin et al., 2021) as well as including "thick descriptions" (Humphreys et al., 2021, p.857) of how the interpretation came to be, triangulating as well as different perspectives, especially also from people with lived expertise. Integrating the methodologies from qualitative research into computational methods, such as topic modeling, can help us face the inevitable need to validate what we find.

Importantly, Humphreys and colleagues (2021, p. 857) note that "it is important to recognize that differences in methods call for different kinds of validity- and credibility-enhancing research practices". In a similar vein, Barberá and colleagues (2021, p.40) assess that "every research question and every text-as-data enterprise is unique", and thus also call for validation decisions to be adapted to each individual research project. This mirrors our findings that there are many validation methods that can be applied in order to assess some aspect of validity.

What we can recommend at the end of our review is that computational social scientists should familiarize themselves with qualitative validation criteria and more openly embrace the inductive nature of topic modeling. This might mean realizing that topic modeling is not an ideal method for measurement (e.g. Grimmer et al., 2022), and thus requires putting extra effort into explaining, why and which validation method is chosen, what it tells us, and why this should bolster the credibility of our findings and interpretation. In this, it is important that researchers do not fall into the trap of arguing ex-post, based on the results, that a hypothesis can be accepted or rejected, if they use the same logic to validate their topic model (i.e. Validation via Downstream Tasks). Instead, as it is more appropriate for qualitative research, we should use the outcome to formulate possible hypotheses, which can be tested in a subsequent step, with a different method. We realize that readers might expect clear guidelines or blueprints on how to validate their topic

models at this place. However, the lack of convergence and lack of clear patterns as to which methods are actually applied by researchers suggest that there is no one way to validate topic models. We stress that it is the responsibility of each researcher to decide what kind of validation their particular research question warrants. Elsewhere, Bernhard and colleagues (2023) have given an outline of possible questions to ask yourself when validating topic models, which can help as a guide for deciding on a validation strategy and Maier and colleagues (2018) specifically propose a guideline on validating LDA models.

We argue that similar to how topic modeling usefulness is dependent on the use case, so should the validation procedures be. Striving for a standardized set of validation criteria, applicable universally across diverse use cases, is perhaps a misguided attempt. Depending on the task of a particular topic model in a particular research design, practitioners should choose the validation method that is best aligned with the task. This emphasis on context-specific validation methods fosters transparency in the research process. By explicitly reporting how the model was validated and detailing for what specific reasons this validation was chosen, researchers provide readers with a deeper understanding of the methodology's appropriateness and relevance to the research objectives. This clarity not only enhances the credibility of the findings but also allows readers to assess the validity of the model's outcomes within the context of the study.

Again, it is important to note, that we do not argue against convergence or standardization of validation strategies per se. Convergence, in the sense of shared methodologies, holds value and is important to the advancements in the field. However, we believe that striving for methodological convergence must not overshadow the need for tailored, task-specific validation in the realm of topic modeling. Unlike in classical quantitative research methods where standardized approaches often prove effective, the diverse nature of textual data and the multifaceted objectives of topic modeling necessitate a more nuanced and adaptable approach to validation.

Overall, scientific progress can only be made if we choose the right method for the research question. This includes taking into account the strengths and weaknesses of a method and catering towards the former. Topic modeling is a computational text analysis method, which helps us *inductively* find patterns in large amounts of text. With the recent surge of digitally available text, it is a great method for identifying underlying themes and thus, gaining insights into rich data. Transparent research and detailed reporting on the application of the method and interpretation of the findings can be an important first step toward credible findings. Applying fitting validation methods, ideally around different validation perspectives can reinforce the validity of a research study using topic modeling.

Our review has certain limitations that need to be considered. Firstly, we only focus on studies using topic modeling that explicitly mentioned validation, which means that we may have overlooked studies that referred to validation only implicitly. It also means that our study cannot speak to the validation gap (i.e., the lack of validation) specifically, rather than the standardization gap in topic modeling research in the social sciences. Secondly, although we conducted an inductive coding process, it was not feasible to analyze all of the coded validation methods. Many methods that researchers referred to as "validation" were only present in very few studies and could, therefore, not be considered. The lack of standardized reporting (i.e., reporting gap) also may have led to an under-identification or at least to some fuzziness, when it came to categorizing some of the methods. Thirdly,

it would have been interesting to separate the analyses further between the subfields in the core social sciences (e.g., political science vs. communication science), however, this was not possible, as there were not enough relevant studies per year and subfield in our sample. Fourthly, while there is a desire for recommendations on the best validation method for a topic model, this is not something we can do in this study. We argue that, while standardization is important and researchers should not be overwhelmed with the choice of validation methods, every researcher has to do the work, to derive which validation methods would be the most applicable to a specific topic modeling use case. We do hope that our study can help get this process started.

Conclusion

In conclusion, our systematic literature review spanning two decades of topic modeling in the computational social sciences reveals a notable absence of standardized validation practices and a lack of convergence toward specific validation methods over time. This discrepancy may be attributed to the inherent inductive and qualitative nature of topic modeling, which does not align with the deductive, quantitative traditions that typically seek standardization. Building on this, we propose a shift in how we perceive validation of topic modeling, particularly in the absence of "ground-truth" data. We advocate for a more qualitative approach to validation, emphasizing the importance of correctly interpreting findings as well as transparency drawing from qualitative research methodologies. Acknowledging the uniqueness of each research question, we recommend that computational social scientists adapt their validation criteria to suit the specific context of their research projects and transparently motivate this choice. Ultimately, our review underscores the importance of selecting the right methods for the research question, understanding the strengths and weaknesses of these methods, and fostering transparency and detailed reporting to enhance the credibility of findings when employing topic modeling in the computational social sciences.

Competing interests: The author(s) declare none

References

Baden, C., Dolinsky, A., Lind, F., Pipal, C., Schoonvelde, M., Guy, S., & van der Velden, M. A. G. (2022, September). *Integrated standards and context-sensitive recommendations for the validation of multilingual computational text analysis* (tech. rep. No. Deliverable 6.2). OPTED.

Baden, C., Pipal, C., Schoonvelde, M., & van der Velden, M. A. G. (2022). Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda. *Communication Methods and Measures*, *16*(1), 1–18. https://doi.org/10.1080/19312458.2021.2015574

Barberá, P., Boydstun, A. E., Linn, S., McMahon, R., & Nagler, J. (2021). Automated text classification of news articles: A practical guide. *Political Analysis*, *29*(1), 19–42. https://doi.org/https://doi.org/10.1017/pan.2020.8

Bernhard, J., Teuffenbach, M., & Boomgaarden, H. G. (2023). Topic Model Validation Methods and their Impact on Model Selection and Evaluation. *Computational Communication Research*, *5*(1), 1–26.

Birkenmaier, L., Lechner, C. M., & Wagner, C. (2023). The Search for Solid Ground in Text as Data: A Systematic Review of Validation Practices and Practical Recommendations for Validation. *Communication Methods & Measures*, *0*(0), 1–29. https://doi.org/10.1080/19312458.2023.2285765

Bonikowski, B., & Nelson, L. K. (2022). From Ends to Means: The Promise of Computational Text Analysis for Theoretically Driven Sociological Research. *Sociological Methods & Research*, *51*(4), 1469–1483. https://doi.org/10.1177/00491241221123088

Boyd-Graber, J., Hu, Y., & Mimno, D. (2017). Applications of topic models. *Foundations and Trends in Information Retrieval*, *11*(2-3), 143–296. https://doi.org/10.1561/1500000030

Chen, Y., Peng, Z., Kim, S.-H., & Choi, C. W. (2023). What We Can Do and Cannot Do with Topic Modeling: A Systematic Review. *Communication Methods and Measures*, *17*(2), 1–20. https://doi.org/10.1080/19312458.2023.2167965

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

Dienlin, T., Johannes, N., Bowman, N. D., Masur, P. K., Engesser, S., Kümpel, A. S., Lukito, J., Bier, L. M., Zhang, R., & Johnson, B. K. (2021). An agenda for open science in communication. *Journal of Communication*, *71*(1), 1–26. https://doi.org/10.1093/joc/jqz052la

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

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, *57*(3), 535–74. https://doi.org/10.1257/jel.20181020

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.

Harrando, I., Lisena, P., & Troncy, R. (2021). Apples to Apples: A Systematic Evaluation of Topic Models. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, 483–493. Retrieved September 14, 2022, from https://aclanthology.org/2021.ranlp-1.55

Heidenreich, T., Lind, F., Eberl, J.-M., & Boomgaarden, H. G. (2019). Media Framing Dynamics of the 'European Refugee Crisis': A Comparative Topic Modelling Approach. [Publisher: Oxford University Press / USA]. *Journal of Refugee Studies*, *32*, i172–i182. https://doi.org/10.1093/jrs/fez025

Hoyle, A., Goel, P., Hian-Cheong, A., Peskov, D., Boyd-Graber, J., & Resnik, P. (2021). Is Automated Topic Model Evaluation Broken? The Incoherence of Coherence. *Advances in Neural Information Processing Systems*, *34*, 2018–2033. https://doi.org/10.48550/arXiv.2107.02173

Humphreys, L., Lewis, N. A., Jr, Sender, K., & Won, A. S. (2021). Integrating Qualitative Methods and Open Science: Five Principles for More Trustworthy Research*. *Journal of Communication*, *71*(5), 855–874. https://doi.org/10.1093/joc/jqab026

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

Jacobs, A. Z., & Wallach, H. (2021). Measurement and Fairness. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 375–385. https://doi.org/10.1145/3442188.3445901

Kirk, J., & Miller, M. L. (1986). *Reliability and validity in qualitative research* (Vol. 1). Sage.

Krippendorff, K. (2013). *Content analysis: An introduction to its methodology* (3. ed.). Sage.

Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. *American political science review*, *97*(2), 311–331.

Lowe, W. (2008). Understanding wordscores. *Political Analysis*, *16*(4), 356–371. https://doi.org/10.1093/pan/mpn004

Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-Assisted Text Analysis for Comparative Politics. *Political Analysis*, *23*(2), 254–277. https://doi.org/10.1093/pan/mpu019

Maier, D., Niekler, A., Wiedemann, G., & Stoltenberg, D. (2020). How document sampling and vocabulary pruning affect the results of topic models. *Computational Communication Research*, *2*(2), 139–152. https://computationalcommunication.org/ccr/article/view/32

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

Margolin, D. B. (2019). Computational Contributions: A Symbiotic Approach to Integrating Big, Observational Data Studies into the Communication Field. *Communication Methods and Measures*, *13*(4), 229–247. https://doi.org/10.1080/19312458.2019.1639144

Nelson, L. K. (2020). Computational Grounded Theory: A Methodological Framework. *Sociological Methods & Research*, *49*(1), 3–42. https://doi.org/10.1177/0049124117729703

Reiss, M. V., Kobilke, L., & Stoll, A. (2022, June). Reporting Supervised Text Analysis for Communication Science [DGPuK Jahrestagung der FG Methoden].

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*, *58*(4), 1064–1082. https://doi.org/10.1111/ajps.12103

Scharrer, E., & Ramasubramanian, S. (2021). *Quantitative research methods in communication : The power of numbers for social justice*. Routledge.

Song, H., Tolochko, P., Eberl, J.-M., Eisele, O., Greussing, E., Heidenreich, T., Lind, F., Galyga, S., & Boomgaarden, H. G. (2020). In Validations We Trust? The Impact of Imperfect Human Annotations as a Gold Standard on the Quality of Validation of Automated Content Analysis. [Publisher: Routledge]. *Political Communication*, *37*(4), 550–572. https://doi.org/10.1080/10584609.2020.1723752

Tolochko, P., Balluff, P., Bernhard, J., Galyga, S., Lebernegg, N. S., & Boomgaarden, H. G. (2024). What's in a name? The effect of named entities on topic modelling interpretability [Publisher: Taylor & Francis]. *Communication Methods and Measures*, 1–22.

Wallach, H. (2018). Computational social science computer science+ social data. *Communications of the ACM*, *61*(3), 42–44. https://doi.org/10.1145/3132698

Watanabe, K., & Zhou, Y. (2022). Theory-Driven Analysis of Large Corpora: Semisupervised Topic Classification of the UN Speeches. *Social Science Computer Review*, *40*(2), 346–366. https://doi.org/10.1177/0894439320907027

Ying, L., Montgomery, J. M., & Stewart, B. M. (2022). Topics, Concepts, and Measurement: A Crowdsourced Procedure for Validating Topics as Measures. *Political Analysis*, *30*(4), 570–589. https://doi.org/10.1017/pan.2021.33

Zhao, H., Phung, D., Huynh, V., Jin, Y., Du, L., & Buntine, W. (2021). Topic modelling meets deep neural networks: A survey. https://doi.org/10.48550/arXiv.2103.00498

Appendix

Appendix A: Reliability Tests

For intercoder reliability, each author re-coded 25 randomly selected studies, 9 that were not included and 14 that were included in the study. Regarding intracoder reliability, the authors also coded 25 randomly selected studies, that they did not yet code. Reliability was coded with Krippendorff's alpha, however, as some variables were only present in very few instances in the reliability sample, we also added pairwise agreement (in percentages) to illustrate the agreement, when Krippendorffs' alpha is around or below 0.6. We did not include reliability tests for variables v1 (authors), v2 (year of publication), and v3 (publication venue) due to them not being coded but directly adopted from the citation manager. Concerning the variables on the text corpora, we did not test reliability on v6 (name), v7(originality), and v10(genre) as this information was added after the initial coding procedure, by one of the authors. Which topic modeling method was applied in the study (v11) was coded inductively and thus not included in the reliability analysis. The same is true for the validation method (v12), however, as this is the central variable of this study, we added reliability tests, on the category level to ensure the quality of our results. The lowest agreement score is the inter reliability on the validation category of "comparing methods and hyperparamters". To mend this the authors went over all methods in this category one by one and discussed the coding scheme of each of them. The authors found that the disagreement was limited to one validation method "splitting documents" and thus this validation method was recoded for each of the articles.

