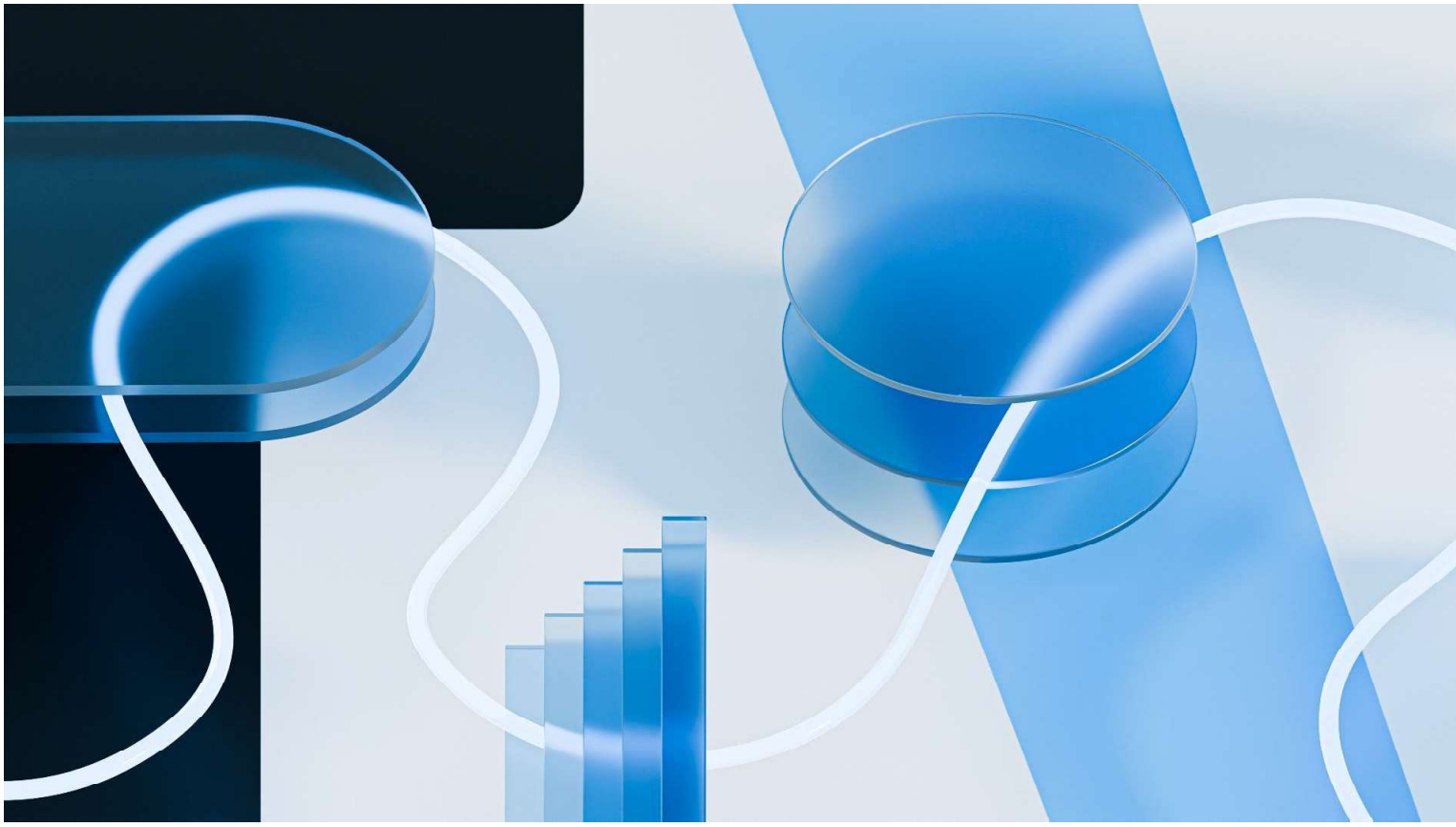
twon
twin of online social networks



TWin of Online Social Networks

Deliverable D5.2

Definition of Metrics

Main Authors: Sjoerd B. Stolwijk, PhD

Funded by
the European Union

twon
twin of online social networks

About TWON

TWON (project number 101095095) is a research project, fully funded by the European Union, under the Horizon Europe framework (HORIZON-CL2-2022-DEMOCRACY-01, topic 07). TWON started on 1 April 2023 and will run until 31 March 2026. The project is coordinated by the Universiteit van Amsterdam (the Netherlands) and implemented together with partners from Universität Trier (Germany), Institut Jozef Stefan (Slovenia), FZI Forschungszentrum Informatik (Germany), Karlsruher Institut für Technologie (Germany), Robert Koch Institute (Germany), Univerzitet u Begogradu - Institut za Filozofiju I Drustvenu (Serbia) and Slovenska Tiskovna Agencija (Slovenia).

Funded by the European Union. Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

Project Name	Twin of Online Social Networks
Project Acronym	TWON
Project Number	101095095
Deliverable Number	D5.2
Deliverable Name	Definition of Metrics
Due Date	31.03.2025
Submission Date	28.03.2025
Type	R — Document, report
Dissemination Level	PU - Public
Work Package	WP 5
Lead beneficiary	1-UvA
Contributing beneficiaries and associated partners	Universität Trier (UT), FZI Forschungszentrum Informatik (Germany), Karlsruher Institut für Technologie (KIT), Robert Koch Institut (RKI), Univerzitet u Begogradu - Institut za Filozofiju I Drustvenu (UoB), Slovenska Tiskovna Agencija (STA), Dialogue Perspectives e.V (DIA)

Executive Summary

This report proposes a set of metrics to determine the deliberative quality of discussions on social media in general and TWON in particular. Chapter 2 lists the key indicators of:

- Exposure to political content
- Engagement with political content
- Contributing political content
- Diversity of exposure
- Quality of exposure

. Chapter 3 explains how and why this set of indicators differs from the typical list of deliberative indicators and proposes to view deliberation from a summative rather than an additive perspective. The traditional additive perspective simply argues that to achieve societal deliberation, people need to deliberate. The best way to strengthen deliberative democracy would thus be to organize good deliberation, which is then "added" (-> additive) to the societal debate. Often, organizing such deliberation is challenging at a large scale, especially on social media platforms. In contrast, the summative approach does not require any one venue of good deliberation to achieve deliberative democracy at a societal scale but rather aims to realise it only at the collective, i.e. summative level. In this view, social media do not need to aim at perfect deliberation within one platform; rather, the goal is to contribute to deliberation at a societal scale via the platform. We propose that social media can contribute especially by offering an avenue for users (citizens, journalists and politicians alike) to be exposed to political debate, but also to engage and participate in that debate. In addition, social media can connect otherwise unconnected users and expose them to ideas they might otherwise have missed. Ideally, these ideas are substantiated with arguments and evidence. Chapter 4 evaluates a large set of automatic classifiers to determine the degree to which social media comments meet several deliberative criteria, specifically whether comments are rational, interactive, diverse and civil. Results show how more modern techniques like fine-tuned transformers and generative large language models have improved our ability to reproduce manual codings automatically, but also that results vary considerably between models. Chapter 5 integrates the aims of Chapter 3 with the results of Chapter 4 and translates them to the case of TWON to arrive at the metrics proposed in Chapter 2. It adds tests of the performance of different classifiers to determine whether a comment is political or not. Chapter 6 takes a look into the future, beyond what is currently feasible for TWON, to explore whether new techniques can help determine

the deliberative quality of online social media debates to the more fine-grained level of specific claims and shows some promising first results.

List of Tables

List of Figures

twon
twin of online social networks

Contents

List of Abbreviations

BERT	Bidirectional Encoder Representations from Transformers
eMFD	extended Moral Foundations Dictionary
FK	Flesch-Kincaid index
GLLM	Generative Large Language Model
IC	Integrative complexity score
LIWC	Linguistic Inquiry and Word Count
LLM	Large Language Model
MFD	Moral Foundation Dictionary
MTEB	Massive Text Embedding Benchmark
mxbai	Mixedbread AI
OSN	Online Social Network
SML	Supervised Machine Learning
STS	Semantic Text Similarity
SVC	Support Vector Classifier
TWON	Twin of Online Social Networks
UvA	University of Amsterdam
WDYL	Google What Do You Love Project

Definition of Metrics

Sjoerd Stolwijk*

March 25, 2025

1 Introduction

To conduct research into the effect of platform mechanics on public discourse, we need a clear conceptualization of what actually is considered a good public discourse. According to the description of the action (p. 5):

> "The second key aim of TWON is to propose metrics to evaluate and optimize the design of online social networks based on epistemic, social, and democratic ideals. This will allow governments to define unequivocal benchmarks that providers of online communication technology must meet to prevent deleterious effects on societies. Plus, it will provide these companies a rigorous tool to optimize their platforms on dimensions that go beyond economic interests and put central the functioning of public debate, societal cohesion, and democratic decision making. OSNs have been criticized, for instance, for contributing to the spreading of falsehoods and allowing foreign powers to influence public debate by disseminating it. Professional and user-based fact-checking has been suggested as a solution, but it is extremely costly, too slow, and hardly achievable for an individual user (Keijzer & Mäs, 2022). But a well-functioning democratic discourse depends on much more than just the absence of falsehoods. Previous work has used manual content analysis to quantitatively measure if user comments conform to Habermasian ideals of deliberative democratic discourse (Trilling et al., 2016; Ziegele et al., 2018), such as bringing in additional knowledge, not using uncivilized language, and staying on-topic. Other work

*This report draws strongly on collaborative research projects. We would like to acknowledge the co-authors of these papers: Mark Boukes, Wang Ngai Yeung, Yufang Liao, Simon Münker, Anne C. Kroon, Damian Trilling, Corinna Oschatz and Michael Heseltine. We would like to thank Simon Münker and Fabio Sartori for reviewing the internal draft. The report has benefited considerably from their comments.

> has grouped variables to measure discussion quality into the dimensions rationality, relevance, reciprocity, and politeness & respect, each consisting of multiple specific indicators (Berg, 2016). Some attempts have been made to automate these metrics, such as using word vector similarities to determine on-topic-ness of a user comment (Park et al., 2016). Yet, a consistent methodology to quantitatively measure debate quality on OSNs at scale does not exist yet." (Project: 101095095 — TWON — HORIZON-CL2-2022-DEMOCRACY-01 ANNEX 1: DESCRIPTION OF THE ACTION, p5)

This report presents the metrics (Chapter 2) we advocate for our project to automatically measure the contribution to deliberative democratic discourse by online social communication. In line with Objective 2.1, and Verifiable Success Indicators VSI-2.1 and VSI-2.2, we will present the theoretical foundation of these metrics (Chapter 3) as well as their empirical validation (Chapter 4) compared to a large inventory of 50+ other metrics relative to a manually coded gold standard. In Chapter 5 we construct the final metrics as presented in Chapter 2 and explain our choices. Finally, we add some first explorations into even better potential future metrics (Chapter 6) which can be further tested and developed within the remaining TWON-project duration or thereafter.

2 Metrics of debate quality

For TWON we propose the metrics for online debate quality presented in Table 1. As will be explained in Chapter 3, the different indicators are non-compensatory: a surplus in one indicator does not automatically make up for a lack in another. Therefore, no single debate quality score will be proposed. Rather, different research projects using this metric can emphasize different aspects/indicators of online debate quality in line with their specific goals. Table 1 presents the metrics in a general form. Specific research projects might be interested in different levels of debate quality. As will be explained in Chapter 5, these metrics can be amended for use at the individual level (what is the quality of debate observed per individual social media platform user), thread level (what is the quality of a particular thread of comments, debate/topic level (what is the quality of all comments connected to a specific debate, for example the war in Ukraine) or platform level (what is the quality of debate on the platform as a whole, for example to compare this quality to alternative platforms or platform mechanics). The choice of the classification model is based on the evaluation presented in Chapter4 and explained in Chapter 5.

In addition to the core metrics in Table 1, we also include a set of additional important metrics relevant to debate quality listed in Table 2. Like the core metrics, these metrics are evaluated in Chapter 4

Table 1: TWON Core Debate Quality Metrics.

Indicator	**Operationalization**
Exposure to political content	Share of comments classified as political with Llama3.1:70b present in the thread to which the participant is exposed
Engagement with political content	Number of political comments liked or shared per participant as classified political with Llama3.1:70b
Contributing political content	Number of comments posted per participant which are subsequently classified as political with Llama3.1:70b
Diversity of exposure	The ideological balance between left, neutral and right-leaning political comments to which a participant is exposed as classified with Llama3.1:70b, if a post is classified as belonging to the minority ideology in a thread it adds a score of 2 to diversity, in the case of a tie it adds 1, otherwise, it adds zero to the cumulative diversity indicator per thread
Quality of exposure	The share of comments to which a participant is exposed which are classified with Llama3.1:70b as substantiating, or expanding on, any claims made within the comment

and explained in Chapter 5.

Table 2: TWON Supporting Debate Quality Metrics.

Indicator	**Operationalization**
Incivility	Share of comments classified as uncivil with Llama3.1:70b to which a participant is exposed
Interactivity	Share of comments classified as interactive with Llama3.1:70b to which a participant is exposed
Novelty	Share of comments classified as having a new topic compared to previous topics to which a participant is exposed with cardiffnlp/tweet-topic-21-multi

We do not consider these metrics the final answer to measuring debate quality for online social media platforms but as a good first step. We are continously exploring how to improve them. In Chapter6, we will present findings of ongoing research which shows potential to further improve the conceptualization and operationalization of the diversity metric.

3 Theory: What is online debate quality?

We propose a set of measures in line with modern interpretations of the role of social media for deliberative democratic theory. Four TWON project members (Sjoerd Stolwijk, Corinna Oschatz, Michael Heseltine, and Damian Trilling) together wrote a conference paper that addresses this question titled "Refining deliberative standards for online political communication: Introducing a summative approach to designing deliberative recommender systems", which is published in Proceedings of NORMalize 2023: The First Workshop on the Normative Design and Evaluation of Recommender Systems, co-located with the ACM Conference on Recommender Systems (RecSys '23), September 18-22nd, 2023, Singapore (https://ceur-ws.org/Vol-3639/paper5.pdf). The following section is a near-copy of this paper.

3.1 Introduction

The development of online debate indicators is a flourishing field (Goddard and Gillespie, 2023). However, existing metrics are not always properly grounded in democratic theory – and this is less straightforward than it may seem. To begin with, anyone who wishes to develop systems that facilitate online discussions along normative democratic lines has to choose which out of various alternative conceptions of democracy they want to rely on (Helberger, 2019; Vrijenhoek et al., 2021). For example, based on so-called liberal theories of democracy one may argue that topics that are prominent on the political agenda should also be prominent in a discussion, while based on so-called critical theories, one would expect marginal topics to be discussed instead; so-called participative models would draw attention to topics citizens "should know", and so-called deliberative models would focus more on the diversity of the discussion (Helberger, 2019). In this paper, we focus on deliberative models of democracy. Deliberative democracy is often considered the most demanding in terms of the required quality of (online) political discussion, and thereby provides a nice ideal point to strive for (cf. Strömbäck, 2005). Its emphasis on discussion also fits well with the nature of online communication platforms.

Although there are many variants of deliberative theory, the work of Habermas (1962, 1984, 1996) is considered a cornerstone in this literature (cf. Caluwaerts et al., 2023; Goddard and Gillespie, 2023). Deliberative democracy is about more than making decisions based on aggregating the preferences of citizens, or even than preferences based on accurate information, but rather holds that (collective) preferences should be formed through inclusive, reasoned debate (Freelon, 2010; Strömbäck, 2005). In this view, online communication needs to provide relevant and accurate information, but also connect citizens and motivate them to share and debate their views in a deliberative way.

Scholars seeking to measure deliberation online have proposed and listed various indicators. They

measure concepts like equality, diversity, rationality, interactivity, civility and reference to the common good (Vrijenhoek et al., 2021; Goddard and Gillespie, 2023; Nelimarkka et al., 2019; Oswald, 2022). However, less attention has been paid to whether and how these indicators can best be put to use to realize deliberative democratic ideals on a societal scale. In political science, scholars have struggled with similar issues when studying the deliberative values of citizen participation initiatives. Although they are using non-computational methods, their insights can be of use to the computational community as well. Similar to computational scholars, deliberative political scientists started out evaluating whether and to what degree the elements of deliberative democratic debate could be found in the exchanges between citizens. After three decades of sustained, wide academic attention for this topic, they are now moving away from what is called an "additive" approach to deliberation and question whether deliberative quality is best treated as a single, unitary, concept consisting of all of its elements to an equal degree regardless of context (Bächtiger and Parkinson, 2019; Thompson, 2008). The additive approach holds that "Deliberation [...] is produced by specific methods or institutions which then add it—inject it, if you will—into the system more broadly" (Bächtiger and Parkinson, 2019, p.7). That is, they sought to construct an ideal forum to foster deliberation in all its facets and use this as a way to produce deliberative democracy at a societal scale. One example is forming a Citizen Assembly in which a stratified sample of the population is invited to deliberate on a policy proposal over several days and then share their arguments and propose their decision to the public at large and a legislative body in particular (e.g., see Már and Gastil, 2021). Deliberation in this view is often seen in a more-is-better fashion, so more diversity in a news feed is better for deliberative democracy than less diversity, more interactivity is better than less and so on (Thompson, 2008). Although intuitive, the additive approach has been criticized for failing to grasp the complexity of human behavior and for being ill-equipped to relate the outcomes of individual (online) interactions back to the over-arcing goals of deliberative democracy on a societal scale (Mutz, 2008; Bächtiger and Parkinson, 2019). Notably, even Habermas himself has recently questioned the approach of using his indicators as a yardstick for societal debate, stating that "I do not see deliberative politics as a farfetched ideal against which sordid reality must be measured" (Habermas, 2022, p. 149).

In the next section, we outline the critique political scientists have leveled against the additive approach. Subsequently, we sketch different views on how (micro) online deliberations fit within the larger deliberative democratic ideal. Finally, we propose an alternative conception of debate quality for online platforms, based on the recently proposed systematic, summative approach to deliberative democracy (Mansbridge et al., 2012; Bächtiger and Parkinson, 2019). The summative approach does not seek to optimize deliberation at any single venue but rather maximizes its value at the aggregate,

societal scale. This alternative treats online platforms as complementary rather than substitutive to traditional media. It seeks to realize deliberative goals at a societal scale (i.e. summative) and social media can act as a contributing factor rather than requiring full deliberation to occur at the micro-level of the platform first and then *adding* (i.e. additive) the outcome into society to produce deliberative democracy.

3.2 More is better, or is it?

Up until recently, either implicitly or explicitly, proponents of computationally measuring deliberative quality often appear to advocate an additive, more-is-better approach. Some describe how a specific debate aspect, like diversity, is helpful for deliberative democracy and then propose how to measure it, others list indicators a good debate needs in order to be deliberative (Goddard and Gillespie, 2023; Oswald, 2022; Vrijenhoek et al., 2021; Helberger, 2019). Like early deliberative political scientists, computational scholars then evaluate whether a particular debate fulfills these criteria and to which degree (Beauchamp, 2020, e.g.,[). Based on this approach one could either aggregate indicator scores into a total deliberative score, or define minimum/maximum values needed to pass platform moderation (such as for civility). However, doing so would imply assuming that deliberation is a unitary concept, i.e. that all criteria need to be fulfilled together to a certain degree for the ascribed benefits to deliberative democracy to materialize. While intuitive, Mutz (2008) argues that one should be careful, since this approach carries some assumptions about human behavior that are unwarranted. Also adding the scores of various criteria together to one overall deliberative score implies that deficiencies on the score for one criteria can be compensated for by higher scores on other criteria, while using indicators for minimum benchmark values implies that failing to meet the minimum level on one criteria disqualifies any achievements on other criteria.

A growing literature questions that deliberation is a unitary concept and all indicators necessarily need to go together (e.g., Mutz, 2008; Bächtiger and Parkinson, 2019). In fact, there are many situations where various indicators of deliberation appear to be at odds with one another (Thompson, 2008). For example, diversity might conflict with inclusion: the more diverse the voices in a debate, the less likely it is that everyone participates, since debate requires conflicting arguments and most people feel uncomfortable being confronted by (too many) opposite opinions (cf. Reuver et al., 2021; **?**). Likewise representation, accountability and openness might conflict with civility: publicity makes representatives more accountable to their supporters, but is also found to be less conducive to mutual respect and constructive politics than deliberation behind closed doors (Bächtiger and Parkinson, 2019). This directly highlights one of the main problems in deliberative democratic theory: that quality aspects

of debates are frequently conflated with positive effects of deliberative debates in definitions of what counts as deliberative democracy (Mutz, 2008; Friess and Eilders, 2015). This lack of conceptual clarity in stipulating causes and effects masks the myriad in conceptualizations of what exactly constitutes deliberation as well as what its effects are supposed to be.

Like deliberative indicators, deliberative effects/outcomes are also more complex empirically than a unitary concept approach would appear to imply. A great many presumed effects can be found in the literature. Bächtiger and Parkinson (2019) group them as: (1) epistemic outcomes (find best possible approach to handle a common problem); (2) ethical outcomes (follow the rational argument-dominated deliberative process as a goal in itself); (3) providing legitimacy (of collective decisions formed through deliberation); (4) emancipation of minority groups (providing a space to make all citizens heard); (5) transformation and clarification of preferences (people learn from the debate and change their views or deepen their perspectives in return); with some also listing (6) consensus as a desired outcome. Although these goals overlap to a certain degree, different deliberative aspects contribute to them in a different degree. Scholars have listed many contradictions between different aspects of deliberation and deliberative outcomes. For example, if the process of rational argumentation is the goal, the resulting formal tone of the debate might be off-putting to some citizens, and dominant groups might use their definition of what counts as a rational argument to suppress minority voices, also requiring each position to come with an elaborate set of supporting arguments might favour established and well-documented positions over new voices that as of yet have not had the time and space to develop such arguments (Sanders, 1997; Dahlberg, 2007).

Given these conflicts between indicators of deliberation both among themselves and in relation to the desired deliberative outcomes, Mutz (2008) proposes to abandon the unitary concept of deliberation and instead investigate which deliberative aspects have which effects under which conditions. She argues that it is likely that different deliberative aspects might interact with each other to reach specific outcomes, and should therefore be studied separately as well as in different combinations. Applying this line of argumentation to social media platform design, we can say that if a system optimizes either for (1) all deliberative indicators, or (2) an indexed value of debate quality based on combining various indicators, or (3) some minimum/maximum values of deliberative indicators, then the contradictions between these indicators and outcomes as described above might lead this system to hinder rather than enable the deliberative process. In the next section, we will elaborate more on the argument that what is seemingly good for deliberation on a micro-level does not necessarily lead to good deliberative outcomes on the societal macro-level.

3.3 The gap between micro and macro

The second main critique of political scientists on using deliberative indicators as a yardstick to measure contributions to normative standards of deliberative democracy is the problematic relation between specific debates (micro) and deliberative democracy on a societal scale (macro) (Chambers, 2009). Most accounts of deliberative democracy aim at the societal level rather than that of the individual debate. They require the deliberative process to be democratic: that it culminates in a collectively binding decision (Thompson, 2008). Habermas (2022) stipulates how issues, information and arguments in debates between citizens are picked up by societal actors, like social movements, PR-organizations, political parties and the media which translate positions into coherent discourses relating distinct positions to relevant arguments, which then feed into the political arena to result in collective decisions. He appears to be unsure, though, how to fit online debates productively into his framework, and argues that such debates might actually be counterproductive warning that "[t]he platforms do not offer their emancipated users any substitute for the professional selection and discursive examination of contents based on generally accepted cognitive standards" (Habermas, 2022, p. 160), and "the increasing dissonance of a strident diversity of voices and the complexity of the challenging topics and positions is leading a growing minority of media consumers to use digital platforms to retreat into shielded echo chambers of the like-minded. For the digital platforms not only invite their users to spontaneously generate intersubjectively confirmed worlds of their own but seem to lend the stubborn internal logic of these islands of communication, in addition, the epistemic status of competing public spheres" (Habermas, 2022, p. 162)[1]. On top of this, deliberation between *all* citizens in a modern society is practically unfeasible due to constraints in time and resources, and can therefore only be realized at the institutions of the state (Habermas, 2022). However, empirical research into the political arena, where the final deliberative debate between contrasting discourses should culminate into collective decisions, finds that parliamentary debate is oriented towards voting rather than aggregating information and participants rarely change their preferences in view of contrasting information (Bächtiger and Parkinson, 2019). It is therefore unclear whether and how online political communication contributes to deliberative *democracy* at all.

This focus on relating the micro to the macro is often labelled as the "systemic turn", as it views deliberative democracy as a larger system rather than a debate at large (Chambers, 2009; Owen and Smith, 2015; Mansbridge et al., 2012). Political scientists have proposed three main theoretical frameworks that specify how specific debates between citizens could translate into collectively binding decisions

[1] Note that Zuiderveen Borgesius et al. (2016) found little empirical support for the relation between recommender systems and echo chambers or filter bubbles.

in a deliberative way. Bächtiger and Parkinson (2019) group them into discursive, sequential and spatial models of deliberation. The discursive model focuses on how people understand and shape society through discourses that find their way into arguments, decisions and policy (Dryzek, 1990). Various versions of the sequential model hold that feedback loops ensure deliberative outcomes over time, where societal debate influences political decisions, which are in turn part of societal debate to critique, alter, maintain or reject them at another round of political decisions. The spatial model specifies distinct deliberative functions of distinct institutions and the proper relations between these institutions to ensure deliberative outcomes. Each of these models has received its fair portion of critique, where the discursive model is unclear on how deliberation contributes in what way to the forming of discourses and how these discourses feed into policy in a deliberative way, the sequential model is often empirically incorrect in policy preceding debate, especially for non-salient issues, and the spatial model is found to be too static to encompass the creatively changing nature of political decision making with new forms or organisation continuously popping up (e.g. `#MeToo`) and actors reinterpreting their role and using their institution in new ways (Bächtiger and Parkinson, 2019).

3.4 Where do online platforms fit in the larger system of deliberative democracy?

Regardless of whether one adopts a discursive, sequential or spatial perspective, for an online platform to contribute to deliberative democracy optimally it thus needs to produce some deliberative contribution and transfer this contribution in some way to the wider society and its political decision-making bodies in particular. So what kind of contribution should online platforms make to the normative ideal of deliberative democracy? Habermas (2022) ascribes them a similar role to traditional media in his (sequential) approach to deliberative democracy and then criticizes online platforms for failing to live up to those expectations: Platforms lack journalistic moderation and do not “qualitatively filter opinions”, like the traditional media, where journalists scrutinize arguments and opinions for facts and counterarguments and professionally select what to present to their audience (Habermas, 2022). In addition, he is wary of personalization, since it may enable selective exposure and echo chambers. On top of this online platforms have not been very successful empirically in producing good deliberative debates. Where additive approaches to deliberative democracy seek to find sites that facilitate optimal deliberation and then to channel the results as best as possible to other parts of the deliberative system, most major online platforms are known to lack those deliberative qualities (Wessler, 2018; Dahlberg, 2001; Esau et al., 2021).

However, there might be a different contribution that online platforms can make to deliberative democracy, which is more feasible and better suited to their qualities. Habermas (2022) notes that the contribution of political communication in the public sphere, where online platforms are located, is inherently limited since only representative bodies make collective decisions. The normative requirements of achieving deliberation in all its facets need thus not be so strict for these platforms. As the discussion in sections 3.2 and 3.3 has shown, it might not be optimal to strive for deliberation in all its facets on online platforms to begin with. Bächtiger and Parkinson (2019) have recently proposed an alternative route to (macro) deliberative democracy which might better fit with the potential of online platforms: the summative approach. They explain that "the deliberative quality may emerge from the complex interactions of a variety of practices and institutions rather than an input generated by one or two of them" (Bächtiger and Parkinson, 2019, p. 14). Simply put: deliberative outcomes may be realized through non-perfect deliberative components, like online platforms.

We propose that by thinking of deliberative democracy as a summative quality, we arrive at other goals that online social media platforms should perform. They no longer need to facilitate *deliberation* between citizens as best as possible but might focus on optimizing the larger goals of *deliberative democracy*: reaching collective decisions on a rational basis involving as many citizens in the most equal way possible (cf. Thompson, 2008). It needs some mechanism to (1) involve citizens including those acting on behalf of social groups, politicians, PR officials etc.; to (2) make them share their views and information; to (3) facilitate them to interact with each other, existing discourses, and actions and words of political actors to develop and question their opinion; to (4) collect and scrutinize arguments and positions into coherent discourses concerning collective issues; to (5) communicate those discourses back to as many citizens as possible, but (6) also to their representatives in the political arena.[2]

3.5 Moving forward: A complementary role within a summative approach

We propose that in this summative understanding of deliberative democracy, it is more helpful to think of online communication as complementary to other forms of political communication, rather than as a substitute for traditional media. Our approach thus allocates a different function to online platforms within the spatial and sequential system of deliberative democracy than an additive approach would. For example, when facilitating debate between citizens, instead of aiming for civil conversation, it might be better for (macro) deliberative democracy, if in some cases people are allowed some incivility to make suppressed voices heard or to create a communicative environment where some might feel more

[2] Wessler (2018) provides an alternative list of possible contributions of non-deliberative media to deliberative democracy.

at home, where they feel they don't need to be eloquent and highly educated to be allowed to speak up. While at the same time, those who might be put off by such discourse could be shielded from exposure to uncivil content (cf. Fernandez and Bellogin, 2020). Where the one-to-many format of traditional media necessitates compromises in form and content to fit a larger audience at the cost of individual differences, personalization enables online platforms to tune into the individual needs of each citizen (Reuver et al., 2021). Hereby, content can be presented to each citizen in a fashion tailored to encourage involvement, both in engaging with arguments and in building the efficacy needed to share one's views and information (cf. Helberger, 2019)). Social media platforms also provide opportunities to go beyond what can be achieved in deliberative terms by traditional media, by directly linking citizens to, for example, journalists and politicians (Wessler, 2018). Hereby they create a crucial link in facilitating deliberative sub-products, like suppressed voices and new positions and arguments, to reach the traditional media and institutional political arena.

Of course, such personalization is exactly what Habermas (2022) criticizes when he warns about the potential of creating parallel public spheres. When each citizen receives her own tailor-made version of online content, this might hinder a common understanding of the main issues, positions and arguments facing a society. However, from the viewpoint of the complementary role of online communication to other media and institutions, this can also be seen as an opportunity for social media platforms to provide that common information through sharing relevant content as provided by, for example, traditional media, politicians or activist groups, in a tailored way to the largest audience. In this way, online platforms can actually improve the deliberative value of intersubjective understanding by involving previously disaffected groups.

Table 3 provides an example of how this summative approach to designing deliberative online social media platforms differs from existing approaches in the parameters that need to be optimized. The left column ("additive deliberative social media platform") shows how additive approaches seek to optimize all aspects of deliberation at once and facilitate deliberative democracy through concrete instances of citizen deliberation, while the right column shows that the summative approach instead optimizes deliberative outcomes at a societal level. Note how the goals listed in the right column match the mechanisms required for deliberative democracy outlined in section 3.4. The right column focuses on optimizing exposure to foster the deliberative value of inclusion; optimizing engagement to get citizens to interact with the debate; optimizing the sharing of information to include as many insights from as many citizens as possible; providing these insights to other users and *political actors* alike; and including fact-check information to debunk misinformation and increase the factual quality of the arguments. The summative column thus seeks to explicitly and directly link citizens and political actors

(e.g., politicians, but also including activist groups or PR-agencies), since connecting the diversity of arguments leveled by both groups to each other is a specific macro deliberative democratic value. While the additive approach (left column) thus seeks to fit the debate into a deliberative mold, the summative approach (right column) seeks to optimize societal deliberative outcomes.

The summative indicators proposed here are familiar ones in the field of social media systems and partly overlap with both additive indicators and those used in commercial revenue-based applications. They are not meant to form a definitive list. They should rather be seen as an invitation to scholars to propose their own more effective set of indicators to make communication on online platforms contribute to the mechanisms required for deliberative democracy outlined in section 3.4. The overlap with existing commercial applications makes the summative approach more in line with existing practices on online platforms and potentially easier to realize (cf. Heitz et al., 2022). It does not try to change what people like about online platforms, but rather to guide them in a normative, societal deliberative direction. The familiarity of these indicators illustrates the feasibility of this alternative route to realizing deliberative values online.

Table 3: Comparison of Additive and Summative Deliberative Metrics for Social Media Platforms.

	Additive Deliberative Online Platform	**Summative Deliberative Online Platform**
Typical Metrics	– Equality – Diversity – Rationality – Interactivity – Civility	– exposure – engagement (likes/comments) – sharing information – diversity of traditional news exposure – diversity of user and political actor exposure – inclusion of fact-check info where possible
Personalization	Metrics matter for everyone equally	Weight of metrics determined on individual basis
Temporal structure	Static, all metrics are important at every point in time	Metrics can also be optimized for sequentially; long-run outcome more important than simultaneously good scores on every metric
Contribution	Realize deliberation within platform	Contribute to societal deliberation

3.6 Conclusion

Computational approaches to measuring deliberative indicators of online communication are a blooming field and much work has been done in constructing indicators for various aspects of deliberation, such as equality, rationality, interactivity, diversity and civility (Goddard and Gillespie, 2023). However, political scientists have levelled two main critiques against the common computational implementation of debate quality: deliberation is unlikely to be related to deliberative democracy in a unitary, more-is-better fashion and (micro) online deliberation is unlikely to contribute in an additive way to (macro) deliberative democracy. So even if computational scholars could find a way to overcome the current technical challenges and construct a perfect set of reliable and valid indicators of deliberative quality (see Goddard and Gillespie, 2023), then still it would be questionable how these indicators could be implemented in social media platforms to attain normative deliberative outcomes.

We propose that one way out of this dilemma could be to build on the summative approach, which seeks to optimize deliberative outcomes at a societal scale, rather than the additive approach which seeks to optimize deliberation at each site/venue. Instead of a straightforward design of online platforms that either keep indicators of (micro) deliberation within acceptable bounds or optimize for them, we have argued that the complexity of human behavior frustrates those efforts and that it might lead to counterproductive results at a societal level. Instead, we propose a summative approach to designing deliberative online platforms. These systems take more account of the place of online platforms within the larger system of deliberative democracy and respect the potential trade-offs between different deliberative values. They select and optimize an alternative set of indicators directed at macro deliberative goals.

This approach aims to be more fitting to the less than pure deliberative nature of online debate (cf. Wessler, 2018; Dahlberg, 2001). In fact, designing a summative deliberative online platform does not have to be at odds with commercial interests. For example, in the summative deliberative framework, one explicit goal is to increase exposure to and engagement with the “long tail” of content to make users aware of perspectives they may not be aware of, and allow them to contribute. But this can perfectly align with commercial interests: In many recommendations scenarios, it is an explicit goal to increase usage of long-tail items that the user would otherwise not find.

Through its better fit with both deliberative democracy at the societal level and the nature of online platforms, the summative approach proposed here can help online platforms to increase the contribution that online communication can make to deliberative democracy and thereby also help reduce negative effects often associated to online communication, such as filter bubbles, selective exposure and misinformation (Pariser, 2011; Sunstein, 2001; Fernandez and Bellogin, 2020).

4 Evaluation: Testing metric performance

Based on the indicators outlined in Chapter 3 we need to find a good way to measure them in order to analyze the results of experiments using our TWON platform. The following chapter will explore the ability of various methods to validly measure these concepts.

Three members of our team (Sjoerd Stolwijk, Simon Münker and Damian Trilling) have collaborated with other scholars (Mark Boukes, Wang Ngai Yeung, Yufang Liao, Anne C. Kroon) to write a paper titled "Can we Automatedly Measure the Quality of Online Political Discussion?" which is currently under review. The following section is largely based on a slightly adjusted version of that paper. Please note that this chapter does not attempt to review all the latest developments within computational linguistics. However, it rather surveys measures used within substantive (i.e. political communication) empirical research, in which automatic measures are still little used, and when they are, studies often rely on rule-based methods. Since TWON needs to have a broader range of metrics than those discussed in that paper under review, in this report, we have added relevant concepts and results where appropriate.

4.1 How to (Not) Measure Interactivity, Diversity, Rationality, and Incivility in Online Comments to the News

Communication scholars have rapidly and in large numbers made a turn to computational methods. This opened up rich possibilities for large-scale analyses and cost-effectiveness (Boumans and Trilling, 2016). So far, the bulk of applications in communication research are centred around classifying topics, frames, and tone of texts. More complex concepts, like the *democratic quality* of online debates, are notoriously hard to measure, even manually (Bächtiger and Parkinson, 2019). Yet, concepts such as these are central to answering big questions about the role of social media in society. Simultaneously, large language models (LLMs) and generative AI have allegedly improved the performance of automated tools, which suggests that these tools might now also have become feasible options for measuring more complex concepts (see Kroon et al., 2023).

Early research demonstrated problems with accuracy across different automated methods to even detect the tone of texts (González-Bailón and Paltoglou, 2015; Ribeiro et al., 2016; Soroka et al., 2015). Rule-based measures, such as dictionaries, are to date still regularly applied to different contexts than for which they were initially developed; even though it is known that they perform rather poorly in detecting the tone (i.e. valence) of texts in a new context. For other applications than tone, the performance of these, and other, automated measures has received little systematic, comparative attention

so far. Given the demonstrated difficulty in constructing automated measures even for conceptually less complex constructs, such as tone, we wonder: Are computational methods ready for use in substantive, rather than methodological, communication research projects on democratic debate quality? And can generative AI open up new potential to further improve the performance of measurements?

From the many theories of how people ought to communicate in a democratic society with each other, deliberative democracy Habermas (1996) is the most demanding in terms of quality standards (Strömbäck, 2005). Since communication is at the core of this theory, it has inspired a large communication science literature (Schmitt-Beck and Grill, 2020). Habermas' theory makes specific normative claims on what is needed for a discussion to contribute to or cause harm to democracy: Thus, it provides testable criteria for the use of (automated) measurements that aim to determine the quality of online debate.

The current study, therefore, concentrates on the performance of automated measurements in the context of *four concepts* that are central to the "deliberative quality" of political speech (Habermas, 1996); these concepts are also frequently used in studies of other prominent features of online communication, such as polarization, partisan selective exposure, filter bubbles, and echo chambers (cf. Friess and Eilders, 2015): interactivity, diversity, rationality and (in)civility of online comments. The current manuscript provides important methodological guidelines for future research on this increasingly studied phenomenon (see, e.g. Stromer-Galley et al., 2023). In our comparisons, we include more than 50 different metrics varying from simple off-the-shelf dictionaries to various machine-learning approaches, including recent (generative) large language models (LLMs), and compare these against a large set of manually coded data (N = 3,862).

We will first introduce our case in further detail, as well as our selection of automated measures. Thereafter, we will describe our manual and automatic annotation procedures and present the performance of the different automated measures. We will conclude by offering extensive recommendations for further research seeking to use automated measurements.

4.2 Deliberative Quality of Online Debate on Social Media

In the theory of deliberative democracy (Habermas, 1991), debate is about more than exchanging views or collecting issue positions; it also requires active listening and a willingness to understand arguments one initially disagrees with, to learn and find common solutions, or reach a consensus. Such active interactions by the public have long been limited by the modality of the traditional (news) media: i.e., mainly one-to-many mass media with sender and receivers being strictly separated. However, with the increasing accessibility and popularity of social media platforms, "the people formerly known as the

audience" (Rosen, 2012) have been given the assets to actively participate in public debate themselves. Scholars have started to investigate whether citizens participate according to the standards expected in a deliberative democracy (Freelon, 2015; Jaidka et al., 2019; Ksiazek et al., 2015).

We examined the four above-mentioned normative standards based on Habermas' work (Habermas, 1984, 1991, 1996), which have been operationalized by many studies afterwards in various ways (Freelon, 2010; Friess and Eilders, 2015; Janssen and Kies, 2005). First, citizens' expressions should not just be monologues but involve an *interaction* between people who are listening and responding to each other. Second, "to argue is to contradict" (Habermas, 2022, p.152) and, therefore, these interactions should include multiple, alternative viewpoints that do disagree with each other (i.e., *ideological diversity*). Third, the dialogue should be composed of comments that compare arguments through logic and argumentation (i.e. *rationality,* (cf. Bächtiger and Parkinson, 2019, p.21)). Fourth, comments should be provided in a *respectful* manner, usually operationalized as incivility being absent (i.e., *incivility*); uncivil language could distract from the logical strength of substantive arguments or potentially silence the voice of opponents with alternative opinions (Mutz and Reeves, 2005).