Table 1

Overivew of Intercoder and Intracoder Reliability

	Intracoder Author 1	Intracoder Author 2	Intercoder
Exclusion	1	1	1
Substantive RQ	0.87	1	0.65 (84.4%)
Methodological focus	0.87	1	0.51(81.3%)
Error Rate Analysis	0.87	1	0.87 (75%)
Qualitative Interpretation (Internal and External)	0.48 (86.6%)	0.00 (93.3%)	0.69
Downstream Tasks	0.77	0.85	0.39 (75%)
Comparing Models	0.87	0.48 (73.3%)	0.09 (62.5%)
Information Theory Metrics	0.64 (93.3%)	1	0.92
Similarity and Distance Measures	0.77	0.64 (93.3%)	0.00 (78.1%)

Appendix B: Overview of Validation Methods

Table 2

Classification of all Validation Methods included in this Study, with the total number of application and studies, in which it is applied

Validation Method	application	studies
Model Comparison	**767**	**486**
Cross-Validation	97	94
Applying different Methods	215	215
Split Train Test Set	307	305
Baseline Model	148	129
Distinctivness of Topwords	**275**	**177**
Coherence Scores	235	170
Exclusivity	26	26
Purity	14	14
Downstream Tasks	**334**	**320**
Error Rate Analysis	**621**	**352**
Accuracy	139	137
Area Under the ROC Curve (AUC-ROC)	30	25
Error 1 and 2	22	12
F-Score	139	135
Mean Absolute Error (MAE)	11	10
Mean Squared Error (MSE)	9	9
Precision and Recall	246	200
Recall-Oriented Understudy for Gisting Evaluation (ROUGE)	11	11
Root Mean Squared Error (RMSE)	14	13
Internal Qualitative Inspection	**637**	**430**
Consulting Topic Experts for Evaluation	36	35
Topic Interpretation	211	187
Reading Top Documents	52	52
Topic Labeling	322	312
Word Intrusion	16	16
External Qualitative Inspection	**210**	**177**
Comparison with inductive corpus coding	85	77
Theoretical Considerations	76	75
Real Life Dynamics / external events	49	48
Information Theory Metrics	**225**	**193**
Entropy	11	11
Jensen-Shannon Divergence (JSD)	26	24
Kullback–Leibler Divergence (KL)	37	36
Perplexity	151	146
Similarity and Distance Metrics	**80**	**66**
Jaccard Coefficient	13	13

Table 2 continued from previous page

Silhouette	15	14
Similarity	52	45

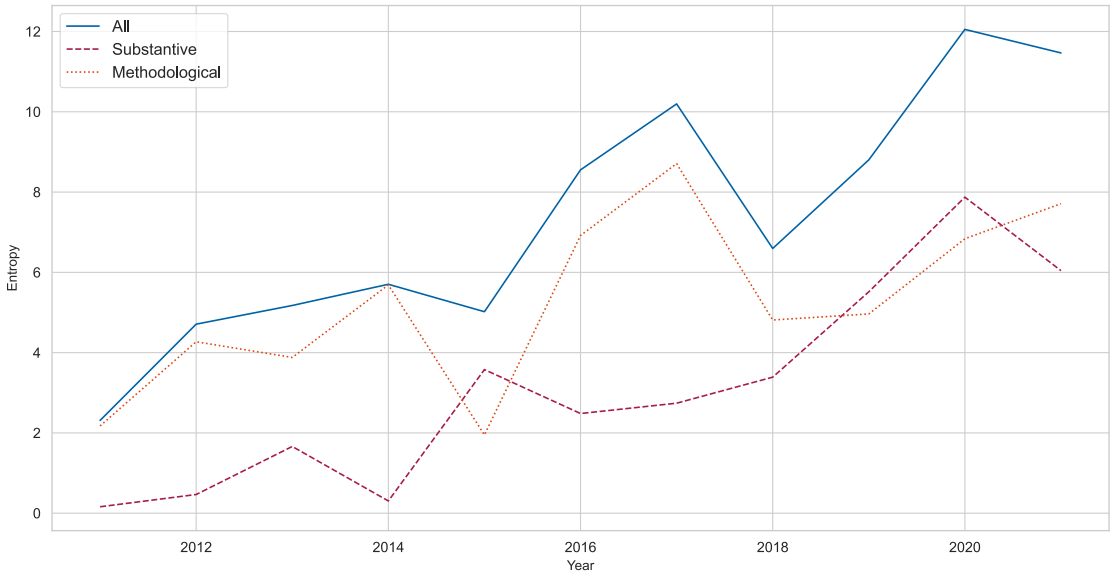
Appendix C: Further Information

Figure 7

Percentage of substantive (left panel) and methodological (right panel) Studies employing validation methods

Figure 8

Information Entropy for validation methods over time



Top Publication Outlets	
ACM International Conference on Information and Knowledge Management	38
International Conference on World Wide Web	34
ACM SIGKDD International Conference on Knowledge Discovery and Data Mining	26
ACM SIGIR Conference on Research and Development in Information Retrieval	22
ACM International Conference on Web Search and Data Mining	18
ACM Transactions on Knowledge Discovery from Data	9
International Journal of Communication	9
Conference on Empirical Methods in Natural Language Processing	9
ACM Transactions on Intelligent Systems and Technology	9
IEEE ACCESS	9

Top Publication Journals	
ACM Transactions on Intelligent Systems and Technology	9
IEEE ACCESS	9
International Journal of Communication	9
ACM Transactions on Knowledge Discovery from Data	9
Journal of Machine Learning Research	8
ACM Transactions on Information Systems	8
IEEE/ACM Transactions on Audio, Speech, and Language Processing	7
Communication Methods & Measures	7
Political Communication	6
Marketing Science	6

Top Publication Conferences	
ACM International Conference on Information and Knowledge Management	38
International Conference on World Wide Web	34
ACM SIGKDD International Conference on Knowledge Discovery and Data Mining	26
ACM SIGIR Conference on Research and Development in Information Retrieval	22
ACM International Conference on Web Search and Data Mining	18
Conference on Empirical Methods in Natural Language Processing	9
ACM Conference on Web Science	7
ACM Conference on Computer Supported Cooperative Work and Social Computing	5
IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining	5
Annual Meeting of the Association for Computational Linguistics	5

Top Publication Outlets Core Social Science	
International Journal of Communication	9
Communication Methods & Measures	7
Political Communication	6
Marketing Science	6
Environmental Communication	4
Journalism Studies	4
Journal of Broadcasting & Electronic Media	3
International Conference on Social Media and Society	2
American Sociological Review	2
International Conference on Digital Government Research	2

Top Publication Outlets Peripheral Social Science	
ACM International Conference on Information and Knowledge Management	38
International Conference on World Wide Web	34
ACM SIGKDD International Conference on Knowledge Discovery and Data Mining	26
ACM SIGIR Conference on Research and Development in Information Retrieval	22
ACM International Conference on Web Search and Data Mining	18
ACM Transactions on Knowledge Discovery from Data	9
ACM Transactions on Intelligent Systems and Technology	9
IEEE ACCESS	9
Conference on Empirical Methods in Natural Language Processing	9
ACM Transactions on Information Systems	8

Table 3

Overview of Top Publication Outlets of Studies in our Systematic Literature Review

Appendix D: Replication Data Set

The replication dataset does not include the source of publications, so that the authors may not be connected to our analysis. However, a list of included studies is given in a second document. Our goal with this paper was not to shame or point fingers towards specific studies and praise others, but to give an overview over which validation methods are being used. The replication file can be used to redo the full descriptive analysis, as well as the visualizations.
https://osf.io/yf47s/?view_only=414af84ab4d54e37910c62ac9a7553c1

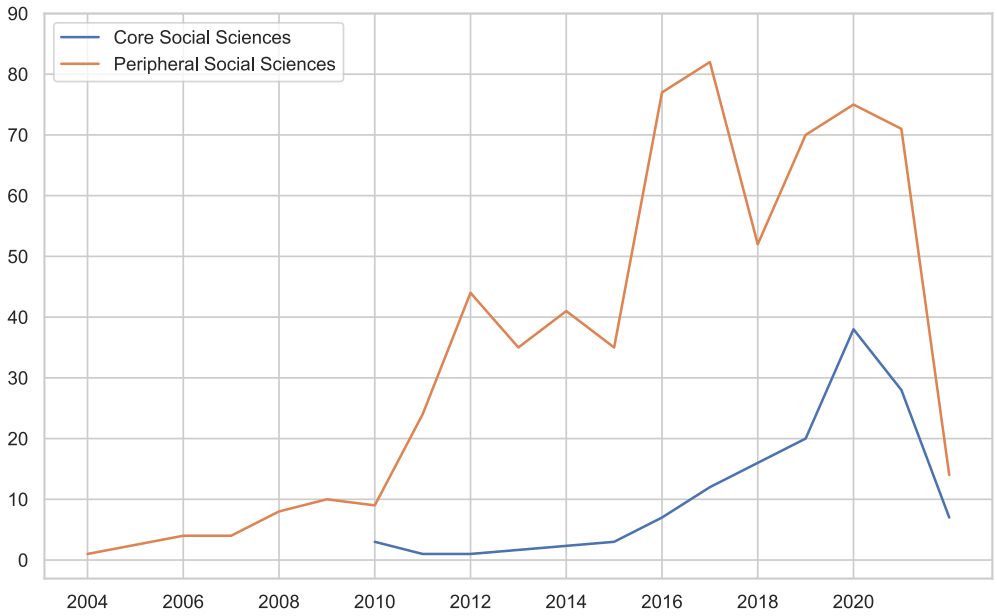
Appendix E: Robustness Analysis

In our literature review on topic modeling validation, we sought to enhance the robustness of our findings by conducting a pooled analysis that differentiates between social science and non-social science studies. Recognizing that methodological and thematic variances might influence the outcomes across diverse research domains, we categorized the publication outlets into these two broad categories. By re-running our entire analysis strategy separately for social science and non-social science studies, we aimed to investigate whether our original conclusions held consistent across different academic disciplines.

This additional analysis serves as a robustness check, ensuring that our findings are not biased by the nature of the publication outlets. The results, presented in the appendix, demonstrate a lack of convergence in topic modeling validation practices across both social science and non-social science studies. This reinforces the validity of our original findings and underscores the widespread challenges in achieving consistent and reliable validation in topic modeling, regardless of the disciplinary context.

Figure 9

Number of Studies from the Core Social Sciences vs. Peripheral Social Science in our sample over time.



Note: As only a quarter of the year 2022 is included in the sample we did not include it in the graph, as it would have resulted in a misleading trend.

Figure 10

Average number of validation categories used per study over time Core Social Sciences vs. Peripheral Social Science.

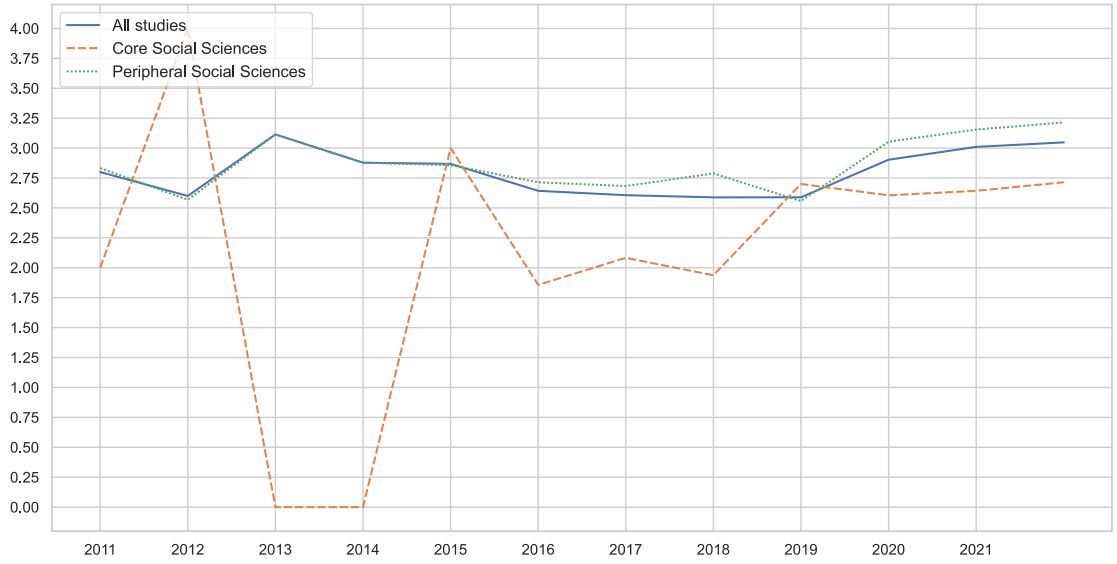
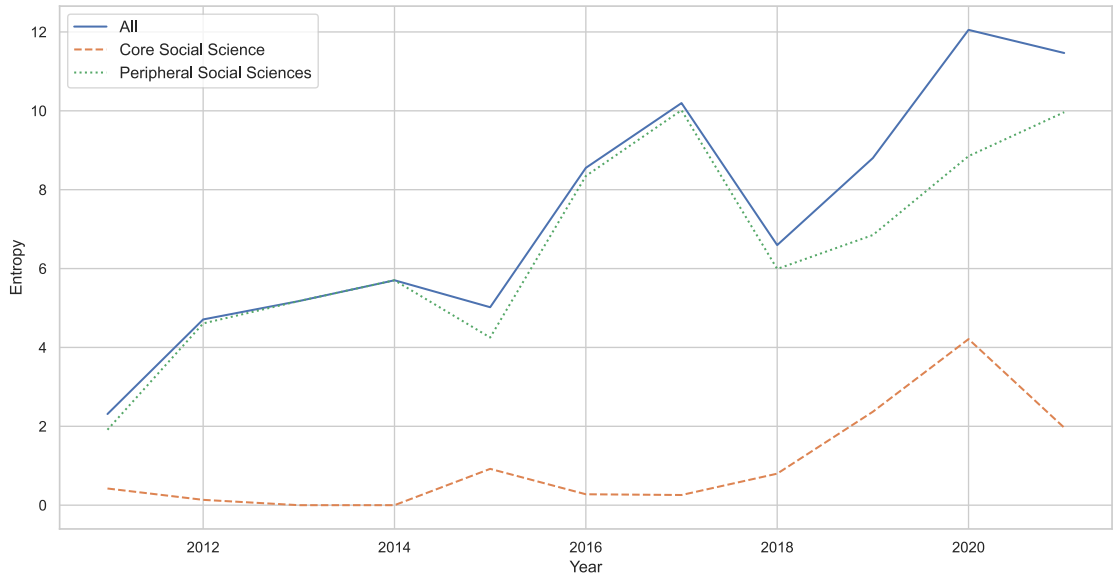


Figure 11

Information Entropy for validation methods over time Core Social Sciences vs. Peripheral Social Science



4.3 Study 3: Topic Model Validation Methods and their Impact on Model Selection and Evaluation

Amsterdam University Press

COMPUTATIONAL COMMUNICATION RESEARCH . (20) 1–26
HTTPS://DOI.ORG/10.5117/CCR????

Topic Model Validation Methods and their Impact on Model Selection and Evaluation

Jana Bernhard
Department of Communication Science, University of Vienna

Martin Teuffenbach
Faculty of Computer Science, University of Vienna

Hajo Boomgaarden
Department of Communication Science, University of Vienna

Abstract
Topic Modeling is currently one of the most widely employed unsupervised text-as-data techniques in the field of communication science. While researchers increasingly recognize the importance of validating topic models and given the prevalence of discussions of inadequate validation practices in the literature, there is limited understanding of the consequences of employing different validation strategies when evaluating topic models. This study applies two different methods for topic modeling to the same text corpus. It uses four validation strategies to assess how the choice of validation method affects the final model selection and evaluation. Our findings indicate that different approaches and methods lead to different model choices and evaluations, which is problematic. This might lead to unwanted results in case the choice of model has a decisive impact on findings and, consequently, on theory development and practical implications.

Keywords: Topic Model, Evaluation, Computational Communication Methods, Validation

Introduction

Topic Modeling (TM) has evolved to become one of the most used computational methods in communication science (Chen et al., 2023). While its versatility allows researchers to apply this methodology to diverse, oftentimes rather descriptive research questions, recent publications have called for computational methods, including topic modeling, to go further into the direction of testing and developing theories (see for example: Bonikowski & Nelson, 2022). Regardless of the goal of the research, a thorough evaluation or validation of the model chosen for the analysis is indispensable (Maier et al., 2018).

Validating computational text-analysis-methods, and especially topic models is not trivial (Grimmer et al., 2022), as the process of applying a topic model leaves a large number of researchers' degree of freedom (Denny & Spirling, 2018; Maier et al., 2018). There is no agreement on what kind of validation steps should be included (e.g. Ying et al., 2022) or how all of these steps are to be reported (Reiss et al., 2022). This lack of standardization (Hoyle et al., 2021) makes the scientific application and interpretation of TM difficult, to say the least.

While these difficulties in validation, or the lack of validation in general, have been discussed in recent literature (Baden et al., 2022), and different prescriptive pieces have been published (most notably Grimmer et al., 2022; Maier et al., 2018), we believe that the consequences as well as the extent of this lacking roadmap to topic modeling are not yet discussed enough in the community. We contribute to "the dialogue about the norms and expectations of using topic modeling and other computational text analysis methods properly at this relatively early stage of adopting the methodology" (Chen et al., 2023, p. 2), by assessing the impact of topic modeling evaluation methods on the subsequent model selection when conducting substantive research. Our aim is to investigate whether, if researcher A decides to employ a given validation method, they would run a different TM specification than researcher B, who relies on a different validation method. Furthermore, we consider whether such differences are different for different TM algorithms. Thus, we showcase how a researcher's choice of a particular validation method over another one can, instead of lending credibility to their results, severely influence and potentially bias the results. Our contribution calls for more careful reflection on how validation methods may lead researchers to consider different TM specifications and hence for the dependency of TM approaches on what validations researchers prefer to employ. In addition, we present a four-step recommendation plan in this

later sections of this paper, offering guidance to researchers on planning their model selection effectively.

Theory and Related Work

While different topic modeling methods have different underlying assumptions, approaches, and needs, all these techniques employ machine learning to extract previously unknown patterns in large text corpora, which are then interpreted by researchers as topics (Boyd-Graber et al., 2017). Studies on topic modeling differentiate between four steps in the process of applying a topic model: first, the pre-processing of text data, second, choosing hyperparameters, third, model selection, and fourth, model validation (Chen et al., 2023; Grimmer et al., 2022; Maier et al., 2018). Steps three and four are somewhat intertwined in praxis, as the selection of one specific model over other alternative models is often based on the same validation methods that are used to validate the final model. Thus the evaluation of multiple, possible models is – or at least can be – done in the same way as the validation of the final model used to address substantive research questions. All preprocessing and hyperparameter choices as well as model selection introduce on the one hand complexity and researcher degrees of freedom, and on the other hand potentially have an impact on the results (for pre-processing and hyperparameter setting see Denny and Spirling, (2018), Maier et al., (2018) and Tolochko et al., (2022) and for model selection see Grimmer et al., 2022).

We argue that, given the multitude of choices and associated researcher degrees of freedom, it is vital in topic modeling to rely on different validation approaches to come to an informed model selection. In such a scenario we would argue that actual validation work is done in step three, which then would yield a choice of the most valid model to be selected, whereas step four then merely evaluates the overall quality of the validation outcomes against an ideal case. Hence, this study primarily focuses on step three - we assess how different validation approaches may or may not lead to different conclusions which model to select eventually. Thus, the decision on selecting which topic model is used for possible substantive analysis is often based on assessing which model looks useful ("face validity") or which models get better scores at various statistical measures ("statistical validity"). Yet we have little systematic knowledge as to whether and how different validation approaches would converge towards the same model selection.

86 While these scholars give us some indicators of what to focus on when

discussing the validity of the results of applying TM, the validation of the models themselves is inherently challenging due to the characteristics and properties of the method as well as its usage. Methodologically speaking, topic models are applied to find useful text classifications based on the topicality or themes of each document. However, what is useful depends on the research problem in question and is strongly dependent on what the model is used for. There are a number of different topic modeling methods, which propose different functions to cluster text. These objective functions are formulated to identify an optimal partitioning, which is determined by a predefined similarity metric, such as the cosine similarity between sentence embeddings. Whether or not this is useful is for the researchers to decide, thus separating the mathematical, formalized "optimal" model from a "model that can answer my research question". This discrepancy between what is mathematically optimal and what is optimal for research introduces additional complexity (and degrees of freedom) to the social scientist since "model fit" essentially depends on an ad-hoc decision and should be thoroughly investigated and justified, which further complicates validation efforts. While these ex-ante decisions are important (Chen et al., 2023; Gentzkow et al., 2019), other researchers have emphasized the importance of post-hoc tests to ensure validity. The application of topic models to a diverse range of text corpora and research questions requires an individual approach to validation, given the specificity of each case (Barberá et al., 2021).

Ballester and Penner (2022, p. 2) argue that "the three properties that functional topic models should have [are]: robustness, descriptive power and reflection of reality." Validation relates to the latter property. Validity in social science refers to the accuracy and truthfulness of the results and conclusions of a study. It's the extent to which a study measures what it claims to measure and that the results are a true reflection of the reality being studied. Social scientists differentiate between types of validity that can be taken into account. In general, Scharrer and Ramasubramanian (2021, p. 62f) explain *face validity* ("the measure maps on to common understanding of the concept"), *criterion-related validity* ("the measurement relates in a logical manner with another variable outside of your study") and *content validity* ("degree to which the full range of meanings of the concept are being reflected in the measurement"). On manual content analysis Krippendorff (2013, p. 319) differentiates three main categories *face validity* ("being obviously true, sensible, plausible"), *social validity* ("addressing important social issues"), and *empirical validity* ("The degree to which available evidence

and established theory support intermediate stages of a research process and its stages"). Regarding the latter, he concludes that this evidence can be based on *content, internal structure* and *relations to other variables*. He further distinguishes these three subcategories to include *sampling, semantic, structural, functional, correlative* and *predictive validity*. This detailed description of different types of validity can function as a guide when thinking about how we validate automated content analysis, such as TM.

The validation of topic models is critical in scenarios in which ground truth labels are not available for the text corpus being analyzed (as arguably true for most TM application scenarios). DiMaggio and colleagues (2013, p. 586), partly relying on Grimmer and colleagues (2011) emphasize that validation should focus on three different points of view:

1. statistical validation: if the model results are consistent with the assumptions of the model
2. semantic or internal validation: whether the model meaningfully discriminates between different senses of the same or similar terms
3. predictive or external validation: attention to particular topics responds in predictable ways to news events

While the first, statistical validation, has a special place due to the mathematical background of TM, the second validation step is closely related to what has been described in general as *criterion-related validity* as well as *content validity* and Krippendorff subsumed in the *internal structure*. The third then connects well to Krippendorffs *relations to other variables*.

Probably the clearest roadmap for TM in communication science was put forward by Maier and colleagues (2018). They describe the following steps in evaluating TM to ensure reliability and validity: 1. Coherence Metrics to identify useful hyperparameter settings and 2. Qualitative judgement of different, but well-performing models (as found in step 1) by experts based on the top words. This leads to the selection of one topic model, which is validated in more depth, by summarizing different statistical values, excluding topics that are not interpretable, reading documents that are related to each topic, and employing hierarchical clustering on the top words, to identify mergeable topics.

The selected topic model or models are then validated in more depth, by summarizing different statistical values, excluding topics that are not interpretable, reading documents that are related to each topic, and employing hierarchical clustering on the top words, to identify mergeable topics. Thus, validation methods are applied in two steps: Model Selection and Model

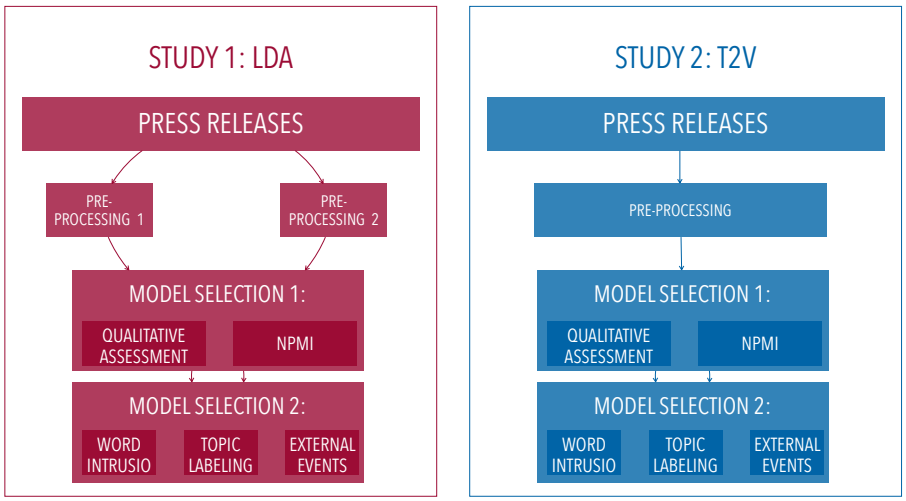
Validation. As it is quite necessary for the research process to decide on one (or at least only very few) topic models to base the substantive results on, this paper aims to showcase that this is often not as straightforward as some scholars have described before. It might be a trivial statement to suggest that each researcher has some influence on the results, however, they should at least aim for as little impact as possible or relatedly for objectivity and transparency throughout all decisions in the research process (Scharrer & Ramasubramanian, 2021). Our goal is to explain how the choice of validation methods can impact the model selection and thus the final results of a research study, that employs TM.

Research Design

We propose an empirical setup to assess how the choice of validation method impacts the model evaluation and selection, and thus, potentially the results of topic models for substantive research questions. Our design, as illustrated in Figure 1, relies on two studies with distinct TM approaches, the results of which in combination would yield insights into our research interest. We apply two topic modeling algorithms, with different pre-processing steps, to one text corpus and then apply different evaluation methods and assess how they would affect model selection. In our first study, we test the impact of validation methods on models generated with LDA, as LDAs are still among the most used methods in the social sciences (Chen et al., 2023; Maier et al., 2018). Second, we use Top2Vec (T2V) (Angelov, 2020), which is an embedding-based model built on a pre-trained neural language model.

As mentioned above, there are many different validation methods and no standards regarding their implementation, not to speak about their potential combination. To choose appropriate approaches for our study, in the first step we looked at prescriptive resources: Maier and colleagues (2018) as well as Ying and colleagues (2022), emphasize the technique of labeling topics by reading the top words or related documents and by relying on automated metrics (topic coherence, mutual information, or hierarchical clustering) regarding internal validity. On external validity, they suggest expert evaluation, manual codings as well as considering external events. Additionally, Ying et al. (2022) refined the intruder method, which was first put forward by Chang and colleagues (2009) to measure semantic coherence in an attempt to create one off-the-shelf validation method that can be used for any topic modeling research question. Grimmer and colleagues (2022) regarding validation without gold-standard data, highlight practices such as

Figure 1: Visualization of the Research Design



intrusion tasks (Chang et al., 2009; Ying et al., 2022), labeling top words for semantic validity, and assessing the correspondence of the model to external events (hypothesis validity). A recent systematic literature review found that the most frequently used validation methods that build on human judgment are: labeling topics based on top words, and human interpretation of topics based on top words and documents, comparing to manual analysis, including theoretical considerations and relating to external events (Bernhard et al., 2022).

Evaluation Methods

In line with the research presented above (DiMaggio et al., 2013; Maier et al., 2018) we chose different points of view of validity: statistical (NPMI) as well as face validity (qualitative judgment of top words and reading documents) to choose useful hyperparameters as well as internal (word intrusion and topic modeling) and external validation (relation to external events) to further evaluate our models. As the steps on internal and external validation are quite labor intensive, we used the information from statistical and face validity to choose two models which were evaluated in more detail. Of course, these different validation approaches provide different perspectives on model validity, yet as a baseline, we would argue that ideally, TM should show high validity on all accounts. If TM is used for theory building and refinement, it appears important to draw on models that do not compromise on certain types of validity but rather converge on high validity in different areas and through different approaches.

Mutual Information

As a first step, adhering to the advice of Maier and colleagues (2018), we took a statistical indicator to determine which hyperparameter settings would lead to the most "useful" topic models. We initially computed various coherence metrics (Röder et al., 2015) for all models to systematically evaluate the hyperparameter settings. Adhering to the recommendation of Hoyle and colleagues (2021) as well as Grimmer and colleagues (2022), we ultimately relied on the Normalized Pointwise Mutual Information (NPMI) metric as well as human judgment (qualitatively checking top words and documents) for deciding which are useful models. The NPMI score is high if the top N words have a high joint co-occurrence probability, i.e. the words often co-occur in the corpus. This is an intuition similar to what most statistical topic models (e.g. LDA) make use of, where topics are generated based on word co-occurrence patterns. Neural-topic models on the other hand rely on text representations generated by neural-network-based models (e.g. transformers). These embedding models are optimized to find semantically meaningful representations of texts. Therefore, we expect statistical models to perform better when compared to neural models in terms of automated topic coherence metrics. Thus, we do not compare the NPMI scores between topic modeling methods but only within one method.

Word Intrusion

We first implement a word intrusion task, as put forward by Chang and colleagues (2009). This evaluation method is extremely versatile and straightforward. The method uses the top words that are calculated to be indicative of each topic and postulates that a human should be able to spot a randomly included word, that is not part of these top words. Thus, it is a test of internal validity (as defined by DiMaggio et al., 2013, or a matter of face or semantic validity as defined by Krippendorff, 2013). We took nine top words from each topic and randomly included one of the top ten words from another topic as the intruder. We instructed three student assistants who were not familiar with the details of the research project but were aware that they were evaluating press releases, to mark the intruder word. Each of them completed this task in two days. We then calculate the percentage of correctly identified intruders, thus, this measure can go from 0% to 100%, allowing us to compare models with a different number of topics.