4.3 Methodological Challenges of Measuring Deliberative Quality

To understand how well discussions on social media comply with deliberative democracyś normative standards, scholars have mostly relied on human coders to measure deliberative quality (Goovaerts, 2021; Stromer-Galley, 2007). Human coders are still viewed as the most reliable method to measure latent constructs (Baden et al., 2022). Yet, human coding does not scale well to large amounts of data: It is costly, time-intensive, and highly dependent on the motivation and perseverance of coders (Beauchamp, 2020). Specifically for online debates, coders may experience mental distress when being asked to rate large amounts of potentially toxic content. Furthermore, and despite intercoder-reliability tests, substantive and unpredictable individual biases may still prevail, and different sets of coders might arrive at different annotations even when following the same training and codebook (Weber et al., 2018). Yet, human annotations—especially when conducted by experts and with resolved disagreements—are still considered the gold standard in content analysis.

Researchers have attempted to overcome the obstacles of manual content analysis (i.e., expensive, labour-intensive and often unreliable) by developing and proposing automated measures to gauge elements of debate quality. Even though it remains an open question whether these computational methods perform well enough to replace human coding, they are replicable in the sense that the same outcome should emerge when another scholar runs the same script. Some scepticism seems to be warranted, though: For the adjacent task of sentiment analysis, different automated approaches have

been demonstrated to *also* yield widely diverging scores with low predictive validity (Boukes et al., 2020; González-Bailón and Paltoglou, 2015). Results of automated methods should, thus, never be accepted at face value; especially not in a new and complex domain, such as the deliberative quality of online public discussion.

Despite the development of automated tools in measuring various aspects of debate quality, Goddard and Gillespieś (2023) review generally found that existing measurements do lack validity. They recommend rigorously testing them against manual content analysis to verify their performance. This is exactly what we set out to do in this study. To provide researchers interested in studying the quality of online debates with better methodological guidance, we investigated a wide range of automated measurements that capture the degree of interactivity, diversity, rationality, and incivility in online comments. We facilitate future use (and non-use) of these methods for substantive communication research, by comparing the results of each method to the gold standard of manual coding as well as to the other tested methods. Before delving into the details of our methods, we now shortly introduce the automated measures selected for this paper.

4.4 Automated Measurements

We selected the automated measures for this study through a thorough review of the literature for existing specific applications and also included much-used general automated approaches. We group them into different generations of automated measures: rule-based, traditional machine learning, transformer models, and generative AI. Model Groups

Rule-Based Measures. The first generation of automated measures used a rule-based approach, but are currently still popular due to their easy use (e.g. Duncan et al., 2024) and replicability. Rule-based measures usually apply a list of words to verify whether any of these words (or word combinations) is present in a text. They use very little computational resources, are easy to interpret and available in pre-assembled lexicons (i.e., dictionaries and corpora) with various scoring systems (Guo et al., 2016). Moreover, rule-based measures are highly *explainable* and *understandable* (Albaugh et al., 2013); researchers can clearly convey the inner workings of the method and explicate *what* is being counted (especially when compared to for example “deep-learning” models). Rule-based measures have been deployed for common classification tasks, including interactivity (Collins and Nerlich, 2015), ideology (Graham et al., 2009), rationality (Nithyanand et al., 2017b) and incivility (Ksiazek et al., 2015; Nithyanand et al., 2017a).

Counting specific words is very straightforward, but might be too simplistic to understand the meaning of texts since it ignores the different possible interpretations of words in specific contexts. This

makes it difficult for rule-based measures to *both* achieve a satisfying precision and acceptable recall when comparing it to human coding (Atteveldt et al., 2022). Moreover, dictionaries are often derived from other genres or subject contexts than the problem to which they are applied (Boukes et al., 2020; Loughran and Mcdonald, 2011). For instance, online language is more casual and contains novel terms (i.e., slang) that are often missed by dictionaries developed for other types of texts (Vidgen and Derczynski, 2020). Within novel or evolving research domains it is, thus, challenging to find validated dictionaries, and this often leads to a rather low predictive power of such measures; at least, if researchers even undertake the effort to compare it against human-coded data (Baden et al., 2022).

Traditional Machine Learning. With supervised machine learning (SML), researchers are not responsible for crafting a finite list of dictionary terms. Rather, the researcher should first collect a manually-coded (or "annotated") dataset to *train* the algorithm (Boumans and Trilling, 2016). During training, the computer learns to identify the *features* (e.g., words in a social media post) that contribute either positively or negatively to the likelihood of a specific *output* (e.g., the presence/absence of *rationality*). In this sense, the supervised classifier learns the rules for how textual input features relate to the presence (absence) of a specific concept. It uses this information to make future predictions about the presence in new texts.

Transformer Models. Traditional machine learning relies on bag-of-word (BoW) representations of textual data. Here, the algorithms' input features are simple word counts (either with or without some term-weighting, such as "term frequency–inverse document frequency": tf-idf). In the traditional SML approach, word order is lost and models cannot benefit from this contextual information. Neural networks, on the other hand, can take word order into account (e.g., convolutional neural networks). More recently, transformer-based models have largely superseded the traditional SML approaches.

Kroon et al. (2023) summarised the advantages of transformer-based machine learning for communication scholars with two arguments of specific importance: First, in contrast to both rule-based approaches as well as traditional machine learning, transformer-based models excel in taking context into account; second, they (typically) require less training data. Both advantages emerge by levering a large language model (LLM) that is pre-trained on a gigantic text corpus, typically scraped from the web. In the pre-training phase, contextualised embeddings are created; these are numeric representations of the "meaning" of words that even can handle that the same word can have a different meaning in different contexts; for instance, as a swear word being either an insult or a positive signifier in some slang.

Pre-trained LLMs are now readily available for many languages. The researcher may then further improve their performance by fine-tuning these models by using in-context examples with a smaller,

annotated training dataset from the corpus to be analysed. This enables the model to *learn* the details of the specific task. Compared to traditional supervised machine learning with bag-of-words (BoW) representations, these models effectively get a head-start by already "knowing" a lot about language. Moreover, they do not suffer from the unsolvable problems of BoW representations, which cannot distinguish homonyms (words with identical spelling, but different meanings) or take word order into account.

Generative AI. With the new advances in generative large language models (GLLMs), such as the *OpenAI ChatGPT* models, both the size of the pre-training data as well as the complexity of the models in terms of the number of parameters estimated has vastly expanded (Minaee et al., 2024). These GLLMs can also be instructed, which allows a researcher to explain the task to the model directly, rather than by asking it to infer the task from the training examples provided. While this is also possible with natural-language inference models, such as BERT-NLI (Laurer, 2024), the possibility to give such instructions (referred to as "prompting") are a key characteristic of generative LLMs (GLLM). These GLLMs have already shown remarkable performance in data annotation, even at times surpassing human annotation quality (Heseltine and Clemm von Hohenberg, 2024; Törnberg, 2024b).

In the next sections, we present our operationalization of the manual coding (gold standard) and the automated measurements.

4.5 Method

4.6 Data and Sampling

We first conducted a manual content analysis on an original dataset of social media comments in response to TV news items posted on *YouTube* and *Twitter* (currently: *X*). A large dataset of user comments was collected in the fall of 2019. These were comments replying to a wide variety of the most popular U.S. news shows at the time of data collection. We included various genres of TV news to maximize the potential variance in the types of comments and audiences that wrote them; thereby, aiming to enhance their generalizability. Included were nine regular news programs, five partisan news shows from both the left and the right (*MSNBC* and *FoxNews*), and seven satirical news programs (defined by Baym (2005) as "the reinvention of political journalism"). For the similar reason of increasing findings' generalizability, we collected user comments from two different social media platforms: *Twitter* and *YouTube*. Eventually, 3,862 user comments were manually coded. A complete overview of the shows, sampling strategy, and the number of manually coded comments is provided in Appendix A.

4.7 Manual Content Analysis (the Gold Standard)

We constructed a codebook based on the coding instructions found to be reliable in existing research (Freelon, 2015; Papacharissi, 2004; Rossini, 2022; Rowe, 2015a,b; Southern and Harmer, 2021; Ziegele et al., 2020).[3] All codebook details of this study are available in Table 24 in Appendix B. Notwithstanding the reliability in previous studies, the current research underwent a rigorous process of coder training and further validation of codebook items. Coding was done by two student assistants. Six rounds of intensive coder training were conducted to achieve the best possible understanding of the coded variables. Four rounds were sufficient for interactivity, diversity, and incivility; a fifth and sixth round of training were necessary for rationality—clearly, the most complicated construct to be coded reliably by human coders.

Table 24 in Appendix B provides an overview of variables and inter-coder reliability scores in terms of both Krippendorff's α-values and %-agreement. While the overall coding procedure yielded reasonable reliability scores, human coders did not achieve satisfactory agreement in all cases. The lower Krippendorff values for some indicators might be explained by its sensitivity to the skewed variables that dominate our dataset. Unless noted otherwise, the answer options were recoded into a binary variable as "no" (0) or "yes" (1). Coding work began in May 2021 and was completed by January 2022.

Interactivity. To determine whether citizens actively engaged with the views of others, we measured whether they *interact* with each other's comments (Rowe, 2015b). The item tapped whether the substance of a comment referred back to a previous comment or claim of another commenter. Thus, it measured whether someone acknowledged the existence of other comments; and, thereby, showed that a real interaction and exchange of ideas could have taken place.

Diversity. Diversity was operationalized in line with the prior literature on partisan selective exposure, polarization, and echo chambers (Pariser, 2011; Stroud, 2010; Sunstein, 2001). Comments were categorized as having no ideological direction (i.e., absence of political opinion), being of a liberal/Democrat nature, conservative/Republican nature, neutral nature with no clear ideology present (when it attacked/supported both sides), or a rest category of an unclear direction (Freelon, 2015; Rowe, 2015b). We stored this information in two dummy variables indicating whether the comments could be classified as being liberal or not, or as conservative or not.

Rationality. Ryfe (2005) operationalized rationality as positions being substantiated with arguments and empirical evidence. For the current study, three items of existing codebooks were combined to measure the presence of rationality in user comments (Freelon, 2015; Rowe, 2015b; Ziegele et al., 2020). We created one dummy variable indicating whether at least one of the following occurred:

[3] The codebook shows considerable overlap with Friess et al.ś (2021); probably, because it was inspired by the same literature.

(1) the commenter used explicit reasoning and/or argumentation, such as through elaboration on the opinion that was put forward, for example by using the word "because" (Camaj and Santana, 2015); (2) the comment analysed the background of the addressed issue or provided background information; and/or (3) external evidence was provided (e.g., with facts and figures, or with verifiable evidence).

Incivility. Following most existing work, we operationalized "civility" by measuring its opposite (i.e., incivility). Incivility was measured in this study following the codebook of Papacharissi (2004). Two items were added from other studies (Southern and Harmer, 2021; Ziegele et al., 2020): (1) accusing others of incompetence or questioning their intelligence, and (2) suggesting or invoking violence. This resulted in nine items that were used to measure incivility: name-calling, vulgarity, questioning intelligence or competency, shouting, sarcasm, attacking a reputation, threatening individual rights, discrimination, and invoking violence. Again, we created a dummy variable indicating whether at least one element was present.

4.8 Employed Automated Measurements

Rule-based measures are specific to the concept they measure, so these will be discussed per concept. SML, transformer, and generative AI models can all be tailored to a concept, but their model variations are more general, so these will be discussed per modelling approach. To select the best model for each concept for each group of models (i.e. rule-based, SML, transformer, and generative AI), we split the dataset at random into a train (N = 3,089) and a test dataset (N = 773). We used the train dataset to find the best model parameters and select the best model per approach. In our choice for the best model we looked at overall performance, but especially prioritized performance on the harder task of correctly identifying positive cases, i.e. the presence of incivility, interactivity, rationality or diversity rather than their absence. We present the performance of these best models in the results section below for the test dataset: the trained models thus need to predict new data to avoid so-called overfitting. A more detailed technical description of the set-up of each model and a discussion on our choice for the best model per group is provided in Appendix C.

Rule-based Measures: Interactivity. No good text-based dictionary seemed to exist to measure interactivity. We followed existing literature and simply used whether @-mentions were present in comments as proxy (Collins and Nerlich, 2015; Gruzd et al., 2011).

Diversity. It was difficult to find an appropriate dictionary for diversity. The best dictionaries available to measure the partisan nature of comments focus on ideology, especially moral values (e.g. see Zhou et al., 2024). We selected three subsequent versions of the *Moral Foundation Dictionary*: MFD, MFD 2.0, and eMFD. The MFD is designed to measure the ideological positioning of the texts by exam-

ining the (moral) languages used in them. The MFD is both theoretically and empirically related to the partisan nature of the text, although the exact nature of that relationship remains disputed (Graham et al., 2009; Haidt and Graham, 2007; Hopp et al., 2021). MFD 2.0 is an updated version with further enhancement of psychometric properties that should improve the normality and predictive validity of the dictionary (Frimer, 2020; Frimer et al., 2019). The extended Moral Foundations Dictionary (eMFD) is the most recent update, which was developed based on crowd-sourced annotated texts (Hopp et al., 2021). Conservative and liberal values are measured in all MFD versions; respectively, by calculating the ratio of corresponding words indicative of liberal values (fairness, care) and conservative values (authority, loyalty, purity).

Rationality. Various formal text metrics from the field of computational linguistics are available for the concept of rationality. These formula-based metrics are easy to implement and bear similarities to the concept, although it was more difficult to find a good dictionary-only measure. We selected the Flesch-Kincaid (FK) index (Flesch, 1948) and language formality (Heylighen and Dewaele, 2002) metrics to measure language complexity and formality of comments (e.g. Nithyanand et al., 2017b). Another index used to measure rationality is the Integrative Complexity (IC) score (Owens and Wedeking, 2011). Unlike the FK score, the IC score attempts to measure the semantic complexity of texts through a formula based on the use of words that belong to the cognitive complexity dimension in the LIWC dictionary (Pennebaker et al., 2007).

Incivility. Multiple dictionaries have been developed to measure the construct of incivility. We identified six different dictionaries to be tested for this manuscript (see Appendix C.1). These dictionaries include the (1) Ksiazek et al.'s (2015) Hostility dictionary and (2) Ksiazek et al.'s Civility dictionary (*reverse-coded*), (3) the Incivility dictionary developed by Muddiman and Stroud (2017), (4) the LIWC-22 (Boyd et al., 2022), (5) *Google* What Do You Love Project (WDYL) Censored wordlist (available at (Dubs)), and (6) the Hatebase wordlist constructed by Hatebase.org (Quinn, 2020). All these dictionaries are recoded as dummy variables (0 or 1), to maximize comparability with the hand-coded data: We defined that a comment shows incivility (score 1) if at least one uncivil word appears.

Traditional Supervised Machine Learning (SML). The traditional SML approaches use bag-of-words representations, which can either be count-based or tf-idf-based. Eight models were estimated for each variable: two vectorizers (Count and tf-idf) × four classifiers (Multinomial Naïve Bayes, Logistic Regression, a support vector machine classifier (SVC) with a radial ("rbf") kernel, and SVC with linear kernel). When training the classifiers towards the best model performance, Each model was further optimized by modifying (1) the number of words considered when tokenizing a sentence; (2) the range of word frequency, and (3) the standard for regularization that aims to avoid over-fitting in the classifier. For

each of the four tested concepts, we then chose the model-variant with the best performance in terms of macro F1 scores and minority class F1 scores in the test dataset for presentation in the results section. See Appendix C.2 for more details on the procedure and individual model performance.

Transformer Model. In addition to training classifiers using the above-mentioned traditional SML techniques, we explore the potential of using transformer-based models in our classification pipeline by using Pythonś *PyTorch* library. Here, we used the uncased version of the English-language BERT model (bert-base-uncased) and fine-tuned it for our classification tasks using the training set from our manually annotated data. During this fine-tuning process, the modelś parameters are updated to better suit the specific task at hand. We again selected the best model parameters based on macro F1 scores and minority class F1 scores.

Generative AI. There are many different GLLMs currently available. Törnberg (2024a) recommends using an open-source model where possible. Other, so-called proprietary, models (e.g., ChatGPT) do not share their weights and are secretive about the training data they used to build their models. Moreover, they require a researcher to share their data with OpenAI by sending it to their API (OpenAI, 2024b,c). Yet, OpenAI promises not to use this data as long as you use the paid API and opt out of data-sharing.

Nevertheless, we believe that data security is better preserved — in terms of the privacy and copyright of analyzed materials —by using a model that can be run in-house without exchanging data with a third party; this is also in line with the spirit of the GDPR. Being free of charge is an additional benefit. Therefore, we evaluate the opportunity costs of opting for an open-source model like Llama instead of the popular GPT models owned by OpenAI. Microsoft offers a paid service on their Azure platform guaranteeing that no data is shared with OpenAI (Microsoft, 2024), although data still needs to be exchanged with Azure to use this option.

Since running GLLMs is computationally heavy, we focused on two specific model families for this project: OpenAIś GPT and Metaś Llama. We chose state-of-the-art variants of these models. For Meta, we used their latest large language model, Llama3.1; more specifically, we used the llama3.1:70b-instruct-q6_K-variant (hereafter Llama3.1:70b). For comparison, we also used the smaller version llama3.1:8b-instruct-q6_K (hereafter Llama3.1:8b). From OpenAI, we used the two most recent and advanced models available through the Azure OpenAI API: GPT-4o (the one released on 2024-08-06) and GPT4-Turbo (the one released on 2024-04-09). Appendix C.4 presents details of the model setup and prompt wording.

We compared the effect of using different prompts. Since running classifications on (large) generative AI models is computationally expensive, and for OpenAI models also financially costly, we first tested the effects of different prompts in Llama3.1:70b. For the instruction (or: prompt) given to Llama3.1,

we first followed the codebook nearly verbatim, often only adding small label specifications (i.e. “not present *(1)*”) to help the model classify the data in the correct classes. For some items, the wording of the codebook appeared to confuse the model, which resulted in high numbers of missing values. In these cases, we asked OpenAIś GPT-4o to reformulate the prompt to make it better interpretable for GLLMs. All prompts were checked manually to contain the same information and examples as the codebook; thus, keeping the information constant for GLLMs versus the human coders. The only changes were in the structuring and wording of the instructions. This procedure resulted in a large number of long prompts since incivility and rationality were measured by multiple indicators.

However, long prompts are often not optimal for GLLM performance and running them is computationally costly since you need to process many runs (one for each prompt-and-comment combination) of many tokens (i.e. many words in each prompt) (Törnberg, 2024a). Therefore we also considered a simpler approach. This simpler approach would be the most likely one a researcher without our codebook would use: short concise single prompts per latent concept (interactivity, diversity, rationality and incivility) rather than one for each specific indicator of the latent concept.

Since the simple prompts consumed fewer tokens and also produced the best results overall (see Appendix D), these simple prompts were used to test the effect of model size and family. Therefore, we ran these simple prompts with GPT-4o, GPT4-Turbo and the smaller Llama3.1:8b in addition to the Llama3.1:70b. Smaller models have smaller hardware requirements in terms of GPU memory, making them more accessible to use for many researchers. However, a reduction in model size often comes at a cost in performance (AI, 2024), and indeed Appendix DD shows this is also the case for our application. As further explained in Appendix C C.4 we selected Llama3.1:70b as the best-performing GLLM based on a combination of macro F1 and minority class F1 performance, financial and ethical considerations. Appendix D lists the classification results for these different prompts and models on the training set. Note that our main results hold regardless of whether we had selected the best model per group on macro F1 or minority class F1.

4.9 Results

4.10 Performance Indicators and First Impression

To evaluate the performance of automated measurements against manually coded data, we calculate precision, recall, accuracy, weighted F1, and F1 macro scores. Precision indicates the method's reliability in correctly *identifying* positive results, while recall measures its sensitivity in *finding all* positive results. Accuracy represents the proportion of correctly classified cases. This is misleading for highly

imbalanced samples, which is why the F1 macro average score (F1 macro: the harmonic mean of precision F1 and recall F1) is often preferred. F1 macro provides an average of two classes (i.e. the concept being present or not present) without considering proportions, while weighted F1 considers proportions (i.e. when one class is overrepresented in the data it also carries extra weight in the F1 calculation).