In the topic modeling process, the LDA model assigns a topic probability to each word in the corpus. For the generation of topic top words, we selected

the top n-words for each topic, representing the words with the highest topic probability. Due to the LDA model's statistical nature, these top words are in general words that often co-occur in the text corpus. T2V aligns words and documents within a shared latent vector space. The algorithm identifies document clusters within this space, defining them as topics. For selecting top words, we extracted the top n-words from this latent space with the highest similarity to documents within the topics. In other words, we selected words that the model represents as semantically similar to the document clusters. These are words that are used in the same context, which is in general not equivalent to the LDAs word co-occurrence approach. We, thus, expect the T2Vs approach to produce superior results for this evaluation, as word co-occurrence often finds words that are not related to the topics.

Topic Labeling

To include further human oversight (Grimmer et al., 2022), we read 10 documents per topic to assess whether they can be meaningfully interpreted (as suggested by Maier et al., 2018). Meaningful in this case is defined as the documents relate to one, distinct issue of Austrian Politics. This was done by one of the authors with an education in political and communication science so that topical expertise is given. To do so, the topic of each document was paraphrased with one or two words before trying to find one label for the topic encompassing all of the ten documents. To compare the number of meaningful topics, we additional distinguished between three cateories: no label found; label found that would relate to all documents, labels found if the texts in the topic included more than one topic. For example, a topic, with documents on health policy and voluntary work. Thus, in the classification of DiMaggio and colleagues (2013) this task points us to internal validity. For Krippendorff (2013) this task would be in the area of face and content validity. Yet, LDA allows for documents to have multiple topics, while T2V classifies each document into one topic. Thus, to get to the documents of each topic, we only chose documents for which the topic made up more than 50% of each press release. However, we do not expect this difference to substantively impact the evaluation method, thus this kind of evaluation can be used within and between the different topic modeling methods.

External Events

The last method, comparing to external trends (Maier et al., 2018; Ying et al., 2022) aims at comparing the findings of a topic model (e.g. the number of topics in a given timeframe) to some kind of external baseline (e.g. official statistics or the occurrence of specific events). Thus, this evaluation method would be classified as external (DiMaggio et al., 2013) or correlative (Krippendorff, 2013). Often this method is only partly implemented, as it is only possible for topics that can be reasonably expected to be related to quantitatively measurable external events. This can either be done for topics that are of specific interest for the analysis or as many topics as feasible. As an *example*, we show how the topic of unemployment develops over time and compare this development to official unemployment statistics (WKO, 2022). We then calculate a correlation index (Person's r) to assess how close the two developments are. We argue that this is a reasonable comparison, as it can be expected that parties talk more about unemployment when it is high, as this also leads to unemployment being discussed in the news. However, as it could also be that unemployment is discussed more when it is exceptionally low, we do not expect a strong correlation. We thus argue that the differences in the correlation should be focused on, not the strength of the correlation itself. The decision of which topics to compare to which statistic has to be taken, in part, after the topic model has been evaluated as to which topics it includes. For this study, we wanted to be able to compare all four models based on the same topic-statistic correlation. We chose unemployment, over the other connecting topics Health (too ambiguous in the LDA models), Pension (lack of external event), and Feminism (lack of suitable external statistics). Regarding the different TM methods we do not have a strong reason to expect this validation approach to work better or worse for one or the other method.

These four validation methods correspond to different kinds of validity, as described above. While mutual information relates to statistical validation, the intention behind the metric can be seen as relating to internal validity as well. This connects it to the task of word intrusion, and topic labeling, which in the classification of DiMaggio (2013) all relate to internal validity, while the comparison to external events, would be external validity. If we take into account the more detailed description of Krippendorff, we can see some differences between the three internal validation methods, as they could relate more to semantic (word intrusion) or content (topic labeling), however, of course, both still go into the direction of internal validity. In sum,

we would expect the results of approaches one to three (mutual information, word intrusion, and topic labeling) to converge more and clearly show which model a researcher should prefer since they arguably relate to the same types of validity. Approach four (external events) may, somewhat in contrast, diverge more from the pattern, as it aims at measuring a different kind of validity. Yet ideally models are valid on all accounts.

Case Description

As a case for this setup we analyze which topics parties in Austria have talked about in the past 15 years. To do so we aim to find the most useful text classification put forward by the method of topic modeling. We define "usefulness" as the number of topics that can a) be meaningfully interpreted by humans and b) are theoretically sensible for the context of Austrian politics between 2004 and 2020. For this analysis, we use 218.471 press releases that have been sent out by the five parties currently in the Austrian Parliament (SPOE, OEVP, FPOE, GRUE, NEOS). [1]

Topic Modeling Methods

Study 1: Latent Dirichlet Allocation(LDA)

The Latent Dirichlet Allocation (LDA) model (Blei et al., 2003) is one of the most widely used topic modeling methods (Bernhard et al., 2022). It is a statistical model that simultaneously estimates a document-topic and a topic-word distribution. With those two functions, one can estimate the membership probability for each document to the topics, as well as the most descriptive words for each topic. The classical LDA model requires the number of topics k to be specified beforehand. The resulting distributions can be adjusted with two parameters, usually denoted as α and β. The parameter α is the prior concentration parameter representing document-topic density. Hence this parameter controls how many topics are assumed to be in a document. High α results in more topics per document. β represents the topic-word density prior, which influences how many words are ascribed to each topic. As with most statistical models, LDA requires pre-processing of the data. Pre-processing has a strong influence on the results (Denny & Spirling, 2018). Therefore we decided to follow best practice conventions (Maier et al., 2018) and performed the following steps:

[1] Replication material for this study can be found on https://doi.org/10.17605/OSF.IO/PYFDT.

1. Removal of punctuation and digits, lowercase all characters
2. Stemming and tokenization
3. Remove the most frequent and least frequent words.

We performed the pre-processing with two settings, once with removing all words that appear in $\geq 95\%$ or $\leq 0.5\%$ of the documents and once with $90\%/1\%$. We evaluated several values of k (4 to 50) and α (0.1 to 1). β was set to $\frac{1}{k}$ (*symmetric prior*, for more information on the parametrization of the LDA model, see Maier et al. (2018)). For every parameter-setting, we performed three runs and averaged the topic coherence score.

For all settings and parametrizations, we validated the models with the Normalized pointwise mutual information (NPMI). This score returns a value between 0 and 1, the higher the better. The upper Figure 2 depicts the achieved results for our two settings. As expected, the LDA model produced topics with nearly perfect NPMI scores (over 0.96 for all parameter settings for all $k \geq 10$). Additionally, we found that this coherence score (1) improved with the number of topics and (2) hardly varied for different parametrizations (less than 0.025 for $k \geq 10$).

Due to the high time consumption of manual validation, we decided to pick only two models for further analysis, one for each pre-processing setting. We chose models with different k so that we get an overview of how k impacts the results (6, 14, 30, 40, 50). We manually inspected ten top words as well as five documents related to each topic. We then chose two models which have the most interpretable topics for further human-based validation.

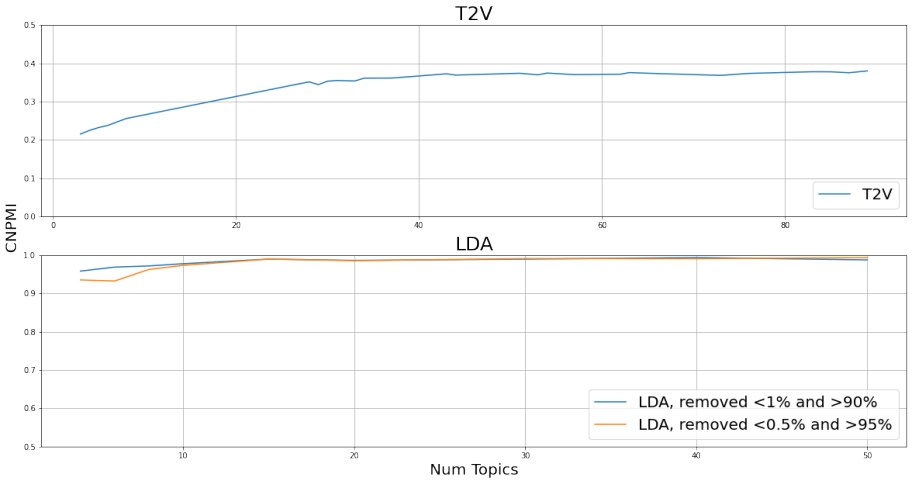
Study 2: Top2Vec (T2V)

The T2V (Angelov, 2020) model is a neural-network-based topic model. In contrast to statistical models, neural language models utilize context-aware embeddings instead of word frequencies. Therefore, these models do not require extensive pre-processing of the input texts. To find topics, T2V embeds the input corpus with a pre-trained embedding model and clusters them. The resulting clusters are interpreted as topics. Next, the vocabulary of the corpus is embedded in the same vector space. For each cluster of documents (i.e. topic), the closest word embeddings based on Euclidean distance in the embedded space are computed and used as topic representatives.

T2V utilizes a density-based clustering algorithm, namely HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise)

(Campello et al., 2013), combined with an additional dimensionality reduction algorithm. The results of T2V can be adjusted with the initialization of the HDBSCAN algorithm. Note that, unlike LDA, this algorithm does not use a pre-defined number of clusters (i.e. topics) k. The framework supports a variety of embedding models. For our experiments, we decided to use an SBERT (Reimers & Gurevych, 2019) model that is pre-trained on a multilingual-text corpus (*distiluse-base-multilingual-cased*), which is a state-of-the-art transformer model. As T2V requires no further pre-processing, to achieve various numbers of output topics k we adjusted the min-cluster-size parameter of the HDBSCAN algorithm (see Campello et al., 2013). After several runs, we again evaluated the NPMI metric (see Figure 2). Similar to the LDA model, the metric increased with the number of topics. However, the results were significantly worse than for the previous model (between 0.22 and 0.38 compared to 0.91 to 0.99). Again, we picked two models for further validation. To do so, we again manually assessed the quality of models with different k (4, 19, 30, 54, 63).

Figure 2: NPMI coherence scores for LDA and T2V



For a summary of the models parameterization and the preprocessing of textual data please refer to Tables 3 and 4 in the Appendix.

Results: Study 1: LDA

Word Intrusion

Three student assistants completed the word intrusion task for both LDA models (see Table 1 for detailed results). We found that the LDA40 TM per-

formed on average a bit better (one-quarter of intruder words correctly identified), but the detailed results of the student assistants differ from each other, which suggests that this estimate is unstable. The LDA50 TM performed worse (one-fifth of intruder words were correctly identified). However, both scores are not good overall, which would suggest that both LDA models are not sufficiently well suited to be used for substantive research.

Table 1: Results of Word Intrusion Task for LDA Models

	TA1	TA2	TA3	mean
LDA40	17.5% (7/40)	32.5% (13/40)	22.5% (9/40)	24.2%
LDA50	18.0% (9/50)	20.0% (10/50)	16.0% (8/50)	18.0%

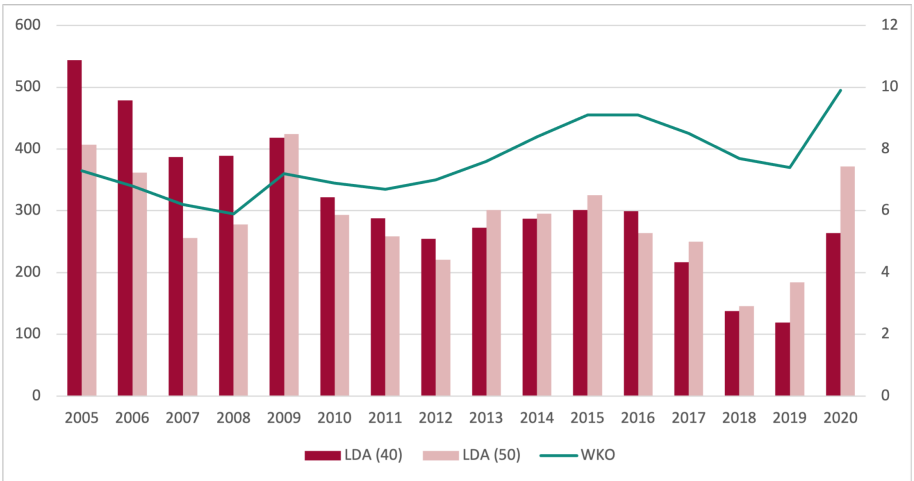
Reading Documents

We found that in the first LDA model (40 topics), 13 topics (32.5%) revolved around a meaningfully interpretable topic. Additionally, 16 topics (40%) could be interpreted, even though they included two topics that were connected but not the same. Only 11 topics (27.5%) could not be interpreted at all. Similarly, the second LDA model (50 topics) included 15 meaningful topics (30%), however, only 13 topics confounded two connected topics (26%). This led to 22 topics (44%) that could not be interpreted. This validation method would suggest that the first LDA model is more suitable for substantive analysis, yet still with about a third of the models representing nonsense.

External Events

Figure 3 shows the development of the *Unemployment*-topic for both models, as well as the official monthly unemployment statistics. We see that although there is some parallel movement in the development of unemployment salience in press releases and the statistics, they do not correlate strongly (LDA40: $r(14) = -0.35, p = .181$ and LDA50: $r(14) = 0.13, p = .63$). More worryingly, however, is that one of these correlations is positive, while the other is negative, however, neither of the correlations is significant. This would suggest that neither of the models adequately captures the topic of *Unemployment*. 97

Figure 3: Number of press releases on the topic of Unemployment in both LDA models versus the official unemployment statistic for Austria



Results: Study 2: T2V

Word Intrusion

Three student assistants completed the word intrusion task for both T2V models (see Table 2 for detailed results). For the top words of the T2V30 TM, 84% of intruders were found consistently by the student assistants, while 75% of intruders were found for the T2V30 TM. Both scores are indicative of models that have coherently clustered documents into topics, but it is clear that, based on these results, the TM with 30 topics would be preferred for further substantive research.

Table 2: Results of Word Intrusion Task for T2V Models

	TA1	TA2	TA3	mean
T2V30	83.3% (25/30)	86.6% (26/30)	83.3% (25/30)	84.4%
T2V63	74.6% (47/63)	76.2% (48/63)	74.6% (47/50)	75.1%

Reading Documents

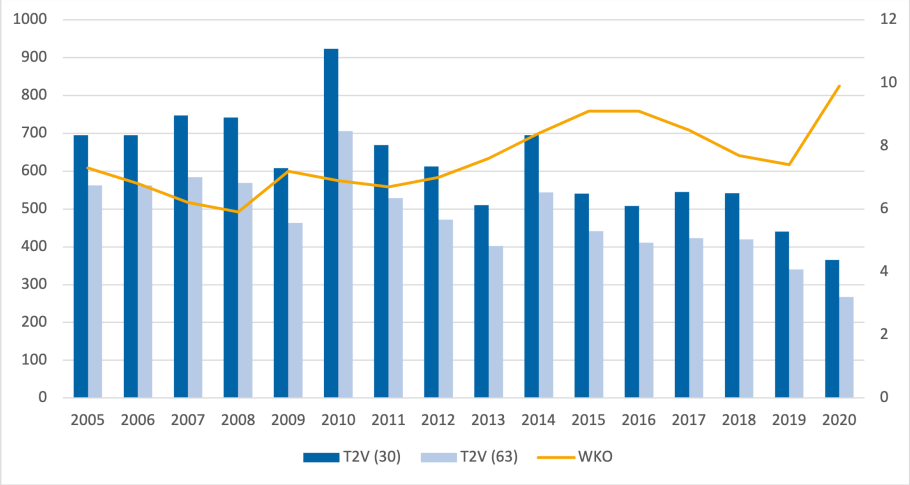
Regarding the first T2V Model (30 topics) we found 26 useful and meaningful topics (86.7%) and only two which confounded two topics, as well as two that could not be interpreted (6.7% each). The second T2V Model (63 topics) included 57 meaningful and useful topics (90.5%) and only three topics that

confounded two topics and another three that could not be interpreted (4.8% each). Thus, both models seem to cluster the press releases in the most meaningful ways, which would allow for further substantive research.

External Events

Again, we want to see how the number of press releases in the *Unemployment*-Topics relates to the official statistics. Figure 4 shows some parallel development, and this time we see a stronger correlation (T2V30: $r(14) = 0.68, p = 0.04$ and T2V63: $r(14) = 0.67, p = .005$). Thus, both models seem to adequately capture this topic.

Figure 4: Number of press releases on the topic of Unemployment in both T2V models versus the official unemployment statistic for Austria



Final Thoughts on Model Evaluation

These validation steps reveal different things about the different topic models. First, on LDA: Our results show that both models score similarly on NPMI, the word intrusion task, and the comparison to external events. However, had we relied on statistical validity only and not taken into account any human validation approach, we would have been confident with our models and would have used them for substantive research. Had we relied on the word intrusion task or the comparison to the Unemployment Statistics we might have concluded that both LDA models are insufficient for further substantive research. The two LDA models seemed to be successful when looking at the topic labeling method, where the LDA40 model performed

slightly better. Thus, depending on which validation method we would put our trust in, we would have come to different conclusions about our models. This has a clear impact on possible substantial results, as the two models give us vastly different topics (see Table 5 in the appendix). Thus, even though the word intrusion task and the topic labeling both aim at face or internal validation they point us in different directions.

Second, on T2V: Our results show that the models get similar results for NPMI, reading documents, and external events. Both models score very high in the validation task that is based on reading documents, which will lead us to believe that we have great models that can be used for substantive research if we were to rely on this approach only. Our models do show differences in their performance of word intrusions and how they compare to external events. This suggests that as researchers we would have to make a trade-off between internal and external validity. Both models seem to work similarly well when compared to unemployment statistics, suggesting that they might be used to some extent, for substantive research on this topic. T2V30 performs better on the word intrusion task, which could sway researchers to choose this model when only considering this validation method. Thus, again, depending on which validation method we choose, we would either assume that both models are equally good, or that the T2V30 is slightly better. For this method, the impact on results is smaller, as the topics are more stable (see Table 5 in the Appendix and Rodriguez and Spirling, 2022).

For both TM methods, we thus see a divergence in terms of how different validation approaches may lead to different conclusions about the substantial usefulness of a particular model. If, as argued above, an ideal scenario would show the strong validity of a model in all accounts, this is something that we do not clearly see in any of the scenarios above. So, in the absence of a standardized approach to topic model validation (which validation approaches to apply, how many of them) our results demonstrate a situation in which different validations, were they used exclusively, would point researchers to use different models for substantive research. This problem, however, appears to be less strong for T2V, since here we see a stronger convergence of different approaches. Thus it seems that in our scenario, the T2V method showed more robust models than the still widely used LDA.

100

Discussion

Validity in the social sciences refers to the accuracy and truthfulness of the results and conclusions of a study and is often defined as the extent to which a study measures what it claims to measure. Especially when talking about computational methods, which utilize algorithms that work as a black box or are applied by researchers without a background in computational science, validation is often performed post-hoc on the models' results. The consequence is that each validation strategy depends on the research question, text corpus, and (maybe) the theory behind the analysis. This setup is less than ideal and the reliance on models to interpret results in light of theory has been named as one of the causes of the replication crisis in psychology (see for example: Wiggins & Christopherson, 2019).

In more straightforward statistical models (like regression models), certain criteria evaluate how well the model fits the data and if these criteria are met — conditional on theoretical expectations — one can be sure that the output is correct. Topic models, however, do not have a method that defines the "correctness" of the model. Regardless of the post-hoc model validation, there are several "useful" models (as demonstrated above). But this means that none of these models can be shown to adhere to a single theory. Ultimately the choice of the model would determine which "theory" we are testing (without our explicit knowledge). Thus, every judgment is dependent on a, more or less, arbitrary model selection, and is therefore post-hoc and not suited for theory building. In a quantitative setting, even if we build on gold-standard data (see e.g., Song et al., 2020) and have a good model fit, researchers have to rely on existing theories for interpreting the results. As argued earlier, in the absence of clear guidelines, topic modeling is not yet a standardized methodology (but first steps are provided by Denny and Sperling (2018) and Maier et al. (2018)).

Topic Model Selection is a crucial step in the topic modeling process, which is often brushed over or presented as being very straightforward (taking the model with the best scores regarding different statistical values). However, as evident in this study, it is not that easy. Above we have shown that the application of different validation methods exclusively would lead researchers to put their faith in different models that at points show vastly different substantial results. Our study also showed that this is more problematic for LDA models, as compared to T2V models. It appears that T2V would show a somewhat better convergence of different validation approaches and therefore might be the preferred modeling method to yield TMs with higher overall validity. Where do we go from here? Planning the

validation of the topic model should start before the application of the topic model. Here are some steps to take when using topic models to research communication scientific phenomena.

Where do we go from here? Planning the validation of the topic model should start before the application of the topic model. Here are some steps to take when using topic models to research communication scientific phenomena.

1. Before starting the application of the method, in the first step researchers should consider several questions that may inform validation steps: What would a good model look like? Although this might seem like a somewhat obvious suggestion at first glance, it is not trivial at all to formulate a short description of a) what topics a good model would include or not include, b) how many topics you would expect at a minimum and maximum, or c) which patterns would you expect to find. These decisions have to come from knowledge of the text corpus, text context, and theoretical considerations. For deductive research, this process is close to hypothesis generation, however, not about the model outcome, but the model itself. This description should be saved, so that it can be used for the upcoming steps.
2. A second step researchers could ask themselves: If these data were created by manual content analysis, that you did not conduct yourself (for example in secondary data analysis), how would you go about checking the quality of the data and the validity of the topic classification? Content Analysis has been applied in communication science for a long time, and as a scientific community, we have found ways of thinking about the validity (Krippendorff, 2013). We can and should use this knowledge to build validation strategies for automated analysis, and topic models. We thus suggest using the validation classifications we already have from manual content analysis (Krippendorff, 2013; Scharrer & Ramasubramanian, 2021), or from prescriptive publications such as (DiMaggio et al., 2013) and looking at different kinds of validation and how they relate to the description made at the first step. Which validation angle is helpful to gain insights into the description we formulated? We suggest deliberately taking into account as many different kinds of validation as possible so that it can be assessed whether a model is good on more than one account. The goal of this second step is to come up with a list of kinds of validations that should be applied to all potential models.
3. Third, researchers have to decide which evaluation method they want

to apply. For this, researchers should use the overview of which validation methods correspond with which validation angle for step two. This gives them a rich list of possible validation steps to take. The researchers can then shorten this list by assessing which methods are feasible (in terms of e.g. time or funding), have been applied by researchers with comparable projects (e.g. Maier et al., 2018) or proposed for a specific topic modeling method (e.g. Zhao et al., 2021). We want to highlight, however, as many have before us, the importance of including human-in-the-loop validation methods.

4. Researchers then have to decide on several validation methods, which they want to apply to their model at the a) model selection and b) model validation stage, to avoid arguing circularly. Additionally, researchers should be transparent about why they chose specific methods and disregarded others, and clear about which benchmarks they set for which validation method so that a model can either pass or fail a specific step in the validation process. At this stage, researchers also need to take into account the possibility, of different methods not converging, and pointing at different models, as could be seen in our example. In this case, researchers need to decide which validation method to prioritize.

When following all the steps in the list, researchers end up with a description of what a good model would look like, which kinds of validation correspond to this description, and which validation methods can be applied to assess these kinds of validation. The researcher also has a set of decisions that were taken for or against a method, as well as benchmarks for them. This can increase the transparency of the decisions taken by the researcher. We recognize that this process is extremely resource-intensive. However, it is important to recognize the impact validation strategies have on model selection when discussing findings that were obtained through topic modeling. As discussed above, we believe it is vital to come to better-informed model selections through the application of different validation approaches in step three, selecting the best-performing alternative. The same validations can then be used to judge, in step four, to what degree the best-performing model can actually be considered a valid representation of the text.

Our study is not without limitations. The first is, that we cannot solve the problem we describe, only give recommendations and demonstrate its implications. Second, we rely on only one text corpus in one language in our demonstration. We thus want to encourage further research in this area,

including different corpora and languages. Third, we showcase the impact of three widely used validation methods, however, there are many more, which were not included (e.g. Bernhard et al., 2022). Fourth, we also had to rely on a pre-selection of topic models, which is based on statistical and face validity, to reduce the number of models that we assessed in-depth.

We see this as yet another indication that we need to shift our attention toward measurement validity (Baden et al., 2022) before we can talk about generating new theories with topic modeling. Indeed, recently, scholars have highlighted how topic models can be used in a qualitative research setting (Isoaho et al., 2021). This allows researchers to put the unsupervised and inductive nature of the method to use. The validation of computational methods, and all methods in general, is an important step in the research process. Continuing our efforts in researching and revising the process of validating is needed if we want to use computational methods to build communication scientific theory.

Conclusion

In conclusion, the validation of topic model selection in communication science research is a crucial step to ensure the accuracy and reliability of study results and any theoretical or practical recommendations derived from them. Computational methods, such as topic modeling, present unique challenges due to their algorithmic nature and reliance on post-hoc validation strategies. We showcased that the choice of validation method has an impact on the selection of the final topic model, which in turn impacts the results. Thus, we argue that topic models offer valuable insights and facilitate exploratory analyses, but their use for theory-building remains problematic. To add to the literature on the validation of computational methods (Baden et al., 2022; Chen et al., 2023), we have proposed an approach to coming up with a validation strategy for topic model selection, emphasizing the importance of formulating a clear description of an ideal model and aligning validation strategies with existing content analysis methodologies. By transparently documenting the decision-making process and benchmarks, researchers can enhance the credibility and replicability of their findings. Additionally, a shift towards measurement validity is essential before topic modeling can become a reliable tool for theory generation. As we continue to explore computational methods' potential, refining and standardizing validation processes will be paramount in advancing communication scientific theory.

References

Angelov, D. (2020, August). Top2Vec: Distributed Representations of Topics. https://doi.org/10.48550/arXiv.2008.09470

Baden, C., Pipal, C., Schoonvelde, M., & van der Velden, M. A. G. (2022). Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda. *Communication Methods and Measures*, *16*(1), 1–18. https://doi.org/https://doi.org/10.1080/19312458.2021.2015574

Ballester, O., & Penner, O. (2022). Robustness, replicability and scalability in topic modelling. *Journal of Informetrics*, *16*(1), 101224. https://doi.org/https://doi.org/10.1016/j.joi.2021.101224

Barberá, P., Boydstun, A. E., Linn, S., McMahon, R., & Nagler, J. (2021). Automated text classification of news articles: A practical guide. *Political Analysis*, *29*(1), 19–42. https://doi.org/https://doi.org/10.1017/pan.2020.8

Bernhard, J., Ashour, R., & Boomgaarden, H. G. (2022, June). Towards Validity Standards of Topic Models in Computational Social Science [Jahrestagung der Fachgruppe Methoden der DGPuK 2022].

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning*, *3*, 993–1022. https://doi.org/10.5555/944919.944937

Bonikowski, B., & Nelson, L. K. (2022). From Ends to Means: The Promise of Computational Text Analysis for Theoretically Driven Sociological Research. *Sociological Methods & Research*, *51*(4), 1469–1483. https://doi.org/10.1177/00491241221123088

Boyd-Graber, J., Hu, Y., & Mimno, D. (2017). Applications of Topic Models. *Foundations and Trends in Information Retrieval*, *11*(2-3), 143–296. https://doi.org/10.1561/1500000030

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. In J. Pei, V. S. Tseng, L. Cao, H. Motoda, & G. Xu (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 160–172). Springer Berlin Heidelberg.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., & Blei, D. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. *Advances in Neural Information Processing Systems*, *22*. Retrieved October 4, 2022, from https://papers.nips.cc/paper/2009/hash/f92586a25bb3145facd64ab20fd554ff-Abstract.html

Chen, Y., Peng, Z., Kim, S.-H., & Choi, C. W. (2023). What We Can Do and Cannot Do with Topic Modeling: A Systematic Review. *Communication Methods and Measures*, *0*(0), 1–20. https://doi.org/10.1080/19312458.2023.2167965

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

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

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, *57*(3), 535–74. https://doi.org/10.1257/jel.20181020

Grimmer, J., & King, G. (2011). General purpose computer-assisted clustering and conceptualization. *Proceedings of the National Academy of Sciences*, *108*(7), 2643–2650.

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.