In the context of debate quality metrics, we prioritize the classification of minority classes (for instance, most comments are *not* uncivil; and yet, we are especially concerned about the uncivil comments). Although we prioritized minority class F1 in the selection of the best measures per model group even further, for reasons of parsimony, we consider F1 macro as the most appropriate measure for the comparison between the best models per model group presented here.

To get a good first impression of the relative performance of each group of measures, we present the F1 macro for the best-performing model variants per concept per group based on the data in the test set (N = 773) in Table 4. The best scores per concept are printed in bold. Rule-based measures overall perform poorly, except for the incivility concept. Supervised machine learning performs better, but performance is still limited. The BERT-transformer models and GLLM achieve the best results. Overall, the performance of the GLLM (Llama3.1:70b) classifiers appears very reasonable and provides a good out-of-the-box option for coding social media comments; they even beat the finetuned BERT-transformer models for two (diversity and incivility) of our four concepts.

Table 4: Overview of difference in macro average F1 macro scores for the best rule-based, SML, Transformer and GLLM models on the test set (N = 773).

	Best rule-based	**Best traditional SML**	**Best transformer**	**Best GLLM**
Interactivity	.55	.62	**.75**	.66
Diversity	.52	.62	.57	**.77**
- Liberal				
- Conservative	.53	.51	.61	**.81**
Rationality	.51	.67	**.72**	.64
Incivility	.67	.66	.73	**.75**

Note. Diversity is split into liberal and conservative to account for the non-binary nature of our operationalization, which includes a neutral category. Bold markings indicate the best F1 macro score per concept.

4.11 Performance of Different Automated Approaches

We will now list the more detailed results per concept. Again, we use the best-performing variant per measurements group (rule-based, SML, transformer, GLLM). All models are evaluated against the test set, to avoid overfitting on the training data.

Interactivity. Neither @-mentions nor SML showed promising results for predicting interactivity (5). With an F1 score of .32, using @-mentions as a proxy for interactivity is not recommended. Llama3.1 and SML perform better and have an acceptable recall for interactivity being present: The models correctly identify 80% (SML, recall = 0.80) and 69% (Llama, recall = 0.69) of all interactive messages. Yet, the low precision (SML: .42; Llama: .46) illustrates that, among the comments classified as interactive by these models, the majority are not interactive according to the manual annotations. A transformer-based classifier – while also not impressive – achieves better and more consistent results across categories. We find that for interactivity, finetuning a BERT transformer still yields the best results.

Table 5: Accuracy, Precision, Recall and F1 scores of interactivity measures against the manually coded interactivity on the test set.

	Precision	Recall	F1 score	N
Rule-based (Mentions of @)				
0 (Non-Interactive)	.75	.83	.79	559
1 (Interactive)	.38	.28	.32	214
Accuracy			.72	
Macro average	.57	.55	.55	773
SML (Logistic Regression)				
0 (Non-Interactive)	.88	.57	.69	559
1 (Interactive)	.42	.80	.55	214
Accuracy			.67	
Macro average	.65	.69	.62	773
Transformer				
0 (Non-Interactive)	.87	.84	.86	559
1 (Interactive)	.62	.68	.65	214
Accuracy			.76	
Macro average	.74	.76	.75	773
GLLM (Llama3:1.70b)				
0 (No)	.85	.70	.77	559
1 (Yes)	.46	.69	.55	214
Accuracy			.72	

Table 5: (continued)

	Precision	Recall	F1 score	N
Macro average	.66	.69	.66	773

Diversity. The best-performing models for all diversity measures are presented in tables 6 and 7 (see Appendix E for full results of all rule-based measures). Tables 6 and 7 show that the main reason for Llamaś superior overall performance (see Table 4) for diversity compared to the other methods is its much stronger ability to correctly identify positive cases (i.e. comment is conservative, or comment is liberal). Still, in absolute numbers, its precision scores of 0.54 (is Liberal) and 0.60 (is Conservative) show that automated methods still struggle in this regard: A little less than half of the comments labelled by Llama as liberal/conservative did not receive this label from our manual coders. The BERT-transformer model shows disappointing performance for diversity and is even outperformed by SML for Liberal/non-liberal (macro F1: 0.62 vs 0.57).

Table 6: Accuracy, Precision, Recall and F1 score of diversity measures against manually coded diversity (Liberal) on the test set.

	Precision	Recall	F1 score	N
Rule-based (MFD 2.0)				
0 (Non-liberal)	.83	.71	.77	633
1 (Liberal)	.22	.36	.27	140
Accuracy			.65	
Macro average	.52	.53	.52	773
SML (Logistic Regression)				
0 (Non-liberal)	.90	.71	.80	633
1 (Liberal)	.33	.66	.44	140
Accuracy			.70	
Macro average	.62	.68	.62	773
Transformer				
0 (Non-liberal)	.92	.59	.72	633
1 (Liberal)	.30	.78	.43	140
Accuracy			.62	
Macro average	.61	.68	.57	773
GLLM (Llama3.1:70b)				
0 (Non-liberal)	.95	.85	.90	633
1 (Liberal)	.54	.81	.65	140

Table 6: (continued)

	Precision	Recall	F1 score	N
Accuracy			.84	
Macro average	.75	.83	.77	773

Table 7: Accuracy, Precision, Recall and F1 score of diversity measures against manually coded diversity (Conservative) on the test set.

	Precision	Recall	F1 score	N
Rule-based (MFD 2.0)				
0 (Non-conservative)	.88	.66	.75	660
1 (Conservative)	.19	.49	.28	113
Accuracy			.63	
Macro average	.54	.57	.51	773
SML (Logistic Regression)				
0 (Non-conservative)	.93	.54	.68	660
1 (Conservative)	.22	.75	.34	113
Accuracy			.57	
Macro average	.57	.65	.51	773
Transformer				
0 (Non-conservative)	.89	.84	.87	660
1 (Conservative)	.31	.41	.35	113
Accuracy			.78	
Macro average	.60	.63	.61	773
GLLM (Llama3.1:70b)				
0 (Non-conservative)	.96	.91	.93	660
1 (Conservative)	.60	.80	.68	113
Accuracy			.89	
Macro average	.78	.85	.81	773

Rationality. Similar to the results of interactivity, the transformer model has the highest macro average F1 score (0.72) for the concept of rationality. In contrast to the other three concepts, Llama performed a little poorer for rationality on the test set than on the training set (F1 macro dropped from 0.69 to 0.64, cf. Table 29 in Appendix D) and was even outperformed by the best SML (SVC 'rbf'). Table 8 shows Llama did slightly better in terms of precision (0.56 vs. 0.54) of positive cases (that a comment is rational), but at the expense of recall (0.30 vs. 0.56). In line with the results for the other concepts,

correctly classifying the positive cases remains a challenge for all models. The performance of rule-based measures was again poor for rationality.

Table 8: Accuracy, Precision, Recall and F1 score of rationality measures against manually coded rationality on the test set.

	Precision	Recall	F1 score	N
Rule-based (FK-score)				
0 (Not rational)	.84	.59	.69	624
1 (Rational)	.24	.53	.33	149
Accuracy			.58	
Macro average	.54	.56	.51	773
SML (SVC "rbf")				
0 (Not rational)	.91	.78	.83	624
1 (Rational)	.41	.66	.51	149
Accuracy			.75	
Macro average	.66	.72	.67	773
Transformer				
0 (Not rational)	.89	.88	.89	624
1 (Rational)	.54	.56	.55	149
Accuracy			.82	
Macro average	.72	.72	.72	773
GLLM (Llama3.1:70b)				
0 (Not rational)	.85	.94	.89	624
1 (Rational)	.56	.30	.39	149
Accuracy			.82	
Macro average	.70	.62	.64	773

Incivility. Table 9 shows that for incivility, the best-performing rule-based measure is Ksiazekś (2015) Hostility dictionary. Rule-based measures perform much better for the concept of incivility, with a macro F1 of 0.67, than for any of the other tested concepts where they only get to a macro F1 of respectively 0.55 for interactivity, 0.52/0.51 for diversity and 0.51 for rationality. Ksiazekś rule-based measurement performed also relatively well (i.e., the performance gap is smaller) compared to other types of classifiers: for incivility, its F1 macro of 0.67 was lacking only 0.08 points behind the macro F1 score of Llama3.1:70b (0.75) and even surpassed that of the best SML (0.66).

The main strength of the dictionary approach was the high precision for the positive class (uncivil: 0.76), higher even than both the BERT-transformer model (0.72) and Llama3.1 (0.69). This means that

if this dictionary classified a comment as uncivil, manual coders very often agreed. This makes sense because incivility often results from using a specific profanity, i.e. a specific word, which can be listed. The problem with dictionaries is the difficulty of listing *all* uncivil words while avoiding words which are either civil or uncivil depending on context. This results in a lower recall of positive cases (uncivil comments: 0.49). Llamaś better ability to take context into account might explain why this same recall of positive cases is the main strength of Llama3.1 (0.88). Overall, the differences between different measure groups are much smaller for incivility than for interactivity, diversity or rationality. Yet, the BERT-transformer model and Llama3.1 still outperform the other classifiers.

Table 9: Accuracy, Precision, Recall and F1 score of incivility measures against manually coded incivility on the test set.

	Precision	Recall	F1 score	N
Rule-based (Ksiazekś hostility dictionary)				
0 (Civil)	.65	.86	.74	408
1 (Uncivil)	.76	.49	.59	365
Accuracy			.68	
Macro average	.71	.67	.67	773
SML (SVC "rbf")				
0 (Civil)	.74	.55	.63	408
1 (Uncivil)	.61	.79	.69	365
Accuracy			.66	
Macro average	.67	.67	.66	773
Transformer				
0 (Civil)	.74	.75	.75	408
1 (Uncivil)	.72	.70	.71	365
Accuracy			.73	
Macro average	.73	.73	.73	773
GLLM (Llama3.1:70b)				
0 (Civil)	.86	.64	.74	408
1 (Uncivil)	.69	.88	.77	365
Accuracy			.76	
Macro average	.77	.76	.75	773

4.12 Discussion and Practical Recommendations

To enable future scholars to draw on computational methods to also measure complex, but important concepts, such as online democratic debate quality, we collected and evaluated a broad range of automated tools on the ability of their classifications to replicate the results of a manual content analysis. We first draw general conclusions from our results, before we present our practical recommendations for future studies on the use of automatic measures, as well as specific recommendations per model group.

Interpreting the results

Despite the wide variety of automatic measures employed and their increasing technological sophistication, it proved difficult to fully replicate manual annotations. The limits of our automatic methods are particularly illustrated by the macro F1 scores for the positive class (i.e. concept is present), which stall at around 0.65 for most concepts. Nevertheless, the average macro F1 scores we report for the best models vary from 0.72 to 0.81.

Compared to state-of-the-art approaches in the field of artificial intelligence, this appears to be a meagre result, with for example reported F1 scores of over 0.90 in the field of hate speech detection (Agrawal and Awekar, 2018; Badjatiya et al., 2017). However, recent work by Arango et al. (2022) suggests that such impressive results are only possible due to various forms of overfitting on some widely used public benchmark datasets (also see Alzahrani et al., 2024). When they correct for the biases they identified both in the calculation of the metrics and the datasets used to evaluate them, the macro F1 of the state-of-the-art models drops to about 0.78: This very much resonates with our results.

The main critique of Arango et al. (2022) on the use of public, so-called benchmark, datasets is the lack of information they provide on sampling methods, coder distribution and quality. We avoid such biases in our study, by having created an original manually coded dataset specifically for this study, for which we followed all standard steps typical in the field of communication science. From that perspective, our performance scores are quite impressive given that we used sophisticated but standard, out-of-the-box and relatively easy-to-implement algorithms that are accessible for general use by computational communication scholars. This is in contrast to the highly specialized computer science approaches, which require advanced computational skills to apply.

In addition, our results outperform those reported in a recent review of sentiment analysis methods by Van Atteveldt et al. (2021), which was also based on a comparison to a high-quality original manual coded dataset. This is surprising given that the field of sentiment analysis is much more developed in computational linguistics, especially compared to concepts like rationality for which hardly any automated measures were available when we started our review. The deep learning classifiers reported by

Van Atteveldt et al. (2021) do not get beyond an F1 of 0.66 even for the easier neutral sentiment category and dictionary measures don't reach above 0.58 for that same category, while we report majority-class F1 scores up to 0.93 (Llama3.1: on comment "not Conservative") for GLLMs, and up to 0.83 for traditional SML (on comment "not Interactive"). This seems to indicate that recent advances in the domain of text classification, which this paper draws on, particularly (G)LLMs, have significantly improved the ability to classify constructs central to the field of communication science in just a few years.

Practical Recommendations

Which measure should be preferred for future empirical applications is likely dependent on the objective of the study in question. We will highlight some considerations for designing future studies in turn, starting with the choice between manual or automatic annotations followed by specific considerations per group of automatic measures.

Manual versus automatic measures. The inability of automatic measures to fully replicate manual coding might suggest that where possible manual coding should be preferred over automatic measures. However, the intercoder reliability of our manual coding proved to be contested as well for some concepts. We struggled to achieve acceptable reliability levels and depending on the concept needed up to 6 rounds of coder training to do so. Even though we reached acceptable reliability levels for most concepts, the complexity of our concepts in combination with their skewed distribution in most samples limits the ability of human coders to reach perfect reliability scores. The Krippendorf α-values for our coded variables frequently dropped below 0.60 and percent agreement often stalled just upwards of 80% (see Appendix B).

Automatic coding measures are, by nature, more consistent than any group of manual coder can be; different coders involved in manual coding will always slightly differ in their annotations from each other for abstract concepts. Therefore, a manually coded sample likely has higher degrees of coincidental inconsistencies than an automatically coded sample. Accordingly, any discrepancy between an automatic measure and our manual coding sample, might thus also be due to such inconsistencies in the manual sample. In any case, the differences between different coders in our manual content analysis might limit the performance of supervised automated methods, since the algorithms do not know which coder annotated which comment in either the training or the test set. The manual annotation thus follows two similar, but different internal logics in arriving at a particular code per comment depending on the coder. By contrast, the automatic methods can only draw on one internal logic and therefore will necessarily be different to at least one of the coders in this respect. This makes it very difficult for automatic models to perfectly replicate such annotations. If this fallibility of human annotations is taken into consideration, the performance gap between human annotations and those generated by

the newer transformer and Llama models appears to be limited.

Furthermore, we repeatedly found that precision on the positive class (i.e. comments automatically identified as interactive/diverse/rational/uncivil are also manually coded that way) is the key limitation of the performance of these automatic measures, while their recall is generally (much) better. The advantage of low precision but high recall on the minority class (in this case the positive class) is that the error is contained in a well-defined, but limited set of comments, namely those predicted by the model as positive for that variable. The high recall means that this set will very likely contain most cases that are actually positive, although it will also contain many negative cases due to the low precision. To reduce the error researchers may manually code this much smaller set of comments, which is often much more feasible than manually coding the entire corpus. This is very different from the alternative scenario of low recall and high precision in the minority class, since in that case, the missing positive cases are (much) more spread out and must be found in the larger majority class. We, therefore, suggest that scholars who use (G)LLMs to classify concepts with high recall and low precision for the minority class consider manually annotating the positively predicted sample (cf. Heseltine and Clemm von Hohenberg, 2024; Van Hoof et al., 2024). In this way, near-human annotator agreement can be attained even for large datasets that require this additional precision.

Altogether, we reiterate previous calls to always validate the chosen automated method with a human comparison rather than just applying them "off-the-shelf" (Boukes et al., 2020; Kroon et al., 2022; Van Atteveldt et al., 2021), and extend this approach to also include GLLM models, such as Llama or ChatGPT. For example, Appendix E shows a lot of variance among different rule-based measures. Just picking one without validating its performance for the dataset in question can thus be problematic, but the same is true for very advanced GLLMs. Appendix D similarly shows that GLLMś performance may vary considerably depending on the tested concept, the model, model size and prompt wording (also see Salinas and Morstatter, 2024).

Therefore, we emphasize that it is crucial to also test the performance of so-called 'zero-shot' (i.e. without training data) classifiers, such as dictionaries or GLLMs, on a training set (or 'validation/selection' set), then select the best-performing model, and always report performance on a test set to avoid overfitting on the test data and thereby inflating the estimated classification accuracy. Random variations in the data used to compare different models (variations) might give one model an edge over other models, which will not replicate in the larger dataset of interest for the research project. When comparing the performance of many model variations, as might likely be the case during prompt engineering of GLLMs, a researcher should make sure that random prompt wording variations did not inflate the performance of a GLLM classifier, simply because they happened to fit the particular characteristics of

the evaluation set. Splitting this evaluation set into a training set for selection and a test set for performance evaluation is an established way to deal with this issue of overfitting (Atteveldt et al., 2022).

Since manual coding is thus advised for any content analysis using latent concepts, including those using automated measures, it might for now remain the go-to approach for small to intermediate-size samples. On the other hand, our results suggest much potential for automated methods, especially for larger samples. We will consider the consequences for rule-based measurements, supervised machine learning, and generative AI methods in turn.

Rule-based measures. A key advantage of using rule-based measures over other automatic approaches is their ease of use, both in terms of computational skills needed and in computational cost in processing time and IT infrastructure requirements; moreover, their results are easy to explain since they can be traced back exactly to the content of a dictionary (Kroon et al., 2022). However, rule-based measures were found to be worse in replicating manual coding than all other classifiers and their performance varied markedly between different dictionaries for the same concept. Therefore, if a researcher decides to use a rule-based measure it is important to select a sufficiently good one. Since even the best ones did *not* meet acceptable performance for our concepts, it might be necessary to follow various approaches to adapt or self-develop a dictionary (cf. Bodrunova et al., 2019; Bos and Minihold, 2022; Muddiman and Stroud, 2017).

At the same time, given that this, if done well, has to include multiple rounds of testing, including hand-coding test cases, this may be much more work than it seems at first. The combination of inferior performance, the need to hand-code a sufficient sample of data for both the selection and the improvement of rule-based measures, as well as the time and effort needed to do all this appears to outweigh the advantages of rule-based measures in ease and cost of use for most research purposes. We, therefore, cannot recommend them, especially those rule-based measures which did *not* reach the performance level of the best rule-based measures, like the LIWC-22 dictionary for incivility (F1 average = 0.52, F1 [presence of incivility] = 0.32, see Appendix E), even though they are still recently advocated for by others (e.g., Duncan et al., 2024)).

Supervised machine learning. SML and transformer models outperformed rule-based measures. Contrary to rule-based measures, Appendix C.2 shows that performance differences between most different SML models were limited. Still, there were exceptions; therefore, researchers should compare at least two or three different models to ensure optimal performance. If the sample is large, we believe supervised machine learning, especially classifiers using transformers, should be preferred over rule-based measures. Since a manually coded sample is needed anyway, why not use that to train or fine-tune a good classifier? A recent review by Kroon et al. (2023) gives a good introduction to the pros,

cons and alternatives of these advanced methods currently available.

Platforms, such as *Hugging Face*, provide tools and resources that help to make implementing these models less complicated and more accessible for a broader range of communication scientists. Still, transformer models in particular require computational expertise and access to a computer or server that is capable enough to support the fine-tuning process. The analyses for this paper, for example, took several days to run. Notwithstanding these difficulties, the performance of the transformer models was impressive for non-standard and rather abstract concepts, such as interactivity and rationality; thereby, even performing better than the much more advanced GLLMs by a considerable margin. The fine-tuning process with thousands of training codings appears to be especially helpful when the definition of a concept is research-specific.

Generative AI. Although Llama3.1:70b did not achieve the best results on all concepts, overall the performance of Llama3.1 shows the potential for GLLM models. The model used here already performs well across concepts; furthermore, faster and still-improving Llama models are being released very regularly. In contrast to transformer models that performed best on the abstract concepts of interactivity and rationality, the GLLMs tested here excelled at the measurement of the more frequently operationalized concepts of diversity (ideology) and incivility (see Appendix D). This makes some sense, intuitively, since many public training sets of these concepts are likely to be part of its training data. For such concepts, we therefore advise using GLLMs.

When it comes to the choice of GLLM, researchers should balance their data ethics considerations versus the ease of use. The Llama3.1:70b model used in this paper is available open source and free of charge, but it needs an IT infrastructure beyond what is currently, typically available on an average laptop. Using the smaller, less demanding, Llama3.1:8b-model significantly reduced the performance, however. The OpenAI GPT models, on the other hand, can be run from any laptop using an (Azure) API. Running a GLLM locally is not as difficult as sometimes assumed, but still requires more programming skills than only using an API (Grüber, 2024a,b).