Hoyle, A., Goel, P., Hian-Cheong, A., Peskov, D., Boyd-Graber, J., & Resnik, P. (2021). Is Automated Topic Model Evaluation Broken? The Incoherence of Coherence. *Advances in Neural Information Processing Systems*, *34*, 2018–2033. https://doi.org/https://doi.org/10.48550/arXiv.2107.02173

Isoaho, K., Gritsenko, D., & Mäkelä, E. (2021). Topic Modeling and Text Analysis for Qualitative Policy Research. *Policy Studies Journal*, *49*(1), 300–324. https://doi.org/10.1111/psj.12343

Krippendorff, K. (2013). *Content analysis: An introduction to its methodology* (3. ed.). Sage. https://ubdata.univie.ac.at/AC08977231

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

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. https://doi.org/10.48550/arXiv.1908.10084

Reiss, M. V., Kobilke, L., & Stoll, A. (2022, June). Reporting Supervised Text Analysis for Communication Science [DGPuK Jahrestagung der FG Methoden].

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 399–408. https://doi.org/10.1145/2684822.2685324

Rodriguez, P. L., & Spirling, A. (2022). Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research. *The Journal of Politics*, *84*(1), 101–115. https://doi.org/10.1086/715162

Scharrer, E., & Ramasubramanian, S. (2021). *Quantitative research methods in communication : The power of numbers for social justice*. Routledge,

Song, H., Tolochko, P., Eberl, J.-M., Eisele, O., Greussing, E., Heidenreich, T., Lind, F., Galyga, S., & Boomgaarden, H. G. (2020). In validations we trust? The impact of imperfect human annotations as a gold standard on the quality of validation of automated content analysis. *Political Communication*, *37*(4), 550–572. https://doi.org/https://doi.org/10.1080/10584609.2020.1723752

Tolochko, P., Balluff, P., Bernhard, J., Galyga, S., Lebernegg, N., & Boomgaarden, H. G. (2022). What's in a Name? The Effect of Named Entities on Topic Modelling Interpretability [[Presented at the annual ICA Conference 2022]].

Wiggins, B. J., & Christopherson, C. D. (2019). The replication crisis in psychology: An overview for theoretical and philosophical psychology. *Journal of Theoretical and Philosophical Psychology, 39*(4), 202.

WKO. (2022, December). *WIRTSCHAFTSLAGE UND PROGNOSE Arbeitslosigkeit* (tech. rep.). https://wko.at/statistik/prognose/arbeitslose.pdf

Ying, L., Montgomery, J. M., & Stewart, B. M. (2022). Topics, Concepts, and Measurement: A Crowdsourced Procedure for Validating Topics as Measures. *Political Analysis*, *30*(4), 570–589. https://doi.org/10.1017/pan.2021.33

Zhao, H., Phung, D., Huynh, V., Jin, Y., Du, L., & Buntine, W. (2021). Topic modelling meets deep neural networks: A survey. *arXiv preprint arXiv:2103.00498*. https://doi.org/https://doi.org/10.48550/arXiv.2103.00498

Appendix

Table 3: Parametrization of models. To implement the LDA model, we utilized pythons gensim library, for T2V the GitHub provided by Angelov, 2020. For more information regarding the parameters please refer to the corresponding implementation

	min_count	**leaf_size**	**min_cluster_size**
T2V30	50	40	850
T2V63	50	40	300
	α	β	
LDA40	0.9	1/40	
LDA50	0.9	1/50	

Table 4: Data statistics. Setting 1 corresponds to 95/0.5 %, setting 2 to 90/1 %. Stopwords removed with pythons $nltk$ library, frequent words with gensims library. For more information please referre to the corresponding documentation

	# of texts	length vocab	avg # tokens	min # of tokens	max # of tokens
tokenized data	24k	68k	207	7	2.9k
setting 1	24k	2.6k	86	2	1.2k
setting 2	24k	1.3k	69	2	858

Table 5: Overview of all topics per model, that could be labeled.

LDA 40	LDA 50	T2V63	T2V30
Agricultural Policy & Climate Change	Attacks on WKSTA & Problems Judicial System	Agricultural Policy	Agricultural Policy
AUGE Union	Care	Agricultural Policy (EU Level)	Alcohol & Smoking Ban
Budget Policy	Corruption & various Political Topics	Alcohol Ban	Antisemitism
Commemoration days	COVID & Government Criticism	Anti-Muslim Racism	Asylum Policy
Criticism of Others 1	Dates 1	Antisemitism	Climate Policy
Criticism of Others 2	Dates 2	Asylum Policy	Congratulations
Danube Island Festival	Discrimination	Austrian Armed Forces	Cultural Policy
Date Announcements	Education Policy	Banks	Democracy
Dates and Announcements	Election Results	Budget Policy	Disputes
Dates and Commemorations SPÖ	Energy & Traffic	Care	Equality for Women
Economic policy	EU Policy	Carinthia	Health Policy
Education policy	Festivals	Christian Trade Union	Islamism
Election Lists & Youth Politics	Financial Policy	Climate Policy	LGBTQIA
Election Results & various Political Topics	Health & Animals	Commitment & Volunteer Fire Brigades	Nuclear Power
Equality Women	Pension Politics	Construction KH Nord	OEVP
EU Politics & Group Members	Promotion (associations & commuters)	Covid Pandemic	Parliamentary Investigations
Health, Development, Nutrition	Public Transport & Rural Areas	Criticism of Others	Pension Policy
Inequalities 1	Rumors & Speculations	Criticism SPOE	Police
Inequalities 2	SPOE 1	Cultural Policy	Press Conference
Names and reports	SPOE 2	Democracy	Reforms
Pension Policy	Suffrage and others	Digitization	Socialist Youth
Renewable Energy & Climate Change	Taxes 1	Drug consumption	Sports Clubs
Rural Area	Taxes 2	Electoral Success	Tax Policy
Scandals 1	Unemployment	European Climate and Energy Policy	Tourism
Scandals 2	Unemployment & Benefits	Floridsdorf Infrastructure	Transport Policy
Security & Social Affairs	Violence against Women	FPOE against all	Unemployment
Unemployment	World trade & Food	Genetic Engineering	Viennese Topics
Violence against women & Antisemitism	Youth Policy	German Language Skills	Youth and Children Policy
Working time & Care		Health Budget	
		Health Policy	
		Housing Communal Building	
		LGBTQIA*	
		New Elections	
		Nuclear Power	
		Obituaries	
		Parking	
		Parliamentary Investigations	
		Pension Policy	
		Police	
		Press Conference 1	
		Press Conference 2	
		Press service 1	
		Press service 2	
		Psychiatry	
		Redesign Mariahilf	
		Reforms	
		Smoking Ban	
		Social Youth	
		Sports Clubs	
		Statistics Austria	
		Tax Policy	
		Tax Policy 2	
		Tourism	
		Transparency	
		Transport Policy	
		Ukraine	
		Unemployment / Labor Policy	
		Vienna	
		Women's Policy	
		Youth and Children Policy	

5 Central Results

The following section discusses the key findings of the three preceding studies, each of which is focusing on different aspects of validating unsupervised computational text analysis methods.

5.1 Study 1: Comprehensive Validation of Word Embeddings for Social Science Research

The study evaluates the performance of various word embedding models trained on a sizeable Austrian news media corpus. It assesses how different hyperparameter settings impact model performance across multiple tasks and compares custom-trained models with off-the-shelf models.

Each parameter setting was run ten times to account for the non-deterministic training process. The stability of the models was then evaluated by examining correlations within "model families," which share the same hyperparameters. It was found that models with larger window sizes and lowercase models were less stable, exhibiting more outliers. Generally, lower correlations were observed with increasing window sizes, a finding attributed to the grammatical structure of the German language, in which verbs are often placed at the end of sentences, which are also typically longer.

The intrinsic validation involved three semantic and one syntactic task, revealing several key points. All models achieved full coverage for semantic tasks, indicating sufficiently large vocabularies for off-the-shelf test questions. For syntactic tasks, coverage varied, with some models achieving only 60-70%. There was no clear impact of lowercasing, minimum count and window size on intrinsic performance. Except for the "opposite" task, lowercasing the training data hurts the performance significantly. Window Size is also negatively related to intrinsic model performance, however, its impact is only significant for the the best match and word intrusion task.

The extrinsic validation focused on three downstream tasks (author classification, topic classification, and sentiment analysis). For author classification, different hyperparameters did not significantly impact performance. In topic classification, results showed variability, but the impact of hyperparameters was quite inconsistent. In sentiment analysis, BERT models outperformed custom models, scoring above 0.59, whereas the Facebook model correctly inferred sentiment only in 45% of the tasks, below the average of the custom models. Overall, it was found that window size and minimum word count had negligible effects, while lowercasing showed mixed results.

The study also included a practical example of differences between the model families, focusing on the terms "Frau" (women) and "Femizid" (femicides) in Austrian news articles,

as word embeddings are often used to assess bias. We found that the top 100 nearest neighbors of each word overlapped, on average, only 50% between the 32 model families. This finding highlights that model choice significantly impacts the substantive results, emphasizing again the need for careful model selection.

Our study concluded that researchers must carefully consider their choice of validation tasks to ensure that it aligns closely with their research objectives and understandings of what they want to validate. Running multiple models and reporting the variation in results enhances the credibility of the findings and helps understand the inherent uncertainty associated with relying on a single model. By prioritizing comprehensive validation strategies and transparent reporting, researchers can notably enhance the trustworthiness and applicability of their models and findings. This approach fosters a deeper understanding of the nuances involved in word embedding validation, ultimately contributing to more robust and credible computational analyses in the social sciences.

5.2 Study 2: Beyond Standardization: A Comprehensive Review of Topic Modeling Validation Methods for Computational Social Science Research

The goal of this study was to systematically review the validation methods used in topic modeling within computational social sciences and assess the extent of standardization across different studies. With the maturation of computational text analysis in the social sciences, topic modeling has become popular for uncovering latent themes in textual data, yet concerns about the validity of its outcomes persist. Our methodology involved a comprehensive search in four scientific databases — Web of Science, Mass Media Complete, ACM Digital Library, and EBSCOhost — using a broad search string for variations of "Topic Model" across titles, abstracts, and keywords, as well as "valid*" in the text, yielding 1,556 studies initially. After applying exclusion criteria, 792 studies were included for further analysis. These studies spanned from 2004 to 2022 and were assessed to determine which validation methods were employed and how these methods evolved over two decades of applying topic models. The findings of our study are important as they shed light on the current state of validation practices in topic modeling, highlighting areas for improvement and future research.

We inductively identified 445 distinct validation methods, later reduced to 138 through a detailed coding process. These methods were grouped into eight overarching categories: Model Comparison, Internal Qualitative Inspection, External Qualitative Inspection, Error Rate Analysis, Distinctiveness of Top Words, Information Theory Metrics, Similarity and Distance Metrics, and Downstream Tasks. Our results revealed a notable absence of standardized validation practices. The analysis showed that most studies (61.4%) mentioned at least one validation method from the Model Comparison category, indicating that comparing different topic models or configurations was a common approach. Internal Qualitative Inspection was the second most frequent category (54.3%), reflecting the importance of qualitative methods to evaluate the relevance and coherence of topics.

Error Rate Analysis and Downstream Tasks were also commonly used, highlighting the reliance on quantitative metrics and practical applications to validate topic models. Over time, we observed shifts in the popularity of different validation methods. The Categories Model Comparison and Error Rate Analysis have declined, while Internal and External Qualitative Inspection have become more prominent. This trend may indicate a growing recognition of the importance of qualitative and context-specific validation approaches in topic modeling. Additionally, Information Theory Metrics, such as Perplexity and Entropy, were less frequently mentioned, suggesting that while these statistical measures are helpful, they are not as commonly relied upon as qualitative inspections. The use of Distinctiveness of Top Words and Similarity and Distance Metrics also varied, with these methods often being applied in conjunction with others to provide a more comprehensive validation. We also calculated Entropy over time to assess whether there is a convergence towards a specific validation method or category. However, we could not find any indices for such convergence when comparing methodological versus substantive studies or core social sciences with peripheral social sciences.

Our review also highlighted that many studies employed multiple validation methods to triangulate their results and enhance robustness. For instance, combining Model Comparison with Internal and External Qualitative Inspections was a common strategy to ensure that the generated topics were statistically sound and contextually meaningful. Nevertheless, there remains significant variability in how validation is reported and conducted, leading to inconsistencies in the reliability and comparability of topic modeling studies.

Our study concluded that the (at least partly) qualitative nature of topic models stands in contrast to the ongoing calls for standardization. We, thus, advocate for including a more qualitative approach to validation, emphasizing the importance of correctly interpreting findings and maintaining transparency throughout the research process. This includes detailed reporting on how validation methods are chosen, what they reveal about the topic models, and why they suit the specific research context. Such practices can enhance the credibility of topic modeling results and foster greater trust in computational social science research.

5.3 Study 3: Topic Model Validation Methods and their Impact on Model Selection and Evaluation

This study aimed to investigate how the choice of validation methods affects model selection and, thus, the outcomes of research applying topic modeling. To test this, we applied two topic modeling algorithms: Latent Dirichlet Allocation (LDA) and Top2Vec (T2V) to the same text corpus of 218,471 Austrian political party press releases from 2004 to 2020. We evaluated the models using different validation strategies: face validity (qualitative inspection), statistical validation (NPMI), semantic validation (word intrusion), content validation (topic labeling), and predictive validation (comparison to external events).

We employed two different prepossessing pipelines for the LDA models, including removing punctuation and digits, lowercasing all characters, stemming, tokenization, and

removing the most and least frequent words. Different hyperparameter settings were tested, and coherence metrics were computed to evaluate the usefulness of the topic models. For T2V, an embedding-based model was used, which relied on a pre-trained neural language model to generate document clusters interpreted as topics.

For LDA, models varied significantly in their interpretability and correlation with external events, with some models showing high coherence scores but poor human interpretability and external correlation. Specifically, the LDA models achieved high NPMI scores, suggesting statistical coherence, but performed poorly in human-based evaluations such as the word intrusion task and topic labeling. For instance, one LDA model (40 topics) had 32.5% of meaningfully interpretable topics, while another (50 topics) had 30% meaningful topics, with a significant portion being uninterpretable or confounding multiple topics.

In contrast, the T2V models were more stable across different validations, suggesting a greater robustness than LDA. The T2V models demonstrated high performance in the word intrusion task, with one model (30 topics) achieving an 84.4% accuracy in identifying intruder words, indicating strong semantic coherence. Additionally, the T2V models showed a higher percentage of meaningful topics when evaluated through topic labeling, with the 30-topic model having 86.7% meaningful topics. The comparison to external events, such as unemployment statistics, also showed better correlations for T2V models, with correlations of $r(14) = 0.68$ for the 30-topic model, indicating a better alignment with real-world data.

When comparing the results from the different validation methods, we could see that different models emerged as performing 'best' depending on which validation method was chosen. However, the overlapping topics identified between the models are limited, suggesting a substantive problem within topic model validation. We highlighted the importance of using a combination of validation strategies to evaluate topic models comprehensively.

Overall, our research emphasizes the need for transparency and rigor in the topic modeling process. We recommend that researchers carefully plan their validation strategies before model application and be explicit about the criteria and benchmarks used for model evaluation. By adopting such comprehensive validation approaches, researchers can better ensure that their topic models provide accurate and meaningful insights, thereby advancing the field of computational text analysis and its applications in social science research.

6 Critical Assessment and Limitations

The dissertation critically examines validation methods for computational text analysis, with a particular focus on unsupervised learning techniques. It was framed by several key decisions and approaches, each of which can be subjected to critical assessment, as well as discussing their implications and potential areas for improvement.

The primary emphasis of this dissertation was on validation, recognizing it as a crucial aspect of ensuring the quality and credibility of computational research. Although validation is undeniably essential, this singular focus meant that other quality criteria, such as reliability, robustness, generalizability, or reproducibility, were not given equivalent attention. This decision was made because of the considerable focus on the difficulty of assessing validation in the field (Baden, Pipal et al., 2022). However, a more holistic approach might have offered a broader understanding of quality assurance in computational methodologies, addressing a broader spectrum of challenges faced by researchers. It would have also provided better insights into the trade-offs and dependencies among these quality criteria.

The dissertation centered on two prevalent computational methods: topic modeling and word embeddings. These techniques are well-established and widely used, contributing significantly to the field (Chen et al., 2023; Grimmer et al., 2022; Rodriguez & Spirling, 2022). Nevertheless, rapid technological advancements have introduced new models and methods that were not included in this study, such as studying the role of validity in research with large language models or chatbots. By not incorporating newer language models and techniques, the dissertation may have missed opportunities to explore more novel approaches that could enhance the understanding and application of these methods. However, this dissertation also underscores broader lessons beyond just topic modeling and word embeddings, particularly the inherent challenges and significant impact of validating computational models, which remain crucial across multiple methodologies.

A significant limitation of this dissertation is its exclusive focus on unsupervised learning methods. This approach was chosen to highlight the unique validation challenges associated with unsupervised techniques. However, this choice excluded discussions on the validity in supervised settings and manual validation processes (]) $\text{song}_{validations}2020.By not addressing these areas, the diss$
$depth exploration of validation's specific challenges and impacts in unsupervised learning, which is particularly c$

The dissertation tackled the issue of selecting appropriate validation methods and their impact on research outcomes, highlighting a critical aspect of the validation process. However, it did not present new validation techniques. Instead, it explored different combinations of existing methods. It found various validation techniques available, with no clear consensus on which ones are best. This diversity in approaches underscores the complexity of the validation process and the need for researchers to carefully consider the specific requirements of their studies when selecting validation methods.

The dissertation's exclusive focus on methodology –without advancing substantive research or theory building– is another area for critical reflection. While methodological rigor is essential, integrating substantive research could have enriched the studies, offering practical examples and applications of the methods discussed. However, it is essential to emphasize that methodological research is crucial as it lays the foundation for all subsequent substantive work. By ensuring that the methods are valid, this dissertation contributes significantly to future research's ability to build substantive knowledge. Integrating substantive questions beyond the case studies could have provided a dual contribution, enhancing methodological understanding and substantive knowledge in the field.

As with any research project, the three studies presented in this dissertation have limitations. One major limitation of the first study on word embedding models is the lack of incorporating human-based validation methods such as the Turing Test, which could offer deeper insights into the human-like quality of the embeddings or alternative validation approaches, such as assessing the distance between cue words, which might provide more nuanced insights. Additionally, the study used a large training corpus, which, while beneficial, also introduced challenges related to computational resources and time constraints and the generalizability of the findings to models trained with smaller corpora. Next, the literature review of the second study only includes studies that explicitly mention validation, potentially overlooking some studies that discuss validation implicitly. The inductive coding process was limited by the feasibility of analyzing all coded methods, leading to the potential under-identification of some methods. Furthermore, the lack of standardized reporting could bias the importance of specific validation methods, and the insufficient separation of analyses between subfields like political science and communication science prevents an assessment of developments in these fields. The third study is based on a single text corpus in one language, limiting the broader applicability of the results and, thus, the generalizability of our findings. Additionally, only three widely used validation methods were examined, leaving out many other potentially relevant methods, and only two topic modeling algorithms were compared (Zhao et al., 2021). The pre-selection of topic models based on statistical and face validity also restricted the scope of the analysis, potentially overlooking other similarities or differences between the model results.

A significant limitation of all three studies is the absence of clear recommendations and blueprints for future research. While the studies critically analyze and identify problems in validating unsupervised computational methods, they fail to provide actionable guidance or standardized methods for addressing these issues. However, the argument made in all three studies –that context-specific validation methods are necessary due to the diverse nature of textual data and varying objectives– is in itself a key recommendation. This suggests that researchers should develop tailored validation strategies sensitive to their studies' specific contexts and goals rather than seeking a one-size-fits-all solution. The variability of social science means that a one-size-fits-all approach to validation is not only impractical and difficult to give, and it might also be a misguided attempt at standardizing validation approaches by overlooking individual details.

Another limitation is that the three studies only look at two methods. Nevertheless, the validation of these two methods has been continuously discussed in the field, with researchers highlighting the need for more methodological research on these two methods. The insights gained here can also inform validation approaches in emerging areas, such as LLM (large-language model) coding, where context-specific validation and unsupervised learning challenges are similarly crucial. These challenges highlight the necessity for ongoing methodological research and the establishment of clear, actionable guidelines to ensure that findings remain robust, generalizable, and relevant amidst the fast-paced evolution of computational techniques.

In summary, the dissertation made some deliberate choices to focus on specific aspects of validation within two methods from the computational social science, offering valuable insights but also facing limitations due to this narrow focus. While it addressed important issues in validation and methodological approaches, it could have been enhanced by incorporating a broader range of techniques, including more methods, proposing and testing novel validation strategies, and integrating substantive research questions. Despite these limitations, the focused approach allowed for a detailed examination of critical validation challenges and contributed to developing nuanced insights for word embeddings and topic Models. The dissertation underscores the importance of robust, context-sensitive validation techniques that can be adapted to evolving computational landscapes and lays a solid foundation for future research, offering a stepping stone for scholars to build upon and expand the understanding of validation in computational social science.

7 Discussion

In contemporary computational research, particularly within the social sciences, the critical importance of robust and transparent model validation cannot be overstated. This dissertation addresses the research question of how to improve the validation of unsupervised computational text analysis methods to achieve credible and consistent results.

All three studies converge on the critical importance of robust validation practices in computational methods. The development, validation, and publication of Austria-specific word embedding models demonstrates the critical role of assessing the stability of results and providing researchers with different models that can be used free of charge. In addition, the comprehensive review of topic model validation methods highlights the need to incorporate more qualitative validation practices in computational social science. Finally, examining the evaluation of topic models highlights the impact of model selection and validation on substantive research outcomes. They highlight the challenges of standardizing validation processes due to the inherent nature of computational models and social scientific research questions, which are often inductive, exploratory and can include multiple different operationalizations of the constructs under study. The studies emphasize the need for context-specific validation, arguing that models should be validated within the specific frameworks of their application. This is not just a technical question; it is a question of ensuring that our scientific efforts make a real contribution to understanding complex social phenomena. Transparency and detailed reporting are recurrent themes in validation, as these practices are essential for enhancing the credibility and replicability of research findings.

The practical recommendations outlined across the three studies emphasize the need for comprehensive approaches to model validation, ensuring reliability, applicability, and transparency in research findings.

Model stability is foundational to producing reliable results. Researchers must evaluate how different hyperparameters affect model stability and performance. Given that these settings significantly impact outcomes, it is imperative to run multiple models to achieve robust results. This iterative process helps to identify the optimal configurations that best capture the nuances of the data, addressing the need for stability and consistency in unsupervised computational text analysis. Regarding word embeddings, this means that employing both intrinsic and extrinsic validation methods is not just recommended –it is essential for a holistic understanding of model performance. Intrinsic methods, which focus on internal consistencies like similarities between words, should be complemented by extrinsic methods that evaluate performance on downstream tasks. Recognizing that a single validation method is often insufficient, combining approaches provides a more comprehensive assessment of a model's efficacy. While the terminology is different this also

holds true for topic modeling. The validation criteria should be adapted to the specific context of the research. It is crucial to align validation methods with the particular tasks and objectives of the study. This context-specific approach ensures that the model accurately captures the intended measurements, enhancing the findings' relevance and applicability. Tailoring validation methods to the research context is key to improving the credibility of computational text analysis. Regarding the unsupervised nature of these methods, which leave humans out of the loop, incorporating qualitative research methodologies into computational methods can significantly bolster the validity of findings. Utilizing thick descriptions and triangulating perspectives from individuals with lived expertise can lead to a more nuanced and accurate interpretation of results. Human-in-the-loop validation methods, where experts review and interpret model outputs, are essential to ensure the models produce meaningful and contextually appropriate results. Enhancing credibility requires detailed and transparent reporting of the application and interpretation of topic modeling methods.

Transparency is a cornerstone of credible research. Documenting the decision-making process, including the rationale behind model selection, validation methods employed, and hyperparameters used, fosters trust and enables others to replicate and assess the findings. Detailed reporting should encompass the entire modeling process, explaining why specific models were chosen and how they were validated. Transparency in these processes is essential for achieving consistent and credible results. Researchers should also explain the reasons behind choosing specific validation methods, explaining what these methods reveal about the model and why they enhance the credibility of the findings. This level of transparency not only bolsters the validity of the research but also facilitates its replication and assessment by others.

Further research using diverse corpora and languages is encouraged to validate findings and improve the generalizability of models. This broader approach helps to understand how models perform across different contexts and datasets, enhancing their robustness and applicability.

A structured approach to validation involves several key steps.

1. Researchers should clearly describe what a good model for a specific study would look like, incorporating theoretical and contextual knowledge.

2. Validation strategies should align with existing theoretical understandings of validity, and utilizing various validation methods to capture different aspects of model validity.

3. Documenting this decision-making process transparently, including the choice of and benchmarks set for validation methods, should receive a definite place in each research publication.

These recommendations, address the research question by underscoring the importance of methodological rigor, transparency, and contextual relevance in computational social science research.

While methodological research is often criticized for lacking direct real-world impact compared to substantive studies, it has significant implications for science and society

at large. Collectively, the reviewed studies highlight the importance of methodological rigor and validation in unsupervised computational text analysis to ensure the quality and applicability of research findings across different domains. This dissertation provides essential tools and frameworks that underpin how we can approach the validity of substantive studies. These contributions are essential for scientific progress and societal advancement, ensuring that research findings are credible and actionable. They help bridge the gap between theoretical research and practical applications, promote trust in scientific results, and support evidence-based decision-making in politics, industry, and beyond. Improving the validity of research methods contributes to better research practices, ultimately leading to more credible scientific results.

Bibliography

Angelov, D. (2020, August). Top2Vec: Distributed Representations of Topics. https://doi.org/10.48550/arXiv.2008.09470

Antoniak, M., & Mimno, D. (2018). Evaluating the Stability of Embedding-based Word Similarities. Transactions of the Association for Computational Linguistics, 6, 107–119. https://doi.org/10.1162/tacl_a_00008

Babbie, E. R. (2020). The practice of social research. Cengage learning.

Baden, C., Boxman-Shabtai, L., Tenenboim-Weinblatt, K., Overbeck, M., & Aharoni, T. (2023). Meaning multiplicity and valid disagreement in textual measurement: A plea for a revised notion of reliability. SCM Studies in Communication and Media, 12(4), 305–326.

Baden, C., Dolinsky, A., Lind, F., Pipal, C., Schoonvelde, M., Guy, S., & van der Velden, M. A. G. (2022, September). Integrated standards and context-sensitive recommendations for the validat (tech. rep. No. Deliverable 6.2). OPTED.

Baden, C., Pipal, C., Schoonvelde, M., & van der Velden, M. A. G. (2022). Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda. Communication Methods and Measures, 16(1), 1–18. https://doi.org/10.1080/19312458.2021.2015574

Balluff, P., Boomgaarden, H. G., & Waldherr, A. (2024). Automatically Finding Actors in Texts: A Performance Review of Multilingual Named Entity Recognition Tools. Communication Methods and Measures, 1–19. https://doi.org/10.1080/19312458.2024.2324789

Balluff, P., Eberl, J.-M., Oberhänsli, S. J., Bernhard, J., Boomgaarden, H. G., Fahr, A., & Huber, M. (2023). The Austrian Political Advertisement Scandal: Searching for Patterns of "Journalism for Sale". https://doi.org/10.31235/osf.io/m5qx4

Baškarada, S., & Koronios, A. (2018). A philosophical discussion of qualitative, quantitative, and mixed methods research in social science. Qualitative Research Journal, 18(1), 2–21. https://doi.org/10.1108/QRJ-D-17-00042

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? Proceedings of the 2021 ACM Conference on Fairn 610–623. https://doi.org/10.1145/3442188.3445922

Bernhard, J., Ashour, R., Tolochko, P., Eberl, J.-M., & Boomgaarden, H. G. (2024). Beyond Standardization: A Comprehensive Review of Topic Modeling for Computational Communication Research [Conference Presentation, ICA, Gold Coast, Australia].

Bernhard, J., Teuffenbach, M., & Boomgaarden, H. G. (2023). Topic Model Validation Methods and their Impact on Model Selection and Evaluation. Computational Communication Research, 5(1), 1–26.

Bernhard-Harrer, J., Balluff, P., & Kathirgamalingam, A. (2024). Training Austrian Specific Word Embeddings from Online News Corpora: OEmbeddings [Conference Presentation, Amsterdam, the Netherlands].

Birkenmaier, L., Lechner, C. M., & Wagner, C. (2023). The Search for Solid Ground in Text as Data: A Systematic Review of Validation Practices and Practical Recommendations for Validation. Communication Methods & Measures, 0(0), 1–29. https://doi.org/10.1080/19312458.2023.2285765

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning, 3, 993–1022. https://doi.org/10.5555/944919.944937

Bonikowski, B., & Nelson, L. K. (2022). From Ends to Means: The Promise of Computational Text Analysis for Theoretically Driven Sociological Research. Sociological Methods & Resea 51(4), 1469–1483. https://doi.org/10.1177/00491241221123088

Boumans, J. W., & Trilling, D. (2018). Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. Digital Journalism, 4(1), 8–23. https://doi.org/10.1080/21670811.2015.1096598

Bowman, S., Goldberg, Y., Hill, F., Lazaridou, A., Levy, O., Reichart, R., & Søgaard, A. (Eds.). (2017, September). Proceedings of the 2nd Workshop on Evaluating Vector Space Repres Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-53

Box, G. E. P., & Draper, N. R. (1987). Empirical model-building and response surfaces. John Wiley & Sons.