Another clearcut advantage of using GLLMs, such as Llama, is that its performance was quite good compared to other models, but it did not require manually coded training data to fine-tune or train it. Manual coding is only needed for the validation of its performance. The simple prompts without example annotations even outperformed a combined measure based on the collection of elaborate prompts, which contained all information for each indicator in the codebook per concept. GLLM models, thus, were demonstrated to offer a cost-accessible option for automatic coding even of sophisticated concepts, such as rationality; as long as the research objective does not require the highest levels of annotation accuracy.

However, caution should be observed when using GLLMs to classify our concepts. The details presented in Appendix D show an even larger variance in precision, recall and F1 for the different positive and negative predicted classes among GLLMs than observed for rule-based measures. For example, the recall for the presence of rationality ranged from 0.14 to 0.93 (for simple prompts in GPT4-Turbo and the combined measure based on verbatim codebook prompts in Llama3.1:70b, respectively). Therefore, we recommend reporting full classification performance results when using a GLLM. If no further manual annotation is added, we also recommend prioritising minority class performance over majority class, since our results show this is the more challenging case for GLLMs. The more evenly a GLLM performs across classes, and across precision and recall, the more likely its results substantially match those of human coders and lead to replicable results across studies using different methods or models.

Comparing the performance of different GLLMs yielded surprising results. Although OpenAI claims both GPT4o and GPT4-Turbo beat open-source models, such as Llama3.1:70b, across industry benchmarks (OpenAI, 2024a), our findings demonstrate that Llama3.1:70b outperforms each of them in terms of macro F1 on rationality and interactivity. Part of this might be due to the lack of prompt engineering to tailor prompts specifically towards the best performance for the OpenAI models because the prompt effect was only explored and tested on Llama3.1:70b here. However, recent work suggests that benchmark performance of GLLMs might be inflated, since these benchmark tasks might be part of the training data (Dong et al., 2024; Mirzadeh et al., 2024). This opens up the possibility that the performance of different GLLMs might vary depending on the task, especially on newly annotated data and concepts that are less common in public datasets.

Therefore, we recommend to not blindly use the most recent and popular GLLM that is available; but rather to evaluate their respective performance against a freshly coded dataset. Also, the strong performance of Llama3.1:70b shows that open-source models can be competitive with proprietary models like those from OpenAI. At least for the four concepts tested in this study, a choice for better data security could go hand in hand with optimal performance.

4.13 Conclusion

We evaluated a large number of automatic measures compared to an original, manually coded dataset of user comments that replied to a wide diversity of news shows; ranging from satirical to hard news on two different online platforms. We provided a challenging case for our automatic measures since word choice and language style are likely dependent on the genre and platform, which would benefit from a human interpretation. By comparing rule-based measures, supervised machine-learning classifiers (traditional and fine-tuned transformer) and modern generative AI, this study demonstrates the poten-

tial of using automatic methods to measure more complex, but important constructs that are central to the democratic quality of online debate: interaction, diversity, rationality and (in)civility. Overall, the modern transformer-based models and GLLMs outperformed the older methods.

Even though all automatic measures included in this study struggled to fully replicate manual coding results, we believe the more recent approaches did a decent job considering the complexity of the tested concepts. Given that computational communication is a fast-moving field and new methods continuously become available, this shows great potential for the application of such measures to support theory-driven communication research. But our results also give a clear warning to not blindly apply what others have used: Despite being widely used, some approaches and implementations performed so poorly that they are not suitable for studying debate quality.

5 Translating findings to final metrics

So far, we have discussed what indicators we aim to measure for the quality of online social media debates and the performance of various classifiers for some key concepts. However, the TWON project has specific needs that warrant a tailored set of indicators. This section will first lay out which indicators are most suitable for the TWON project and then discuss the performance of some additional indicators that we feel are useful in the TWON context. It will also discuss how the proposed metrics can be applied to research questions at different levels of debate, i.e., at the user, thread, topic, or platform levels.

5.1 Selecting metrics for summative deliberation

Building on Chapter 3 we would like to pursue a summative route to measuring debate quality for TWON. To do so we need to fit specific metrics for each indicator described in the summative column of Table 3. The example indicators listed there are:

- exposure
- engagement (likes/comments)
- sharing information
- diversity of traditional news exposure
- diversity of user and political actor exposure
- inclusion of fact-check info where possible

In Chapter 3 we explain how this initial list of indicators must be fitted to the research project and might be further refined. For TWON we propose to group them into exposure, engagement, participation, diversity of exposure and quality of exposure. While mostly similar this list deviates in some respects

from that of Table 3. First, we now group the diversity of news, user and political actor exposure into one diversity category. This is mainly a simplification based on the currently foreseeable TWON studies that will focus on general public discussions among general participants centred around system-fed news articles [see first and second field study]. Second, we replaced the suggested quality indicator of fact-checking with the more encompassing quality of exposure. This is because we are not just interested in debating on factually correct information, but in initiating a proper debate in the sense of exchanging arguments more broadly. Especially, since previous research (Wessler, 2018; Dahlberg, 2001; Esau et al., 2021) showed that such debate is often lacking on social media, the presence of elements of reasoning (described as 'rationality' in Chapter 4) can be seen as a signal of some quality within a comment. However, if any TWON or related studies would choose or find it appropriate to add metrics for fact-checking/misinformation, or to calculate individual diversity metrics for exposure to political actors or news content we would highly support that. The proposed diversity metric is suitable for such applications.

In addition to simplifying some indicators, we also add a new one: whether content is political or not. This is because the TWON project is specifically interested in political debate. Therefore it makes sense to measure whether a participant is not just exposed to, engaged with or shares content, but also whether and to what degree this content is political. We also define diversity in terms of ideological diversity, in line with the measurements discussed in Chapter 4 to reflect this political dimension. However, as will be proposed in Chapter 6 full deliberation requires more than the balanced inclusion of ideological perspectives but rather needs a variety of arguments related to an issue of common concern (see Chapter 3. We currently do not have such a metric that is also adequately validated for operational use, but we will present some initial tests in Chapter 6. Finally, we need to propose which classifiers to use for political content, diversity/ideology and quality/rationality. The following paragraph will discuss these issues.

5.2 Classifying content as political

We thus need a classifier to determine whether content is political or not. We tested the performance of a variety of prompts in both Meta's Llama3.1:8b, Llama3.1:70b and OpenAI's GPT4, GPT4-Turbo and GPT4o on the German X-data collected and manually annotated by Heseltine and Clemm von Hohenberg (2024) (N = 700). We used the same model specifications and setting as described in Appendix C.4 except for the Llama3.1 temperature setting which we further reduced from 0.1 to 0, results showed this had no measurable consequence on performance. We used the same prompt as Heseltine and Clemm von Hohenberg (2024). Table 10 shows the performance of these classifiers.

Table 10: Precision, Recall, F1-score, and N for classifying a post as political or not using different models.

	Precision	Recall	F1 score	N
Llama3.1:8b				
0 (No)	.81	.83	.82	202
1 (Yes)	.93	.92	.93	489
Accuracy			.89	691
Macro average	.87	.87	.87	691
Llama3.1:70b				
0 (No)	.86	.78	.82	202
1 (Yes)	.91	.95	.93	498
Accuracy			.90	700
Macro average	.89	.86	.87	700
GPT4				
0 (No)	.66	.97	.79	202
1 (Yes)	.98	.80	.88	498
Accuracy			.85	700
Macro average	.82	.88	.83	700
GPT4-Turbo				
0 (No)	.66	.96	.78	202
1 (Yes)	.98	.80	.88	498
Accuracy			.85	700
Macro average	.82	.88	.83	700
GPT4o				
0 (No)	.60	.99	.75	202
1 (Yes)	.99	.73	.84	498
Accuracy			.81	700
Macro average	.80	.86	.80	700

The table shows how performance is quite good across all these models, but perhaps surprisingly, similar to rationality and interactivity in Chapter 4, Llama3.1 outperformed OpenAI's GPT4-models. One potential reason could be that the performance we report here for GPT4 is lower than that reported in the original article by Heseltine and Clemm von Hohenberg (2024). This could be due to chance since our random seed might have differed from that of the two runs considered by Heseltine and Clemm von Hohenberg (2024). Another possibility could be that GLLM performance might vary over time even

within a specified model-version (Barrie et al., 2024). In any case, the performance of Llama3.1:70b is good and the model is open source and installed on the server where we run our TWON field studies, so its performance should be reliable for our purposes. We therefore recommend using this model.

Although the performance is already quite good, we considered whether a German language prompt would work better. The prompt used here is worded in English, to see if this made a difference our native German-speaking TWON colleague Simon Münker translated the prompt to German. We tested the performance of this prompt using the best-performing model from Table 10 (Llama3.1:70b). The performance of this prompt was relatively speaking disappointing with a macro F1 of 0.77, ten percentage points below that of the English language prompt.

5.3 Classifying ideological leaning of content in German

As explained above, we define diversity in line with the analysis in Chapter 4 in terms of ideology. To arrive at a metric suitable for TWON we considered two additional steps on top of the results in Chapter 4. We tested the performance of different prompts and conceptually translated the metric from the comment to the thread level. Both will be discussed in turn.

Since we already have political content annotated from the metric described in the preceding subsection, we simplified the prompt used for ideology to the one used by Heseltine and Clemm von Hohenberg (2024). This prompt differs from the prompt used in Chapter 4 which included more detail than needed (i.e. categories 'ideology absent' and 'ideology present but unclear direction') now that we have already classified political content. Also, we now test our metric on German data and the prompt used by Heseltine and Clemm von Hohenberg (2024) specifically refers to the German political context. Their prompt was worded in English, to make sure this did not affect the performance we considered three German translations which differed in the degree in which political ideology labels were converted to the German political context (i.e. (1) Links/Mitte/Rechts; (2) Linksliberal/Mittlere Mitte/Konservativ; (3) Liberal/Moderate/Konservativ, rather than the English liberal/neutral/conservative).

We tested the performance of these prompts using Llama3.1:70b using the Heseltine and Clemm von Hohenberg (2024) prompt again on the German X-data collected and manually annotated by Heseltine and Clemm von Hohenberg (2024) (N = 700), for model specifications and setting see Appendix C.4.

Table 11: Precision, Recall, F1-score, and N for different ideology prompts.

	Precision	Recall	F1 score	N
English prompt				
Liberal	.62	.82	.71	115

Table 11: (continued)

	Precision	Recall	F1 score	N
Neutral	.92	.85	.88	503
Conservative	.66	.67	.67	82
Accuracy			.82	700
Macro average	.73	.78	.75	700
German prompt 1				
Links	.76	.55	.64	113
Mitte	.84	.95	.89	498
Rechts	.87	.46	.60	74
Accuracy			.83	685
Macro average	.82	.65	.71	685
German prompt 2				
Linksliberal	.64	.79	.71	114
Mittlere Mitte	.90	.90	.90	486
Konservativ	.86	.53	.65	68
Accuracy			.84	668
Macro average	.80	.74	.75	668
German prompt 3				
Liberal	.62	.56	.59	108
Moderate	.86	.91	.88	475
Konservativ	.73	.55	.63	66
Accuracy			.81	649
Macro average	.74	.67	.70	649

Again, the English-worded prompt performed best. The German-worded prompt 2, using the Linksliberal/Mittlere Mitte/Konservativ labelling has about equal performance and better accuracy, but this is mostly due to the easier majority 'Mittlere Mitte'-class, which is the substantially less interesting one. We therefore prefer the slightly superior performance of the English-worded prompt on the minority class of Conservative (F1 macro 0.67 versus 0.65), but arguably, this can be considered a matter of taste. We then tested which model performed best using this prompt. To increase reproducibility we decreased the temperature setting from 0.1 to 0, similar to the procedure used above for the political nature of comments.

Table 12: Precision, Recall, F1-score, and N for ideology using different models (temperature zero).

	Precision	Recall	F1 score	N
Llama3.1:8b				
Liberal	.55	.55	.55	111
Neutral	.85	.85	.85	499
Conservative	.46	.44	.45	81
Accuracy			.76	691
Macro average	.62	.62	.62	691
Llama3.1:70b				
Liberal	.62	.81	.70	115
Neutral	.92	.86	.89	503
Conservative	.70	.68	.69	82
Accuracy			.83	700
Macro average	.75	.78	.76	700
GPT4				
Liberal	.59	.88	.70	115
Neutral	.95	.78	.86	503
Conservative	.57	.78	.66	82
Accuracy			.80	700
Macro average	.70	.81	.74	700
GPT4-Turbo				
Liberal	.51	.86	.64	115
Neutral	.94	.79	.86	503
Conservative	.66	.70	.68	82
Accuracy			.79	700
Macro average	.71	.78	.73	700
GPT4o				
Liberal	.64	.88	.74	115
Neutral	.94	.83	.88	503
Conservative	.63	.76	.69	82
Accuracy			.83	700
Macro average	.74	.82	.77	700

Table 12 shows that changing the temperature setting had little effect on the performance of the Llama3.1:70b model. It also shows that the performance of all large GLLMs is fairly similar. Only the smaller Llama3.1:8b performs slightly less well. Given the better replicability of Llama3.1:70b com-

pared to the OpenAI models, and since we have a stable version running on the TWON server at the University of Trier, it's better security in terms of data privacy and lower costs per iteration since we host it locally, we prefer Llama3.1:70b for TWON.

5.3.1 Classifying the quality of content

As explained above we will build on the results of Chapter 4 to classify the quality of content and define it in line with the rationality measure defined and discussed there. In terms of classifiers Table 8 in Chapter 4 shows that the fine-tuned BERT-model achieves the best performance, however, this method requires manual coding of a suitable sample of any TWON simulation. Therefore, TWON researchers might opt for the best GLLM instead, since it can be applied out of the box.

Table 8 shows that the performance for this GLLM, Llama3.1:70b, is inferior to that of the fine-tuned BERT transformer. However, this could be partly due to particularities in the relatively small test set (N = 773). Results on the training set (N = 3089) reported in Table 29 in Appendix D are much better and the F1 macro average of Llama3.1:70b is with 0.69 only 3 percentage points removed from the macro F1 average of 0.72 of the fine-tuned BERT (see Table 8 in Chapter 4). As discussed in Chapter 4) the decrease in performance between the training and test set could be due to overfitting since various prompts were used to find the best fit. However, in our case, the number of prompts tested for rationality was only four, so the danger of overfitting is limited. Also if overfitting was a problem in our case we would expect the performance in the test set to be lower than the training set across all or most concepts, since we followed the same procedure for each of them. However, a comparison between the performance of Llama3.1:70b on the other variables reported in Chapter 4) on test and training set gives little reason to assume structural overfitting. For interactivity, incivility and diversity:liberal the macro average F1 was equal in the training and test sets, while the macro average F1 for diversity:conservative even increased from 0.78 on the training set to 0.81 on the test set.

In sum, we propose to use Llama3.1:70b to classify quality/rationality for TWON, but if researchers want to be on the safe side they should finetune a BERT model (see Weber and Reichardt (2023) for a similar argument).

5.4 From comment to debate

So far this chapter has explained we propose to use exposure, engagement, participation, diversity of exposure and quality of exposure as debate quality indicators for TWON and that English-worded prompts for Llama3.1:70b work best to classify political content and ideology. Exposure would then be operationalized as exposure to political comments, engagement as liking or sharing such comments,

participation as posting such comments, diversity as the ideology of such comments and quality as the degree to which these comments are rational. However, TWON is geared towards improving the debate in general, not only for a single user. It can therefore be helpful for specific research applications to broaden these indicators to apply to the thread level or higher. In this way, they can help evaluate which discussion threads have a higher debate quality than others, or which platform settings help foster better debate.

This move involves aggregating information from the individual to the aggregate level. We will do so at the thread level here, but the same procedure can be used for other levels, such as on a topic or platform. Threads contribute more to *relevant* exposure if they contain more political comments, likewise if more political comments in a thread are liked or shared it contributes more to engagement, if more people contribute posts in a thread it helps more for participation and if a thread contains more rational comments contributing arguments and evidence it is of higher quality. Things are a little more complex for diversity/ideology.

5.4.1 Diversity

For diversity in terms of ideology, a good balance between left and right-leaning content is often preferred (Loecherbach et al., 2020). However, there is a fundamental difference between such common applications and the TWON setting. The traditional balance approach is focused on perception, i.e. the diversity of the comments available to a reader or watcher, which is fitting for one-to-many media like newspapers or TV. Social media platforms, on the other hand, are many-to-many and offer the opportunity to participate in the debate. To reflect the diversity of participation, having a diversity metric that measures the contribution of a particular comment to the diversity of a thread can be helpful since it allows diversity to be applied to exposure, engagement and participation while acknowledging the context of the thread.

A naive balanced ideology approach could measure the diversity of participation by whether a single person has posted comments from different ideologies, but this would obviously be an odd requirement. A more reasonable traditional balanced approach would be to collapse diversity of exposure with diversity of participation and measure whether people from different ideological backgrounds contributed to a thread. However, from the perspective of deliberative democracy, it should also matter whether people from one ideological background only participate with like-minded people in threads – or echo chambers – of like-minded comments or whether they actively engage with comments from the other ideological side and contribute to threads where their views are the minority. This appears to be a key requirement for the ‘openness’ and ‘inclusion’ criteria of deliberative democracy to be enacted

(Leeper and Slothuus, 2018). In this way, diversity of participation can be measured simultaneously at the thread and the person level.

We therefore propose to measure diversity at the comment level to reflect the contribution a comment makes to the ideological balance in the preceding comments in the thread. We proposed parameter δ to reflect this contribution. As a first step, we propose a simple implementation to give each comment a δ of 2 if it expresses a viewpoint from the minority ideology within a thread and 0 if it expresses a viewpoint from the majority ideology. For example, if someone posts a right-leaning comment to a thread of 3 preceding right-leaning comments, this post does not add to the ideological diversity of the thread. If the same comment was added to a thread of two left-leaning and one right-leaning comment, it adds a diversity of $\delta = 2$ to the thread. After this latter comment has been added to the latter thread, this thread now consists of two left-leaning and two right-leaning comments. If another right-leaning comment is added we still consider it to be more diverse since the debate is progressing, however, it also deviates from the ideal balance of equal attention to both sides, therefore, we add a δ of 1 rather than 2 for such ties. Finally, we consider comments that balance ideological leaning within the comment, or political comments without a clear ideological leaning equally relevant for the diversity of a thread, since bridging ideological divides and adding factual or neutral evidence to a debate can be just as conducive to the deliberative goal of searching solutions to common problems (see f.e. Dahlberg, 2007), therefore we consider the majority ideology not just in terms of left and right, but left, right and centre. Future work may consider the effects of further refined versions of δ.

5.4.2 A note on visibility

In the move from measuring debate quality for legacy media like newspapers or TV to doing so for online social media platforms, there are two more differences to consider. First, content analyses of deliberation often implicitly assume the content to be coded post-hoc, i.e. after a debate is finished. Researchers can investigate whether all sides had equal opportunity to present their arguments. On social media platforms, like TWON, this can be problematic since debates there have no fixed end date. Social media debates can be continuous and unfolding, new comments can be added to threads years after they have been initiated, and researchers might want to measure diversity at different time intervals and even include it in real-time recommendation metrics. Second, on social media platforms, users often don't see all comments in a thread. Top comments are likely to be more visible and therefore matter more for the ideological diversity of a thread to most users. Order of comments matters, if a thread only has left-leaning comments for the first twenty comments or so, before another twenty right-leaning comments follow such thread might be perceived differently than threads that alternate

left and right-leaning comments. Therefore considering whether a user or TWON participant was actually exposed to a certain post at a relevant time interval would matter for the debate quality. Since the example above is on debate quality at the thread level such exposure is assumed, but other research objectives like evaluating debate quality at the platform level should take visibility into account for the diversity and rationality metrics.

6 Additional explorations

As explained in 5 we propose ideological diversity as the diversity metric for TWON, but this is mainly due to the lack of better alternatives. This chapter will describe the results of our effort so far to construct and validate such a metric. The first paragraphs will show how the main metric we propose failed to pass the usability tests and is hence not ready for TWON. Furthermore, we will test some varieties upon that initial metric that show some promise, but as of yet, do not meet the standards required for TWON application. Most importantly they are not yet validated against a tailormade hand-coded validation set.

6.1 Introducing claim diversity

Diversity is a concept central to many key problems within current democracies: A lack of diversity is the key concern of studies on filter bubbles and echo chambers. Deliberative perspectives on democracy require citizens to cooperate to solve common issues by sharing insights and being open to learning from others. Diversity of opinions and information is a key requirement for this to happen. It, therefore, may be surprising that various recent literature reviews conclude that, although a plethora of methods have been proposed, the field has yet to come up with a valid general metric for deliberative diversity (Bächtiger and Parkinson, 2019; Goddard and Gillespie, 2023; Joris et al., 2020; Loecherbach et al., 2020). To address this gap, we propose a theoretically grounded, yet general, computational metric of deliberative diversity for social media platforms. We use a state-of-the-art, open-source, local GLLM (Llama 3.1) to find the claims made in each comment and evaluate the diversity of these claims via the distances between their embeddings using a state-of-the-art open source embedding model (Mixedbread AI: mxbai) [4].

[4] https://huggingface.co/mixedbread-ai/mxbai-embed-large-v1, this model is currently on the leaderboard of best models on benchmark dataset performance for STS (Semantic Text Similarity) and Pair classification see https://huggingface.co/spaces/mteb/leaderboard

6.2 The difficulty of measuring diversity

A key problem why diversity is hard to pin down is that it depends so much on context: it is not an attribute of a specific text in itself. Diversity is an aggregate quality of, for example, a conversation, news report or outlet, not a property of any text in isolation. Traditionally, though, communication scientists are accustomed to treating single texts as the unit of analysis. Scholars have taken one of three different routes to tackle this issue: They propose a measure of diversity for their specific case, which would then need to be redefined to generalize to other contexts (e.g., Voakes et al., 1996); they reduce diversity to some larger, easier to measure and more fixed set of categories like partisan slant, actors or topics (see Joris et al., 2020); or they abandon any specific substantial criteria and look for anything indicating the opposite of similarity (e.g., Welbers et al., 2018). While the first approach is difficult to generalize and apply to vast amounts of (online) data, from a deliberative perspective the top-down nature of the second approach is too coarse and fails to incorporate how new information is introduced and shared, and the third approach does not discriminate between theoretically meaningful and more arbitrary differences like the writing style.

6.3 Our metric

We propose that the abilities of Generative Large Language Models (GLLMs) might present an opportunity to combine the tailor-made and granular advantages of context sensitivity from the first approach, with the scope of the more generalized second and third approaches. GLLMs can analyze the information presented in an unseen text and provide us with the categories needed for diversity in a bottom-up, yet computationally scalable way. We draw on work on argument quality (e.g., Wachsmuth et al., 2017) and deliberation (Stromer-Galley, 2007) and select “claims” as our basic unit to construct our list of categories. Bächtiger and Parkinson Bächtiger and Parkinson (2019, pp 24-5) argue that “deliberation can be defined according to a relatively strict ‘core’ of values which are reason-giving linked to claim-making and listening”. Claims are thus central units of information within the deliberative process. We then propose to either prompt the GLLM directly for the difference between the claims or to calculate the deliberative diversity of a corpus, as the average distance between the embeddings of the claims found in that corpus. In this way, we can differentiate corpora with many similar claims from those with more mixed or different claims. This method can, in principle, be used on any text and produce diversity statistics on the go, for example, to use in a ranking algorithm for a social media feed.

6.4 Procedure

We calculate our metric using the following steps:

1. We give the following system prompt[5] to Llama3.1:70b:

   > Instruction:
   >
   > You are a text annotation assistant. Analyze a social media comment enclosed in chevrons <..>. Identify and list the claims within this comment. Claims can be related to events, issues, opinions or concerns. Claims are defined as the main assertion or conclusion of an argument. You summarize each claim into a short simple sentence.
   >
   > Response format:
   >
   > You provide only the list of claims, separated by commas, without any additional text or explanations. If no claims can be identified, return an empty list [].
   >
   > Response format template:
   >
   > [”claim 1”, ”claim 2”, ... ”claim x”]

2. We pass the following user prompt including the comment in question:

   > The following set of social media posts are replies to a news- or infotainment-post. Check whether your answer strictly adheres to the specified format. ”Posts”: <row[”commentText”]>

3. We store the claims per comment in a pandas dataframe.

4. We pass the set of claims per comment to mixedbread using the following prompt:

   > You help me get embeddings for a sentence. I provide you with a context and a sentence and you reply only with that exact sentence. Context = ‘claims made in social media replies to a news- or infotainment-post’; Sentence: <claims>

5. We store the embeddings in a new column in the dataframe.

6. We calculate mean pairwise Euclidian distances between embeddings per group.

6.5 Test and validation

To validate this metric we first explore whether it can meaningfully distinguish between the diversity of debate on different platforms (YouTube vs Twitter/X) topics or threads present in the dataset described

[5] We used GPT4o for prompt improvement

in Section 4. As a benchmark, we provide similar metrics for the distances between the mxbai embeddings of the original posts and also the distances between tf-idf representations of post and claims. Such tf-idf representations have been used in earlier work on diversity (de Vries et al., 2022).

6.5.1 Diversity of posts

Table 13 shows the average distances between each comment from each platform to each of the comments of the other platform and vice versa. The results show that the distances between comments within a platform are about as large as those between comments between the platforms. Consequently, either the content on both platforms is similar, or this metric cannot meaningfully discriminate between the content of these platforms. Perhaps the content does vary, but this variance isn't captured by these basic representations. Table 14 shows the same distances, but now for mxbai embeddings. These embeddings have been specially developed for similarity tasks[6] so perhaps they can better grasp the differences between the content on the different platforms. However, the results in Table 14 show this is not the case: the average distance among YouTube comments is larger than that between YouTube and Twitter comments. Tables 15 and 16 show similar results for different topics (see Appendix A.2 for details on the construction of topics). Although we need a more detailed manual content analysis to verify these results and find out whether they are due to the inability of tf-idf and mxbai to pick up on relevant differences or whether results correctly reflect the absence of such differences, these results question whether tf-idf representations and mxbai embeddings of posts could be useful for TWON.

	YouTube	Twitter/X
YouTube	1.3871	1.3785
Twitter/X	1.3785	1.3664

Table 13: Average pairwise distances between tf-idf of comments between platforms

	YouTube	Twitter/X
YouTube	8.2750	7.5134
Twitter/X	7.5134	6.0605

Table 14: Average pairwise distances between mxbai embeddings of comments between platforms

6.5.2 Diversity of claims

Perhaps the posts themselves differed in too many unrelated aspects like style or grammar which distracted the embeddings from focussing on the substantial arguments used. Therefore our metric does

[6] https://www.mixedbread.ai/docs/embeddings/overview

	YT general	YT_Mueller/Comey	YT_Economy	YT_Middle East	Twitter general
YT general	1.3646	1.3802	1.3802	1.3804	1.3677
T_Mueller/Comey	1.3802	1.3905	1.3933	1.3935	1.3814
YT_Economy	1.3802	1.3933	1.3913	1.3934	1.3813
YT_Middle East	1.3804	1.3935	1.3934	1.3910	1.3817
Twitter general	1.3677	1.3814	1.3813	1.3817	1.3664

Table 15: Average pairwise distances between tf-idf of comments between topics

	YT general	YT_Mueller/Comey	YT_Economy	YT_Middle East	Twitter general
YT general	7.7612	8.0753	8.0740	8.2375	7.1791
YT_Mueller/Comey	8.0753	8.1638	8.3296	8.4557	7.5444
YT_Economy	8.0740	8.3296	8.3254	8.4921	7.5422
YT_Middle East	8.2375	8.4557	8.4921	8.5103	7.7297
Twitter general	7.1791	7.5444	7.5422	7.7297	6.0605

Table 16: Average pairwise distances between mxbai embeddings of comments between topics

not consider the original posts, but rather the claims made within them. Table 17 shows the pairwise differences between claims made about different topics. Again the results show that the embeddings fail to differentiate meaningfully even between topics with the largest average pairwise distances among YT Middle East comments (8.5) rather than between any of the topics. This is quite surprising, given that the similarity of embeddings also underly topic models such as BERTopic[7]. The embeddings of both our posts and our claims cannot meaningfully distinguish between different topics.

Figure 1 confirms that claims added little to explain the differences between comments on top of their original texts, as the cosine similarities of both mxbai embeddings are extremely large.

	YT general	YT_Mueller/Comey	YT_Economy	YT_Middle East	Twitter general
YT general	17.0677	17.2220	17.2865	17.4035	17.1921
YT_Mueller/Comey	17.2220	16.9826	17.3383	17.3903	17.3090
YT_Economy	17.2865	17.3383	17.3873	17.5042	17.3870
YT_Middle East	17.4035	17.3903	17.5042	17.3745	17.5199
Twitter general	17.1921	17.3090	17.3870	17.5199	17.1842

Table 17: Average pairwise distances between mxbai embeddings of *claims* between topics

Table 19 shows that this result is not due to the large number of comments per topic in which more detailed differences cancel out. The table shows that mxbai embeddings of claims also can't differentiate between the different threads related to different YouTube videos. Although for each thread there are threads to which its comments differ more than among themselves, differences are small.

In sum, we must conclude that these first tests for the usability of embedding distances of either posts or claims with simple tf-idf or advanced mxbai embeddings hold little promise for TWON.

[7] see `https://maartengr.github.io/BERTopic/getting_started/embeddings/embeddings.html`

Label	Original Video Title
Haley: Airstrikes	Haley: Airstrikes ”crippled” Syria’s chemical weapons program
Duck Dynasty	‘Duck Dynasty’ stars on dangers of the ‘liberal left’
Hannity Panel	‘Hannity’ panel on the important questions Mueller needs to answer
AOC’s Chief	AOC’s chief of staff resigns amid multiple controversies
Anderson Cooper	Anderson Cooper lays out questions surrounding Mueller report
Hannity: Trump	Hannity: Trump puts Iran on notice after drone shot down
John Berman	John Berman: Is it even news when the President lies?
Monologue: Trump’s	Monologue: Trump’s ”Got Away with Treason” Tour | Real Time with Bill Maher (HBO)
Sanders: Bolton	Sanders: Bolton is a guy who likes war
Source: Trump	Source: Trump attended 2015 hush money meeting
Teacher Fired	Teacher who said she was fired over topless selfie says she ‘lost everything’
Trump Talks	Trump talks race, football, foreign policy and more ahead of the Super Bowl
Trump: US, France	Trump: US, France and UK launch strikes on Syria

Table 18: Labels and Original Video Titles

	Haley: Airstrikes	Duck Dynasty	Hannity Panel	AOC’s Chief	Anderson Cooper	Hannity: Trump	John Berman	Monologue: Trump’s	Sanders: Bolton	Source: Trump	Teacher Fired	Trump Talks	Trump: US, France
Haley: Airstrikes	5.5076	6.6594	6.3064	5.9904	5.8897	6.2361	5.9038	6.0900	6.2891	6.1456	6.6184	6.0661	5.9634
Duck Dynasty	6.6594	6.2181	6.6248	6.4182	6.2806	6.8820	6.3138	6.3693	6.7278	6.4959	6.6786	6.4088	6.1688
Hannity Panel	6.3064	6.6248	5.9051	5.9218	5.7841	6.3854	5.6944	5.8915	6.3142	5.9617	6.5590	6.0015	6.2121
AOC’s Chief	5.9904	6.4182	5.9218	5.3846	5.5930	6.1263	5.4380	5.6616	6.0361	5.7046	6.3729	5.7572	5.9580
Anderson Cooper	5.8897	6.2806	5.7841	5.5930	5.0361	6.0620	5.1652	5.5170	5.9695	5.5436	6.1840	5.5480	5.7750
Hannity: Trump	6.2361	6.8820	6.3854	6.1263	6.0620	6.0443	5.8998	6.1455	6.3461	6.2116	6.9459	6.1634	6.4436
John Berman	5.9038	6.3138	5.6944	5.4380	5.1652	5.8998	4.6841	5.3307	5.8454	5.3695	6.2559	5.4352	5.8054
Monologue: Trump’s	6.0900	6.3693	5.8915	5.6616	5.5170	6.1455	5.3307	5.2788	6.0189	5.7047	6.3361	5.6006	5.8148
Sanders: Bolton	6.2891	6.7278	6.3142	6.0361	5.9695	6.3461	5.8454	6.0189	5.9489	6.1213	6.7696	6.0803	6.2384
Source: Trump	6.1456	6.4959	5.9617	5.7046	5.5436	6.2116	5.3695	5.7047	6.1213	5.4819	6.3995	5.8173	6.0593
Teacher Fired	6.6184	6.6786	6.5590	6.3729	6.1840	6.9459	6.2559	6.3361	6.7696	6.3995	5.7365	6.3672	6.1645
Trump Talks	6.0661	6.4088	6.0015	5.7572	5.5480	6.1634	5.4352	5.6006	6.0803	5.8173	6.3672	5.3631	5.8562
Trump: US, France	5.9634	6.1688	6.2121	5.9580	5.7750	6.4436	5.8054	5.8148	6.2384	6.0593	6.1645	5.8562	5.3023

Table 19: Diversity in average pairwise distance of mxbai embeddings of claims per video-thread for threads with 20 comments or more

6.6 Improving prompt diversity

We considered several potential improvements to our metric. Perhaps the problem of our initial approach detailed above was that embeddings are just too complex and multifaceted, capturing everything from grammar to syntax, style and word use, to capture the diversity we are interested in. Another possibility is that the meaning of a comment might be context-dependent and therefore difficult to grasp by an embedding model that only considers that comment. Consider the comment ‘I disagree’: It is unclear from this comment what the author disagrees with. A third possibility is that our procedure above embedded all claims per comment at once, but perhaps some were diverse and others were not. Grouping them might wash out the difference of that one claim. To test whether any of these factors contributed to the results described above, we propose and test several variations upon our metric. This paragraph will explain our procedure and first results, comparing whether different versions of our metric are associated with manual annotated disagreement. Tests again use the dataset described in

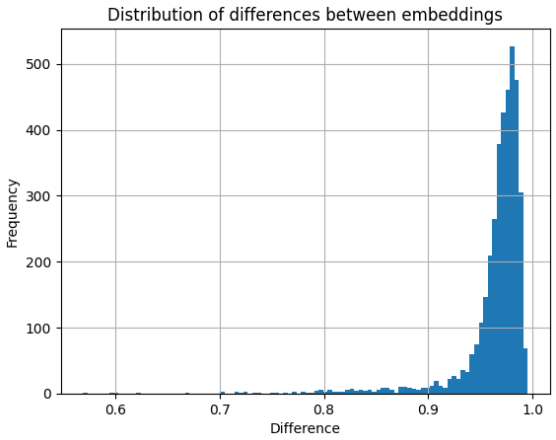


Figure 1: Distribution of cosine similarities between comment and claim embeddings per comment over Public Sphere corpus

Chapter 4, but since thread context will now be taken into account we only include YouTube comments of threads of 2 comments or more and exclude the first comment per thread (which serves as a reference and has no preceding reference in a thread to be compared with). The analyses in this section use OpenAI's GPT4o, for specification and settings, see Appendix C.4.

A first improvement for the metric could be to prompt the GLLM directly for any new information in a comment rather than rely on embeddings for this, perhaps embeddings are too hard to interpret for our use case and prompting the GLLM directly might make the task clearer for the model. We therefore developed the following system prompt[8]:

Instruction

Identify the degree of similarity of the information presented in a social media comment with respect to preceding comments in a thread on an integer scale of 1 (completely dissimilar) to 10 (identical). A score of 5 indicates that the comments share context or topic, but otherwise present different information. Follow these steps:

1. You will receive a target comment in double chevrons <<...>> along with a JSON containing the preceding comments and their `comment_index` enclosed in single chevrons <...>.

2. Determine the most similar comment to the target comment in the preceding comments in terms of the information they present. If only one preceding comment is provided pick this comment as the most similar comment.

3. Find the `comment_index` of this most similar comment. If you can't decide which com-

[8] We used GPT4o for prompt improvement

ment is most similar, pick a preceding comment at random.

4. Identify the degree of similarity of the target comment with respect to that comment on a range of 1-10.

5. If no target comment is provided or only '[]' return an empty list `[]` as value for both the `most_similar_comment_index` and the `similarity_score`, if you can't decide on the similarity score return an empty list for that value.

6. Always and only respond with the `comment_index` and similarity score.

Response format in JSON:

```
[
    {
        "most_similar_comment_index": "1",
        "similarity_score": "1"
    }
]
```

The prompt assumes an input of both the preceding comments in a threat as well as a target comment and then evaluates to what degree this target comment presents any new information. We used the following user prompt:

```
"Target comment":<<{row["commentText"]}>>,
"preceding comments":\n<{df.loc[:,["dataset_index", "commentText"]]
    [:index].to_json(orient="records")}>
```

We did the same for the claims mined by GPT4o from all comments in YouTube threads with 2 posts or more. The similarity scores per comment indicate how different this comment is compared to the most similar preceding comment in a thread. The similarity scores per claim do the same for each claim compared to the preceding claims. If a target comment contained more than one claim, the most diverse claim would be the claim with the lowest similarity score to the preceding most similar claim. The similarity score of this claim was then used as the claim diversity for this comment. All similarity scores are therefore at the comment level to facilitate comparison.

The second improvement we considered was to give more context to the model on the meaning of a target post in relation to the preceding comments in a thread. To do so we developed a procedure we

call ‘post expansion’. This consists of prompting a GLLM (in this case GPT4o) with the following system prompt[9]:

```
# Instruction
Rephrase a social media post to reflect its meaning within the context
of a conversation thread:
1. You'll receive a thread in chevrons `<...>' and a target post in
double chevrons `<<...>>'.
2. If the post is clear without context, repeat it as-is.
3. If the post depends on context, expand it to include necessary details.
4. Respond with only the expanded post.
5. If the post does not refer to context provided in the thread,
or if in doubt, respond with the exact target post as you
received it. If no target post was presented reply with an empty list []
## Example
**Input**:
- Thread: <`Comment 1', `Comment 2', `Comment 3'>
- Target reply: `<<This is so relatable!>>'
**Output**:
- `[context from previous comments] is so relatable!'
**Text:
```

along with this user prompt:

```
"Thread":\n<{df["commentText"][:threadset].to_list()}>,
"Target reply":<<{df["commentText"][index]}>>
```

Since post expansion might affect which claims are found in a post we also ran the claim mining prompt described in paragraph 6.4 on the expanded posts with GPT4o.

The third improvement we considered was that perhaps embeddings could work, but embedding all claims at once might be too distracting for the model, therefore we introduced an intermediate step of finding the most different claim of a comment compared to the claims in the preceding comments in a thread. To do so we embedded all comments and all claims individually using OpenAI’s ADA[10] model again using the secure Microsoft Azure platform. We used these embeddings to calculate the Euclid-

[9] We used GPT4o for prompt improvement

[10] ”text-embedding-ada-002” see `https://platform.openai.com/docs/guides/embeddings`, the model has strong performance on the MTEB industry benchmark, see `https://huggingface.co/spaces/mteb/leaderboard`

ian distances for each comment/claim with respect to all preceding comments/claims in a thread. For comments we then choose the minimal distance between the target comment and any of the preceding comments as the ADA diversity distance for that comment. For claims we first need to determine which claim is the most diverse per target comment. For each comment, we selected the claim with the largest (i.e. most diverse) pairwise distance to the closest (i.e. most similar) preceding claim.

Tables 20 and 21 show the results of our first validation tests of these improvements. We used manually coded disagreement as a baseline. Disagreement was defined in the codebook as:

> "Does this commenter disagree with the comment of another discussant?"

Results should thus be interpreted with caution since this concept is related to but different from diversity. Also, the coders were asked to evaluate disagreement by only considering the comment itself. Thus, disagreement should be expressed in the text rather than inferred from the context of the thread, as we do in our approach. Still, we would expect that comments that show this disagreement should, on average, be more diverse relative to the preceding thread than comments that did not show this disagreement.