Bryman, A. (2006). Integrating quantitative and qualitative research: How is it done? Qualitative research, 6(1), 97–113.

Burawoy, M. (2008). Open the social sciences: To whom and for what? Portuguese Journal of Social Scienc 6(3), 137–146. https://doi.org/10.1386/pjss.6.3.137_1

Chen, Y., Peng, Z., Kim, S.-H., & Choi, C. W. (2023). What We Can Do and Cannot Do with Topic Modeling: A Systematic Review. Communication Methods and Measures, 17(2), 1–20. https://doi.org/10.1080/19312458.2023.2167965

Coenen, L., & Smits, T. (2022). Strong-Form Frequentist Testing In Communication Science: Principles, Opportunities, And Challenges. Communication Methods and Measures, 16(4), 237–265. https://doi.org/10.1080/19312458.2022.2086690

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

Dienlin, T., Johannes, N., Bowman, N. D., Masur, P. K., Engesser, S., Kümpel, A. S., Lukito, J., Bier, L. M., Zhang, R., Johnson, B. K., Huskey, R., Schneider, F. M., Breuer, J., Parry, D. A., Vermeulen, I., Fisher, J. T., Banks, J., Weber, R., Ellis, D. A., … de Vreese, C. (2021). An Agenda for Open Science in Communication. Journal of Communication, 71(1), 1–26. https://doi.org/10.1093/joc/jqz052

DiMaggio, P. (2015). Adapting computational text analysis to social science (and vice versa). Big Data & Society, 2(2), 2053951715602908. https://doi.org/10.1177/2053951715602908

Domahidi, E., Yang, J., Niemann-Lenz, J., & Reinecke, L. (2019). Computational Communication Science | Outlining the Way Ahead in Computational Communication Science: An Introduction to the IJoC Special Section on "Computational Methods for Communication Science: Toward a Strategic Roadmap". International Journal of Communication, 13(9), 3876–3884. Retrieved July 7, 2024, from https://ijoc.org/index.php/ijoc/article/view/10533/2761

Doogan, C. (2022). A Topic is Not a Theme: Towards a Contextualised Approach to Topic Modelling [PhD].

European Commission & Directorate-General for Communication. (2021). European citizens' knowledge and attit Publications Office of the European Union. https://doi.org/10.2775/303708

Faruqui, M., Tsvetkov, Y., Rastogi, P., & Dyer, C. (2016). Problems With Evaluation of Word Embeddings Using Word Similarity Tasks. Proceedings of the 1st Workshop on Evaluating Vector-S 30–35. https://doi.org/10.18653/v1/W16-2506

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. Proceedings of the National Academy of Sciences, 115(16), E3635–E3644. https://doi.org/10.1073/pnas.1720347115

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. Journal of Economic Literature, 57(3), 535–74. https://doi.org/10.1257/jel.20181020

Gladkova, A., & Drozd, A. (2016). Intrinsic Evaluations of Word Embeddings: What Can We Do Better? Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP, 36–42. https://doi.org/10.18653/v1/W16-2507

Godfrey-Smith, P. (2009). Theory and reality: An introduction to the philosophy of science. University of Chicago Press.

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). Text as data: A new framework for machine learning and Princeton University Press.

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

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794.

Guo, L., Mays, K., Lai, S., Jalal, M., Ishwar, P., & Betke, M. (2020). Accurate, Fast, But Not Always Cheap: Evaluating "Crowdcoding" as an Alternative Approach to Analyze Social Media Data. Journalism & Mass Communication Quarterly, 97(3), 811–834. https://doi.org/10.1177/1077699019891437

Heft, A., & Buehling, K. (2022). Measuring the diffusion of conspiracy theories in digital information ecologies. Convergence, 28(4), 940–961. https://doi.org/10.1177/13548565221091809

Heft, A., Buehling, K., Zhang, X., Schindler, D., & Milzner, M. (2024). Challenges of and approaches to data collection across platforms and time: Conspiracy-related digital traces as examples of political contention. Journal of Information Technology & Politics, 21(3), 323–339. https://doi.org/10.1080/19331681.2023.2250779

Hilbert, M., Barnett, G., Blumenstock, J., Contractor, N., Diesner, J., Frey, S., Gonzalez-Bailon, S., Lamberso, P. J., Pan, J., Tai-Quan Peng, Shen, C., Smaldino, P. E.,

Van Atteveldt, W., Waldherr, A., Zhang, J., & Zhu, J. J. H. (2019). Computational Communication Science: A Methodological Catalyzer for a Maturing Discipline. International journal of communication, 13, 3934.

Hovy, D., & Prabhumoye, S. (2021). Five sources of bias in natural language processing. Language and Linguistics Compass, 15(8). https://doi.org/10.1111/lnc3.12432

Hoyle, A., Goel, P., Hian-Cheong, A., Peskov, D., Boyd-Graber, J., & Resnik, P. (2021). Is Automated Topic Model Evaluation Broken? The Incoherence of Coherence. Advances in Neural Information Processing Systems, 34, 2018–2033. https://doi.org/10.48550/arXiv.2107.02173

Humphreys, L., Lewis, N. A., Jr, Sender, K., & Won, A. S. (2021). Integrating Qualitative Methods and Open Science: Five Principles for More Trustworthy Research. Journal of Communication, 71(5), 855–874. https://doi.org/10.1093/joc/jqab026

Ignatow, G. (2016). Theoretical Foundations for Digital Text Analysis. Journal for the Theory of Social Be 46(1), 104–120. https://doi.org/10.1111/jtsb.12086

Isoaho, K., Gritsenko, D., & Mäkelä, E. (2021). Topic Modeling and Text Analysis for Qualitative Policy Research. Policy Studies Journal, 49(1), 300–324. https://doi.org/10.1111/psj.12343

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. Springer.

Kathirgamalingam, A., Lind, F., Bernhard-Harrer, J., & Boomgaarden, H. G. (2024). Exploring Coder Bias: An Investigation of Human and LLM-based Classification of Racism in News Media [Conference Presentation COMPTEXT 2024, Amsterdam, the Netherlands].

Kathirgamalingam, A., Lind, F., & Boomgaarden, H. G. (2024). Measuring Racism and Related Concepts Using Computational Text-as-Data Approaches: A Sytematic Literature Review. [Conference Presentation, ICA 2024, Gold Coast, Austrlia].

Kirk, J., & Miller, M. L. (1986). Reliability and validity in qualitative research (Vol. 1). Sage.

Krippendorff, K. (2013). Content analysis: An introduction to its methodology (3. ed.). Sage.

Kroon, A., Trilling, D., & Raats, T. (2020). Guilty by Association: Using Word Embeddings to Measure Ethnic Stereotypes in News Coverage - Anne C. Kroon, Damian Trilling, Tamara Raats, 2021. Jouralism & Mass Communication Quarterly, 89(2). https://doi.org/10.1177/1077699020932304

Kruschke, J. K., & Liddell, T. M. (2018). Bayesian data analysis for newcomers. Psychonomic bulletin & re 25(1), 155–177.

Lazer, D. M. J., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., Freelon, D., Gonzalez-Bailon, S., King, G., Margetts, H., Nelson, A., Salganik, M. J., Strohmaier, M., Vespignani, A., & Wagner, C. (2020). Computational social science: Obstacles and opportunities. Science (New York, N.Y.), 369(6507), 1060–1062. https://doi.org/10.1126/science.aaz8170

Lehmann, C. (2004). Data in linguistics. 21(3-4), 175–210. https://doi.org/10.1515/tlir.2004.21.3-4.175

Lewis, N. A. (2020). Open Communication Science: A Primer on Why and Some Recommendations for How. Communication Methods and Measures, 14, 82. https://doi.org/10.1080/19312458.2019.1685660

Licht, H., & Lind, F. (2023). Going cross-lingual: A guide to multilingual text analysis. Computational Communication Research, 5(2). https://doi.org/10.5117/CCR2023.2.2.LICH

Lind, F., Gruber, M., & Boomgaarden, H. G. (2017). Content Analysis by the Crowd: Assessing the Usability of Crowdsourcing for Coding Latent Constructs. Communication Methods & Measur 11(3), 191–209. https://doi.org/10.1080/19312458.2017.1317338

Lind, F., Schoonvelde, M., Baden, C., Dolinsky, A., Pipal, C., & Van Der Velden, M. (2023). A validation framework for multilingual computational text analysis in the social sciences [Conference Presentation ECPR 2023, Prague, Czech Republic].

Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-Assisted Text Analysis for Comparative Politics. Political Analysis, 23(2), 254–277. https://doi.org/10.1093/pan/mpu019

Maier, D., Niekler, A., Wiedemann, G., & Stoltenberg, D. (2020). How document sampling and vocabulary pruning affect the results of topic models. Computational Communication Research, 2(2), 139–152. https://computationalcommunication.org/ccr/article/view/32

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., & Adam, S. (2018). Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology. Communication Methods & Measures, 12(2-3), 93–118. https://doi.org/10.1080/19312458.2018.1430754

Margolin, D. B. (2019). Computational Contributions: A Symbiotic Approach to Integrating Big, Observational Data Studies into the Communication Field. Communication Methods and Measur 13(4), 229–247. https://doi.org/10.1080/19312458.2019.1639144

McKee, A. (2003). Textual Analysis : A Beginner's Guide. Sage Publications.

Montgomery, A. C., & Crittenden, K. S. (1977). Improving coding reliability for open-ended questions. Public Opinion Quarterly, 41(2), 235–243.

Nelson, L. K. (2020). Computational Grounded Theory: A Methodological Framework. Sociological Methods & Research, 49(1), 3–42. https://doi.org/10.1177/0049124117729703

Okasha, S. (2002). Philosophy of science: A very short introduction. Oxford Paperbacks.

Pääkkönen, J. (2021). Data Do Not Speak for Themselves: Interpretation and Model Selection in Unsupervised Automated Content Analysis. In Composition and Big Data. University of Pittsburgh press.

Peter, J., & Lauf, E. (2002). Reliability in cross-national content analysis. Journalism & mass communication quar 79(4), 815–832.

Pipal, C., Song, H., & Boomgaarden, H. G. (2023). If You Have Choices, Why Not Choose (and Share) All of Them? A Multiverse Approach to Understanding News Engagement on Social Media. Digital Journalism, 11(2), 255–275. https://doi.org/10.1080/21670811.2022.2036623

Plank, B. (2022). The'Problem'of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. arXiv preprint arXiv:2211.02570. https://doi.org/10.48550/arXiv.2211.02570

Reiss, M. V. (2023, April). Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark. https://doi.org/10.48550/arXiv.2304.11085

Reiss, M. V., Kobilke, L., & Stoll, A. (2022, June). Reporting Supervised Text Analysis for Communication Science [Conference Presentation DGPuK Jahrestagung der FG Methoden, München, Germany].

Rodriguez, P. L., & Spirling, A. (2022). Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research. The Journal of Politics, 84(1), 101–115. https://doi.org/10.1086/715162

Romney, D., Stewart, B. M., & Tingley, D. (2015). Plain Text: Transparency in the Acquisition, Analysis, and Access Stages of the Computer-assisted Analysis of Texts. Qualitative and Multi-Method Research, 13(1), 32–37.

Scharrer, E., & Ramasubramanian, S. (2021). Quantitative research methods in communication : The po Routledge.

Song, H., Tolochko, P., Eberl, J.-M., Eisele, O., Greussing, E., Heidenreich, T., Lind, F., Galyga, S., & Boomgaarden, H. G. (2020). In validations we trust? The impact of imperfect human annotations as a gold standard on the quality of validation of automated content analysis. Political Communication, 37(4), 550–572. https://doi.org/https://doi.org/10.1080/10584609.2020.1723752

Tolochko, P., Balluff, P., Bernhard, J., Galyga, S., Lebernegg, N. S., & Boomgaarden, H. G. (2024). What's in a name? The effect of named entities on topic modelling interpretability. Communication Methods and Measures, 1–22. https://doi.org/10.1080/19312458.2024.2302120

Tolochko, P., & Boomgaarden, H. G. (2019). Determining Political Text Complexity: Conceptualizations, Measurements, and Application. International Journal of Communication, 13(0), 1784–1804. Retrieved June 27, 2024, from https://ijoc.org/index.php/ijoc/article/view/9952

Törnberg, P., & Uitermark, J. (2021). For a heterodox computational social science [Publisher: SAGE Publications Ltd]. Big Data & Society, 8(2), 20539517211047725. https://doi.org/10.1177/20539517211047725

van Atteveldt, W., Trilling, D., & Calderon, C. A. (2022). Computational Analysis of Communication. John Wiley & Sons.

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

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

Van Atteveldt, W., Strycharz, J., Trilling, D., & Welbers, K. (2019). Computational communication science| toward open computational communication science: A practical road map for reusable data and code. International Journal of Communication, 13, 20.

Waldherr, A. (2014). Emergence of News Waves: A Social Simulation Approach. Journal of Communication, 64(5), 852–873. https://doi.org/10.1111/jcom.12117

Wallach, H. (2018). Computational social science computer science + social data. Communications of the ACM, 61(3), 42–44. https://doi.org/10.1145/3132698

Welbers, K., Van Atteveldt, W., & Benoit, K. (2017). Text Analysis in R. Communication Methods & Measures, 11(4), 245–265. https://doi.org/10.1080/19312458.2017.1387238

Wiggins, B. J., & Christopherson, C. D. (2019). The replication crisis in psychology: An overview for theoretical and philosophical psychology. Journal of Theoretical and Philosophical Psycholog 39(4), 202.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3(1). https://doi.org/10.1038/sdata.2016.18

Williams, M. L., Burnap, P., & Sloan, L. (2017). Towards an ethical framework for publishing Twitter data in social research: Taking into account users' views, online context and algorithmic estimation [Publisher: Sage Publications Sage UK: London, England]. Sociology, 51(6), 1149–1168.

Winter, G. (2000). A Comparative Discussion of the Notion of 'Validity' in Qualitative and Quantitative Research. The Qualitative Report, 4(3-4). https://doi.org/10.46743/2160-3715/2000.2078

Ying, L., Montgomery, J. M., & Stewart, B. M. (2022). Topics, Concepts, and Measurement: A Crowdsourced Procedure for Validating Topics as Measures. Political Analysis, 30(4), 570–589. https://doi.org/10.1017/pan.2021.33

Zhao, H., Phung, D., Huynh, V., Jin, Y., Du, L., & Buntine, W. (2021). Topic modelling meets deep neural networks: A survey. https://doi.org/10.48550/arXiv.2103.00498