The first two rows in Table 20 show the mean similarity score (with respect to the most different preceding comment/claim in the thread) for all comments labelled as 0='No' or 1='Yes' on disagreement. The next two rows show the same but now based on the expanded posts, and the last two rows again show the similarity scores for the expanded posts, but now only including posts manually coded as 'political' (both as included comments and as potential preceding comments in the thread). Since lower similarity scores point to more diverse comments, we expect that comments labelled as 'Yes' for disagreement have significantly lower similarity scores than those labelled as 'No'. Except for similarity scores of original posts, all rows show a relation in the expected direction, but the single-sided t-test is only significant for differences between claims based on the original posts. Based on these results there is little reason to assume that prompting GLLMs for similarity scores is better than using embeddings to determine diversity. Note that calculating similarity scores yielded a small number of missing values, where the model did not return a usable score. For claims only posts could be used that included at least one claim. However, more precise tests using a manually coded baseline of the diversity of a post with respect to its political information relative to preceding posts in a thread would be a better and more conclusive test of performance.

Table 21 shows similar t-tests but now based on ADA embedding distances for both posts and claims. As a reference, we also included whether a claim was present in a target comment at all and the number of claims. This is because it is conceivable that many comments could contain little information, for example, one comment consists of just 'sad'. Such comments are less likely to be disagreeing, which

might explain why claim embeddings predict disagreement. Since larger embedding distances indicate more diversity, we expect larger average distances for comments labelled as 1='Yes' compared to 0='No' for disagreement. This is the case for all metrics, both based on original and expanded posts. Single-sided t-tests show that differences are (nearly) significant for all but the embeddings for the original posts, although the absolute differences in distances are limited. We do observe a strong effect on the presence and number of claims. Note that the sample sizes for the different t-tests vary for several reasons: the number of claims could be calculated regardless of whether a comment was preceded by other comments in a thread, while a distance requires at least one preceding comment; some posts had no claims, so distances to other claims could not be calculated; the post-expansion, claim mining and embedding distance calculation procedures each yielded some missing values. The number of comments labelled as disagreeing is also very limited, so all in all these results should be interpreted with caution.

There does not appear to be a clear implication from these findings that expanded comments work better than original comments, although claims appear to be a little stronger associated with disagreement than the full comments. The results look promising but further testing is needed to ascertain this new approach to measuring diversity validly captures manually coded diversity.

baseline	disagreement						
	mean_disagreement=0	mean_disagreement=1	p_value	t_stat	n_disagreement=0	n_disagreement=1	n_total
original posts							
similarity_score_comment	3.774	3.863	0.661	-0.416	1874.0	117.0	1991.0
similarity_score_claim	4.393	3.675	0.000	3.466	1781.0	117.0	1898.0
expanded posts							
similarity_score_comment	4.216	4.115	0.319	0.470	1827.0	113.0	1940.0
similarity_score_claim	4.098	4.012	0.352	0.381	1079.0	84.0	1163.0
expanded political posts							
similarity_score_comment	4.816	4.615	0.196	0.860	724.0	65.0	789.0
similarity_score_claim	3.912	3.612	0.117	1.199	637.0	67.0	704.0

Table 20: t-tests between the mean similarity scores and manually coded disagreement according to GPT4o

baseline	disagreement						
	mean_disagreement=0	mean_disagreement=1	p_value	t_stat	n_disagreement=0	n_disagreement=1	n_total
original posts							
claim_dum	0.576	0.804	0.000	-6.461	2323.0	138.0	2461.0
claim_count	1.551	2.913	0.000	-5.528	2323.0	138.0	2461.0
closest_comment_embedding	0.677	0.679	0.344	-0.401	1878.0	117.0	1995.0
closest_claim_embedding	0.651	0.662	0.048	-1.679	943.0	82.0	1025.0
expanded posts							
claim_dum	0.605	0.790	0.000	-5.092	2323.0	138.0	2461.0
claim_count	1.621	2.855	0.000	-5.127	2323.0	138.0	2461.0
closest_comment_embedding	0.662	0.671	0.051	-1.649	1878.0	117.0	1995.0
closest_claim_embedding	0.633	0.657	0.000	-3.392	1009.0	82.0	1091.0

Table 21: t-tests between claim counts and ADA embedding distances and manually coded disagreement

7 Conclusion

This report has discussed the potential contribution online social media discussions can make to (summative) deliberation and proposed a set of indicators to measure this contribution tailored to the TWON project. By taking a summative approach to deliberation TWON can yield refined and novel insights on the contribution of social media to deliberative democracy (Chapter 3). Chapter 2 presents a hands-on and detailed implementation which can be used to apply these metrics to both TWON and external data. Within TWON these metrics can help test the effect of different algorithms and dynamics on debate quality. Outside of TWON they can be used to audit the ability of existing platforms to fulfill deliberative goals. Chapter 6 looks beyond what is currently feasible and shows how advanced use of generative AI might enable more fine-grained analysis in the future.

8 Appendices

A Overview of Manual Content Analysis

Table 22: Overview of included TV news shows and the number of manually annotated comments per show.

Regular news	**n**
60 Minutes (CBS)	150
ABC Nightline	172
CBS Evening News	170
Face the Nation (CBS)	155
Meet the Press (NBC)	51
NBC News	230
PBS NewsHour	30
The 11th Hour (MSNBC)	127
World News Tonight (ABC)	117
$n_{\text{regular-news}}$	**1202**
Partisan news	
Anderson Cooper 360 (CNN)	316
Hannity (FoxNews)	304
Hardball with Chris Matthews (MSNBC)	317
The Rachel Maddow Show (MSNBC)	37
Tucker Carlson Tonight (FoxNews)	371
$n_{\text{partisan-news}}$	**1345**

Table 22: **Table A1.** (continued)

Satirical news	
Full Frontal with Samantha Bee (TBS)	124
Last Week Tonight (HBO)	136
Late Night with Seth Meyers (NBC)	216
Late Show with Colbert (CBS)	251
Patriot Act with Hasan Minhaj (Netflix)	92
Real Time with Bill Maher (HBO)	242
The Daily Show (Comedy Central)	254
n $_{\text{satirical-news}}$	**1315**

For both *Twitter* or *YouTube*, data were collected that contained the text that TV shows' accounts had posted to the respective platform, but also included all available additional details (i.e. metadata), such as whether the post contained a video (for *Twitter*). The latter was used as a requirement (a) to hold the conditions for both platforms comparable because *YouTube* per definition carries a video and (b) to increase the likelihood that comments were directed at a news item and not to, for example, a schedule announcement.

A.1 Details of Data Retrieval Procedures

YouTube. The *YouTube* Data API (v3) was employed to access the relevant information of *YouTube* videos. The *YouTube Data API* has a default quota allocation of 10,000 units per day, and each API request incurs at least one unit quota (quota cost is determined by the request type). Due to the quota limit, the collection of data lasted several weeks. Posts were collected that were sent between 2011 and 2019.

Three functions were developed to gather the data. First, video IDs of all the channels and playlists were collected using the channels.list and playlistItems.list methods with the part parameter set to "id". The maximum number of items per page was set to 50. Second, with the video-IDs fetched from the first step, video information details (i.e., the title, description, datetime, view count, like and dislike count, and comment count) were collected using the videos.list method with the part parameter set to "id", "snippet" and "statistics". In total, the details of 58,252 news videos were stored.

Third, the top-level comments (comments that reply directly to these videos; thus, the *user-comments* that are concentrated on in this study) were collected using commentThreads.list-method with the fetched video IDs from the first step. Replies (comments that reply to the top-level comments) were collected using comments.list method with the parent comment IDs fetched from the top-level comments. To get the respective username, like-count and reply-count of each comment, the part parameter was

set to “snippet” for both the top-level comments and replies. The maximum number of items per page was set to 100.

In the end, a while-loop was created, within which all three functions were called. As a result, three separate data frames (video_ids, video_info, and video_comments) were generated and stored in an SQL database. A sleep time of 24 hours was also included at the end of the while loop; so, every time the scraper hit the quota limit, it stopped calling the API for a day and re-fetched the API on the next day. In total, 33.640.673 *YouTube* user comments were stored in the database.

Twitter. The *Twitter* API—access granted on academic research grounds—allowed the collection of the most recent 3,200 tweets from a single user and an equal number of corresponding replies to those tweets. The *Python* library “Twython” was used to collect tweets posted by the selected news shows via the “get_user_timeline”-function with “tweet_mode”-parameter set to “extended” and “count” was set to the maximum of 200 per request. A “while”-statement was used with the “get_user_timeline”-function set to have a “max_id”-value of the last tweet collected to eventually reach the 3,200-tweet-limit. The oldest tweet dated back to 2016.

Tweets by the news shows were collected along with all the variables and their metadata. The data was saved in a data frame and later stored at an external SQL server as advised by Oussalah et al. (2013). In total, 4,895 tweets posted by the news accounts were stored.

The reply tweets (i.e., the *user comments* that we are interested in for this study) were, subsequently, collected via the “search”-function set to search any tweets that were directed at the TV shows’ *Twitter* handle. Further code was added to filter out the majority of replies tweets that were not direct replies to our sample of collected tweets that included a video. A maximum of 3,200 replies were acquired *per post* with a “while”-statement with similar parameters to the one described above. The “tweet_mode”-parameter was set to “extended” and the “count”-parameter was set to the per-request maximum of 100. Retweets were not included in the collected tweets nor the replies to retweets and both datasets were cleared of any duplicates. All the available variables found in the replies’ .json-file were collected and stored at an external server before they were uploaded to the SQL database (n = 2.950.500 tweets).

A.2 Sampling

First, a stratified random sample was drawn from the large databases of tweets and *YouTube* posts of the TV shows. For this selection of *Twitter* and *YouTube* posts by the TV shows, we then retrieved the earliest comments (up to 20 maximum). The sample was stratified according to the three news genres that were included, so this was evenly distributed among regular news, partisan news, and news satire shows.

To augment the similarity of what the comments are discussing, we used the Latent Dirichlet Allocation Topic Model (LDA) approach to determine the most prominent themes in the scraped *YouTube* videos (see Appendix C of Authors, 2022, for the full procedure). Three themes occurred as most prominent in U.S. TV news: the Mueller/Comey investigation, Economy, and the Middle East. All *YouTube* comments responding to videos about these themes were selected employing a keyword search (Table 5c of Authors, 2022), after which a second stratified sample was taken for each of the three topics to ensure that an equal distribution was achieved of comments under videos of regular news, partisan news, and satirical news shows. Approximately 3100 user-comments were selected for YouTube, and another 700 for Twitter, see overview in Table 23). The full sample of 3,862 comments was later manually annotated.

Table 2. Overview of where user comments originate from.

Table 23: Overview of where user comments originate from.

Sample	**n**
YouTube comments (general)	679
YouTube: Mueller/Comey investigation	828
YouTube: Economy	800
YouTube: Middle East	825
Twitter replies (general)	730
Total (*n*)	**3,862**

B Codebook items

Table 24: Overview of coded variables, their origin, and intercoder reliability.

Coding instruction	**K-α**	**%-agreement**	**Occurrence (% of comments)**
Interactivity:			
Does this comment **acknowledge** a previously posted user comment or claim of another discussant?	0.56	80.1%	26.0%
Diversity:			
How can the comment be categorized in terms of **ideological direction**?			
Answer options:			
o Absent (no political opinion)	0.58	-	57.9%
o Neutral (attacks both parties)			1.3%

Table 24: (continued)

Coding instruction	**K-α**	**%-agreement**	**Occurrence (% of comments)**
o Left/Liberal/Democratic			20.4%
o Right/Conservative/Republican			15.3%
o Has direction, but unclear in which direction			5.1%
Rationality:			
Does this comment try to justify its comment with explicit **reasoning or argumentation**?	0.30	81.1%	8.9%
Does this comment analyze the **background** of the issue?	0.23	83.2%	10.1%
Does this comment try to justify its argument with an **external evidence**?	-	93.3%	2.5%
Incivility:			
Does this comment contain targeted **name-calling**?	0.57	88.4%	14.5%
Does this comment contain **vulgar** language?	0.69	88.9%	10.1%
Does this comment **attack someone's reputation or integrity**?	0.59	83.9%	22.2%
Does this comment question other people's **intelligence** or accuse their **incompetence**?	0.71	95.0%	8.7%
Does this comment use **all-caps function** to express SHOUTING?	0.78	95.5%	10.5%
Does this comment contain **sarcasm** or satire to target somebody/something else?	0.30	84.4%	10.3%
Does this discussant threaten **individual rights** (e.g. one's personal freedom to speak or vote)?	0.32	98.0%	1.5%
Does this comment contain **discriminatory** intent toward other people or groups?	0.14	94.9%	2.2%
Does this comment suggest or invoke **violence**?	0.50	99.0%	1.5%

C Details of Model Construction and Selection

C.1 Rule-Based Measurements

Before applying the rule-based measures, we lowercased, tokenized, and stemmed all comments in the corpus using the NLTK-package (i.e., TreebankWordTokenizer and PorterStemmer) (Bird et al., 2009). All dictionaries were stemmed as well to avoid a mismatch between comments and dictionaries. For each concept (interactivity, incivility, rationality, and diversity), we selected multiple automated measurements that appeared to fit our purpose and could reasonably be selected by a researcher interested in studying the normative standards of public discussion under the model of deliberative democracy.

Interactivity. No good text-based dictionary seems to exist to measure interactivity. Similar to existing literature (e.g., Collins & Nerlich, 2015; Gruzd et al., 2011), therefore, the present study attempts to simply capture interactivity by detecting @-mentions in comments. Note that for the *Twitter*-corpus (i.e., replies under a show's tweet), each entry already consisted of at least one mention (i.e., responding to the TV show's original tweet), which was first removed before any further investigation.

Diversity. Again, it was difficult to find an appropriate dictionary for diversity. The best dictionaries available to measure the partisan nature of comments focus on ideology, especially moral values (f.e. see Zhou et al., 2024). We selected three versions of the Moral Foundation Dictionary (MFD). The MFD is designed to measure the ideological position of the texts by examining the languages used in them, and are both theoretically and empirically related to the partisan nature of text although the exact nature of that relationship remains disputed (Graham et al., 2009; Haidt & Graham, 2007; Hopp et al., 2021). MFD 2.0 is an updated version with further enhancement of psychometric properties that should improve the normality and predictive validity of the dictionary (Frimer et al., 2019). The extended Moral Foundations Dictionary (eMFD) is the most recent update, which was developed based on crowd-sourced annotated texts (Hopp et al., 2021). Conservative and liberal are measured, respectively, by calculating the ratio of corresponding words indicative of liberal values (fairness, care) and conservative values (authority, loyalty, purity). Two dummy variables were created to represent the ideology of each version of MFD: If a comment has more conservative words, the conservative variable is coded as 1, and the liberal variable as 0, and vice versa. If the counts of conservative and liberal words are equal, both variables are coded as 0.

Rationality. For rationality various formal text metrics from the field of computational linguistics are available which are easy to implement and bear similarities to the concept, although it was more difficult to find a good dictionary-based measure. We selected the Flesch-Kincaid index (FK, see Flesch, 1948) and language formality (Heylighen & Dewaele, 2002) metrics to measure language complexity and

formality of comments. In addition to the original calculation, all scores from automated approaches were later transformed to dummy variables as well for further analysis (i.e., to calculate F1 scores, precision and recall). The Flesch-Kincaid index was recoded by calculating the difference from the maximum value, so higher scores indicated more reading difficulty. Since the FK score per se has no boundary between high and low, the mean of all data was used to create a dummy variable.

Another index used to measure rationality is the Integrative Complexity score (IC, Owens & Wedeking, 2011). Dissimilar to the FK score, the IC score attempts to measure the semantic complexity of texts, as reflected using certain words belonging to a prescribed category in the LIWC dictionary. Precisely, the IC score is obtained by subtracting the number of words belonging to the negative category (e.g., exclusiveness, certainty, etc.) from the positive category (e.g., inclusiveness, causation, etc.). As counting the positive category of *sixl* (i.e., words with 6 or more letters) is very sensitive to text length (r = .98), we have taken the percentage of *sixl* rather than the absolute count of the category. The resulting correlation has decreased to r = .07. The mean of all data was used to create a dummy variable.

Also for language formality the scale is normalized to a range from 0 to 100 (see formula below), where the higher score indicates a stronger level of rationality. Such score reflects the *deeper formality* of language, which is often utilized to achieve mutual understanding by reducing the fuzziness and context-dependent words, and simultaneously reinforcing the objectivity and accuracy. Therefore, formality score is assumed to be a suitable measure for rationality. Frequency in the formula below denotes the "percentages of the number of words belonging to a particular category with respect to the total number of words in the excerpt" (Heylighen & Dewaele, 2002, p. 309).

Formality = $\frac{Freq_{noun}+Freq_{adj}+Freq_{prep}+Freq_{art}-Freq_{pron}-Freq_{verb}-Freq_{adv}-Freq_{intj}+100}{2}$

The dummy variable for formality score was split on theoretical grounds at 65, scores above that indicate a formality level on par with scientific texts (Heylighen & Dewaele, 2002). Similarly, the split for the FK score was set at 60, denoting a 10th – 12th grade readability difficulty.

Incivility. Multiple dictionaries have been developed to measure the construct of incivility. We identified six different dictionaries to be tested for this manuscript. These dictionaries include (1) Ksiazek et al.'s (2015) Hostility dictionary and (2) Ksiazek et al.'s Civility dictionary (*reverse-coded*), (3) the Incivility dictionary developed by Muddiman and Stroud (2017), (4) the LIWC-22 (Boyd et al., 2022)[11], (5) Google's What Do You Love Project (WDYL) Censored wordlist (Lewis, n.d.), and (6) the Hatebase wordlist constructed by Hatebase.org (2020). All these dictionaries are recoded into dummy variables (0 or 1), to maximize the comparability with the hand-coded data. A comment was coded as uncivil if it had at least one uncivil word.

[11] We used 'simple swear' which lists a comment as uncivil if the LIWC-22 swear score > 0.

The rationales and contexts in which these dictionaries were created differ from each other. For example, the Ksiazek et al.'s dictionaries are built to measure user comments on news platforms and social media, meanwhile *LIWC-22* and *Hatebase* wordlists cover a wide range of texts sourced from online and offline texts. Distinctly, *Google* WDYL censored words are not scientifically validated, meaning that whether a word is considered "bad" is based on *Google*'s judgment.

C.2 Traditional Supervised Machine-Learning

We used various kinds of traditional supervised machine learning (SML) to train and build specific classifiers for each of the debate quality concepts using the manually coded data. In this context, our approach involves using models that leverage bag-of-words representations, which can either be count-based or tf-idf-based.

We used an 80:20 train-test split. Since the sample was highly imbalanced for some concepts (e.g., 20.83% of the sample was coded as 1 in the rationality dimension, and 14.88% were coded as Conservative in the diversity dimension), the *resample* function in the *sklearn* package was used to *undersample* the majority class to generate a more balanced training set.[12] The evaluation metrics are calculated on the untouched, fully random, test set.

When training the classifiers towards the best model performance on the training set, eight models were estimated for each variable: two vectorizers (Count and tf-idf) × four classifiers (Multinomial Naïve Bayes, Logistic Regression, SVC with a radial ("rbf") kernel, and SVC with linear kernel). Each model was further tuned by modifying (1) the number of words considered when tokenizing a sentence (ngram_range); (2) the range of word frequency (max_df and min_df), and (3) the standard for regularization that aims to avoid over-fitting in the classifier (classifier_C).

We then needed to select the best-performing traditional supervised machine-learning (SML) classifiers for each variable. A 5-fold cross-validation with grid search was then conducted for each tuning. After finding the best parameter with the function GridSearchCV of each model (8 models × 5 variables), the model was validated with a corresponding validation set and its classification scores were saved. Table 25 shows the performance of the best parameter settings for each model on each variable in the test set in terms of macro F1, thus across classes, to enable easier comparison with the results reported in the main manuscript. Within each variable, the model with the highest F1 score of the label 1 (the positive class: a variable was present) among the 8 models was selected as the best SML model for the results section of the main manuscript. Table 26 lists the models that performed best per concept in

[12] We also tried *Imblearn* for the re-sampling, but this was eventually dropped—and will not be further discussed in this manuscript—since it performed not better than machine-learning without re-sampling and worse than resample function in *sklearn*. This indicates that *Imblearn*, given its numeric nature, might not be able to fit textual data well.

terms of F1 score on this positive class along with their performance in this class on the test set.

C.3 Fine-Tuned Transformer Model

In addition to training classifiers using the above-mentioned traditional techniques, we explore the potential of using transformer-based models in our classification pipeline using Python's PyTorch library. Here, we use the uncased version of the English-language BERT model (bert-base-uncased) and fine-tune it for our classification tasks. During this fine-tuning process, the model's parameters are updated to better suit the specific task at hand.

To address the issue of strong class imbalance, we have implemented a WeightedLossTrainer class during the training phase to account for disparities in class representation. Additionally, and for most concepts, optimize training based on the F1 score of the minority class during training. For incivility and liberal, this strategy proved insufficient. Here, we used a down-sampling strategy for the majority classes in the training dataset and subsequently optimized training on a weighted F1 score. We performed hyperparameter tuning, including exploring a range of learning rates and batch sizes. We assessed the model's performance using macro F1 scores and minority class F1 scores on the training set.

C.4 Generative AI

Since running GLLMs locally is computationally heavy, we focused on two specific model families for this project: OpenAI's GPT and Meta's Llama. We chose state-of-the-art variants of these models. For Meta we used the latest large model, Llama3.1, more specifically we used the llama3.1:70b-instruct-q6_K-variant (hereafter Llama3.1:70b). For comparison, we also used the smaller version llama3.1:8b-instruct-q6_K (hereafter Llama3.1:8b). From OpenAI we choose the two most recent and advanced models available through the Azure OpenAI API: GPT4o (the one released on 2024-08-06) and GPT4-Turbo (the one released on 2024-04-09).

We compared the effect of using different prompts. Since running classifications on (large) generative AI models is computationally expensive, and for the case of OpenAI is also financially costly, we tested the effects of different prompts in Llama3.1:70b. For the instruction, or prompt, given to Llama3.1, we first followed the codebook nearly verbatim, often only adding small label specifications (i.e. "not present *(1)*") to help the model classify the data in the correct classes. For some items, the wording of the codebook appeared to confuse the model, which resulted in high numbers of missing values. In these cases, we asked OpenAI's GPT4o to reformulate the prompt to make it better inter-

pretable for GLLMs. All prompts were checked manually to contain the same information and examples as the codebook. The only changes were in the structuring and wording of the instructions. This procedure resulted in a large number of long prompts since incivility and rationality were measured by multiple indicators.

However, long prompts are often not optimal for GLLM performance and running them is computationally costly since you need to process many runs (one for each prompt-and-comment combination) of many tokens (i.e. many words in each prompt) (Törnberg, 2024a). Therefore we also considered a simpler approach. This simpler approach would be the most likely one a researcher without our codebook would use: short concise single prompts per concept rather than indicator.

To test the effect of model size and family, we ran these simple prompts with GPT4o, GPT4-Turbo and the smaller Llama3.1:8b. At the time of writing GPT4o and GPT4-Turbo are the two most advanced OpenAI models available via the Azure OpenAI service (Microsoft, 2024b). We followed advice by Törnberg (2024a) and used a low-temperature setting in combination with a seed to improve the replicability of our results, although the creative nature of GLLMs makes perfect replicability impossible. All Llama classifications were run on a server hosted by the University Trier which utilizes four NVIDIA L40S cards with 192 GB of video memory, two Intel(R) Xeon(R) Silver 4310 CPU @ 2.10GHz with 48 threads, and 256 GB RAM capacity. We used the default model parameters, except for temperature and seed which were set at 0.1 and 42 respectively (see Morgan & Chiang, n.d.). We used the Microsoft Azure OpenAI service to run the classifications for GPT4o and GPT4-Turbo setting the parameters to default, except for temperature (0) and seed (42). To make sure no data is shared with any third parties the University of Amsterdam (UvA) opted out of abuse monitoring and content logging in addition to the guarantees offered by Microsoft that no data is shared with third parties (Microsoft, 2024a).

Appendix D below shows that Llama3.1:70b attained the best minority class F1 scores for three (rationality, incivility and interactivity) out of our four concepts on the training set. The differences were at times substantial, like for rationality where Llama3.1:70b had a minority F1 of 0.46 versus only 0.30 (GPT4o) and 0.24 (GPT4-Turbo). Llama3.1:70b also reached higher F1 macro scores for rationality and interactivity than GPT4o and GPT4-Turbo. For the variables where GPT4o or GPT4-Turbo did outperform Llama3.1 differences were marginal. For diversity, GPT4o had a slightly higher minority F1 for both liberal 0.67 and conservative 0.67 versus 0.65 and 0.64 respectively for Llama3.1:70b. Likewise, GPT4o had slightly better F1 macro scores for this concept (liberal: 0.79 vs 0.77; conservative 0.80 vs 0.78), while GPT4-Turbo beat it marginally for F1 macro on incivility (0.78 vs 0.75). Llama3.1:70b thus performed comparable or better than the OpenAI models (see Appendix D for full results of all models on the training set). Based on this performance and the financial and ethical considerations mentioned above, the

results for Llama3.1:70b are presented in the main results section.

Note that our main conclusions and recommendations hold regardless of whether we had selected the best model per group on macro F1 or minority class F1. Code and all full prompts are available on github [13]. The wording of the prompts used for the main analysis, i.e. the simple prompts per concept are listed in Table 27.

[13] https://github.com/cl-trier/TWON-Exploration

Table 25: Performance of different supervised machine-learning (SML) classifiers for each variable based on the test set in macro average F1, Precision, Recall and Accuracy.

Variable	Vectorizer	Classifier	F1 score	Precision	Recall	Accuracy
Interactivity	Count	Multinomial NB	0.514	0.611	0.620	0.515
		Logistic Regression	0.614	0.625	0.655	0.640
		SVC(”rbf”)	0.607	0.619	0.648	0.631
		SVC(”linear”)	0.564	0.589	0.611	0.583
	TfIdf	Multinomial NB	0.551	0.635	0.652	0.554
		Logistic Regression	0.623	0.651	0.688	0.636
		SVC(”rbf”)	0.621	0.649	0.686	0.635
		SVC(”linear”)	0.625	0.651	0.688	0.640
Liberal	Count	Multinomial NB	0.405	0.572	0.595	0.410
		Logistic Regression	0.616	0.617	0.682	0.697
		SVC(”rbf”)	0.601	0.613	0.684	0.669
		SVC(”linear”)	0.611	0.614	0.680	0.690
	TfIdf	Multinomial NB	0.395	0.576	0.597	0.398
		Logistic Regression	0.544	0.592	0.654	0.589
		SVC(”rbf”)	0.532	0.589	0.650	0.572
		SVC(”linear”)	0.544	0.589	0.650	0.591
Conservative	Count	Multinomial NB	0.392	0.570	0.614	0.404
		Logistic Regression	0.557	0.571	0.631	0.665
		SVC(”rbf”)	0.528	0.555	0.606	0.627
		SVC(”linear”)	0.543	0.564	0.621	0.647
	TfIdf	Multinomial NB	0.401	0.558	0.600	0.418
		Logistic Regression	0.511	0.573	0.647	0.572
		SVC(”rbf”)	0.505	0.571	0.641	0.563
		SVC(”linear”)	0.514	0.571	0.643	0.578
Rationality	Count	Multinomial NB	0.316	0.599	0.573	0.318
		Logistic Regression	0.626	0.621	0.672	0.710
		SVC(”rbf”)	0.670	0.657	0.718	0.750
		SVC(”linear”)	0.661	0.649	0.700	0.750
	TfIdf	Multinomial NB	0.384	0.605	0.608	0.384
		Logistic Regression	0.570	0.609	0.675	0.612
		SVC(”rbf”)	0.581	0.614	0.683	0.625
		SVC(”linear”)	0.579	0.613	0.682	0.622
Incivility	Count	Multinomial NB	0.563	0.661	0.609	0.592
		Logistic Regression	0.683	0.683	0.683	0.684
		SVC(”rbf”)	0.660	0.660	0.660	0.661
		SVC(”linear”)	0.657	0.657	0.657	0.658
	TfIdf	Multinomial NB	0.561	0.656	0.606	0.590

Table 25: (continued)

Variable	Vectorizer	Classifier	F1 score	Precision	Recall	Accuracy
		Logistic Regression	0.644	0.663	0.654	0.647
		SVC(”rbf”)	0.657	0.674	0.666	0.660
		SVC(”linear”)	0.646	0.662	0.655	0.648

Table 26: The best traditional supervised machine-learning (SML) classifiers for each variable on the positive class (i.e. variable is present) on the test set and their performance on these metrics.

Variable	Original ratio	Vectorizer	Classifier	F1	Recall	Precision	Accuracy
Interactivity	.28	Tfldf	Logistic Regression	.55	.80	.42	.64
Liberal	.18	Count	Logistic Regression	.45	.66	.33	.70
Conservative	.15	Tfldf	Logistic Regression	.34	.75	.22	.57
Rationality	.20	Count	SVC(”rbf”)	.51	.66	.41	.75
Incivility	.47	Tfldf	SVC(”rbf”)	.69	.79	.61	.66

Table 27: Prompt wording simple prompts.

variable	instructions
interactivity	Does this comment acknowledge or respond to another user’s comment? Instructions: Code Yes (1) if the comment shows agreement or disagreement with a specific user’s statement, often signaled by a username or phrases like ‘Yes,’ ‘No,’ or ‘I agree.’ Code No (0) if it lacks a clear acknowledgment or is only an insult. Respond with only the predicted class (0 or 1) of the request. Text: {text} Class: ”0”: ”No”, ”1”: ”Yes”
diversity	Classify the following message as ideologically liberal (0), ideologically neutral (1), or ideologically conservative (2). Ideology here is defined in the context of the US political system. Messages with no ideological content are classified as neutral. Respond with only the predicted class (0 or 1 or 2) of the request. Text: {text} Class: ”0”: ”liberal”, ”1”: ”neutral”, ”2”: ”conservative”
rationality	Does this comment provide rational analysis? Instructions: Code Yes (1) if the comment includes: Context or background, Evidence (facts, sources, authorities), Reasoning or structured argument. Code No (0) if these are absent. Respond with only the predicted class (0 or 1) of the request. Text: {text} Class: ”0”: ”No”, ”1”: ”Yes”

Table 27: (continued)

variable	instructions
incivility	Does this comment display incivility? Instructions: Code Yes (1) if the comment includes name-calling, insults, inflammatory language, sarcasm, shouting (ALL CAPS), vulgarity, discrimination, threats, or restrictions on rights. Code No (0) if none of these are present. Respond with only the predicted class (0 or 1) of the request. Text: {text} Class: ”0”: ”No”, ”1”: ”Yes”

D Individual classification results of different Generative AI prompts and models on the training set

Table 28: Precision, Recall and F1 score of simple, short prompts in Llama3.1:8b against manually coded comments in the training set.

	Precision	Recall	F1 score	N
Diversity: Liberal				
0 (No)	.88	.92	.90	2440
1 (Yes)	.64	.51	.57	649
Accuracy			.84	
Macro average	.76	.72	.73	3089
Diversity: Conservative				
0 (No)	.92	.81	.86	2611
1 (Yes)	.37	.60	.46	478
Accuracy			.78	
Macro average	.64	.71	.66	3089
Rationality				
0 (No)	.84	.98	.91	2541
1 (Yes)	.66	.15	.24	548
Accuracy			.83	
Macro average	.75	.56	.57	3089
Incivility				
0 (No)	.65	.93	.77	1567
1 (Yes)	.87	.50	.63	1522
Accuracy			.72	
Macro average	.76	.71	.70	3089
Interactivity				
0 (No)	.83	.63	.71	2297

Table 28: (continued)

	Precision	Recall	F1 score	N
1 (Yes)	.36	.62	.46	792
Accuracy			.63	
Macro average	.60	.62	.59	3089

Table 29: Precision, Recall and F1 score of simple, short prompts in Llama3.1:70b against manually coded comments in the training set.

	Precision	Recall	F1 score	N
Diversity: Liberal				
0 (No)	.93	.85	.89	2440
1 (Yes)	.57	.77	.65	649
Accuracy			.83	
Macro average	.75	.81	.77	3089
Diversity: Conservative				
0 (No)	.95	.90	.92	2611
1 (Yes)	.57	.73	.64	478
Accuracy			.87	
Macro average	.76	.81	.78	3089
Rationality				
0 (No)	.87	.96	.92	2541
1 (Yes)	.66	.36	.46	548
Accuracy			.85	
Macro average	.77	.66	.69	3089
Incivility				
0 (No)	.85	.62	.72	1567
1 (Yes)	.69	.89	.78	1522
Accuracy			.75	
Macro average	.77	.75	.75	3089
Interactivity				
0 (No)	.87	.71	.78	2297
1 (Yes)	.45	.70	.55	792
Accuracy			.70	
Macro average	.66	.70	.66	3089

Table 30: Precision, Recall and F1 score of near-verbatim codebook-based prompts in Llama3.1:70b aggregated similarly to the manually coded concepts against manually coded comments in the training set.

	Precision	Recall	F1 score	N
Diversity: Liberal				
0 (No)	.89	.90	.89	2440
1 (Yes)	.60	.59	.60	649
Accuracy			.83	
Macro average	.75	.74	.75	3089
Diversity: Conservative				
0 (No)	.96	.77	.86	2611
1 (Yes)	.40	.83	.54	478
Accuracy			.78	
Macro average	.68	.80	.70	3089
Rationality				
0 (No)	.97	.57	.72	2541
1 (Yes)	.32	.93	.47	548
Accuracy			.64	
Macro average	.75	.56	.57	3089
Incivility				
0 (No)	.90	.47	.62	1567
1 (Yes)	.63	.94	.76	1522
Accuracy			.70	
Macro average	.77	.71	.69	3089
Interactivity				
0 (No)	.85	.71	.77	2297
1 (Yes)	.42	.62	.50	792
Accuracy			.69	
Macro average	.63	.67	.64	3089

Table 31: Precision, Recall and F1 score of simple, short prompts in GPT4o against manually coded comments in the training set

	Precision	Recall	F1 score	N
Diversity: Liberal				
0 (No)	.91	.93	.92	2440
1 (Yes)	.71	.63	.67	649
Accuracy			.89	
Macro average	.81	.78	.79	3089
Diversity: Conservative				
0 (No)	.94	.93	.94	2611

Table 31: (continued)

	Precision	Recall	F1 score	N
1 (Yes)	.64	.70	.67	478
Accuracy			.88	
Macro average	.79	.82	.80	3089
Rationality				
0 (No)	.85	.99	.91	2541
1 (Yes)	.80	.19	.30	548
Accuracy			.85	
Macro average	.82	.59	.61	3089
Incivility				
0 (No)	.68	.90	.78	1567
1 (Yes)	.85	.57	.68	1522
Accuracy			.74	
Macro average	.77	.74	.73	3089
Interactivity				
0 (No)	.81	.77	.79	2297
1 (Yes)	.42	.48	.44	792
Accuracy			.69	
Macro average	.61	.62	.62	3089

Table 32: Precision, Recall and F1 score of simple, short prompts in GPT4-Turbo against manually coded comments in the training set.

	Precision	Recall	F1-score	N
Diversity: Liberal				
0 (No)	.92	.88	.90	2440
1 (Yes)	.61	.71	.65	649
Accuracy			.84	
Macro average	.76	.79	.78	3089
Diversity: Conservative				
0 (No)	.94	.92	.93	2611
1 (Yes)	.60	.66	.63	478
Accuracy			.88	
Macro average	.77	.79	.78	3089
Rationality				
0 (No)	.84	.99	.91	2541
1 (Yes)	.84	.14	.24	548
Accuracy			.84	
Macro average	.84	.57	.57	3089

Table 32: (continued)

	Precision	Recall	F1-score	N
Incivility				
0 (No)	.74	.87	.80	1567
1 (Yes)	.83	.69	.75	1522
Accuracy			.78	
Macro average	.79	.78	.78	3089
Interactivity				
0 (No)	.83	.79	.81	2297
1 (Yes)	.46	.53	.49	792
Accuracy			.72	
Macro average	.65	.66	.65	3089

E Individual classification results of different rule-based measures

Table 33: Precision, Recall and F1 score of rule-based diversity measures against manually coded diversity scores (Liberal)

	Precision	Recall	F1 score	N
MFD 1.0 (Liberal)				
0 (Non-liberal)	.82	.84	.83	633
1 (Liberal)	.19	.17	.18	140
Accuracy			.72	
Macro average	.50	.50	.50	773
MFD 2.0 (Liberal)				
0 (Non-liberal)	.83	.71	.77	633
1 (Liberal)	.22	.36	.27	140
Accuracy			.65	
Macro average	.52	.53	.52	773
eMFD (Liberal)				
0 (Non-liberal)	.82	.46	.59	633
1 (Liberal)	.19	.56	.28	140
Accuracy			.48	
Macro average	.51	.51	.43	773

Table 34: Precision, Recall and F1 score of rule-based diversity measures against manually coded diversity scores (Conservative)

	Precision	Recall	F1 score	N
MFD 1.0 (Conservative)				
0 (Non-conservative)	.86	.86	.86	660
1 (Conservative)	.19	.20	.19	113
Accuracy			.76	
Macro average	.53	.53	.53	773
MFD 2.0 (Conservative)				
0 (Non-conservative)	.88	.66	.75	660
1 (Conservative)	.19	.49	.28	113
Accuracy			.63	
Macro average	.54	.57	.51	773
eMFD (Conservative)				
0 (Non-conservative)	.84	.67	.74	660
1 (Conservative)	.12	.26	.16	113
Accuracy			.60	
Macro average	.48	.46	.45	773

Table 35: Precision, Recall and F1 score of rule-based rationality measures against manually coded rationality score

	Precision	Recall	F1 score	N
FK-score				
0 (Irrational)	.84	.59	.69	624
1 (Rational)	.24	.53	.33	149
Accuracy			.58	
Macro average	.54	.56	.51	773
Language formality				
0 (Irrational)	.77	.68	.72	624
1 (Rational)	.11	.17	.13	149
Accuracy			.58	
Macro average	.44	.42	.43	773
Integrative Complexity				
0 (Irrational)	.78	.57	.66	624
1 (Rational)	.16	.35	.22	149
Accuracy			.53	
Macro average	.47	.46	.44	773

Table 36: Precision, Recall and F1 score of rule-based incivility measures against manually coded general incivility

	Precision	Recall	F1 score	N
Ksiazek's hostility dictionary				
0 (Civil)	.65	.86	.74	408
1 (Uncivil)	.76	.49	.59	365
Accuracy			.68	
Macro average	.71	.67	.67	773
Ksiazek's civility dictionary (reverse code)				
0 (Civil)	.50	.62	.56	408
1 (Uncivil)	.43	.31	.36	365
Accuracy			.48	
Macro average	.46	.47	.46	773
Google What Do You Love Offensive Wordlist				
0 (Civil)	.57	.98	.72	408
1 (Uncivil)	.87	.17	.28	365
Accuracy			.60	
Macro average	.72	.57	.50	773
Muddiman's incivility dictionary				
0 (Civil)	.56	.99	.71	408
1 (Uncivil)	.94	.12	.21	365
Accuracy			.58	
Macro average	.75	.56	.46	773
LIWC-22 'simple swear'				
0 (Civil)	.58	.98	.73	408
1 (Uncivil)	.89	.19	.32	365
Accuracy			.61	
Macro average	.73	.58	.52	773
Hatebase wordlist				
0 (Civil)	.55	.96	.70	408
1 (Uncivil)	.72	.11	.19	365
Accuracy			.56	
Macro average	.63	.54	.45	773

References

Sweta Agrawal and Amit Awekar. Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms. In Gabriella Pasi, Benjamin Piwowarski, Leif Azzopardi, and Allan Hanbury, editors, *Advances in Information Retrieval*, pages 141–153, Cham, 2018. Springer International Publishing. ISBN 978-3-319-76941-7. doi: 10.1007/978-3-319-76941-7_11. 42

Meta AI. Introducing Llama 3.1: Our most capable models to date, 2024. URL `https://ai.meta.com/blog/meta-llama-3-1/`. 35

Quinn Albaugh, Julie Sevenans, Stuart Soroka, and Peter John Loewen. The automated coding of policy agendas: A dictionary-based approach. In *The 6th Annual Comparative Agendas Conference, Antwerp, Belgium*, 2013. URL `https://www.almendron.com/tribuna/wp-content/uploads/2017/05/CAP2013v2.pdf`. 28

Norah Alzahrani, Hisham Abdullah Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Alsubaie, Yusef Almushaykeh, Faisal Mirza, Nouf Alotaibi, Nora Altwairesh, Areeb Alowisheq, M. Saiful Bari, and Haidar Khan. When Benchmarks are Targets: Revealing the Sensitivity of Large Language Model Leaderboards, July 2024. arXiv:2402.01781. 42

Aymé Arango, Jorge Pérez, and Barbara Poblete. Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). *Information Systems*, 105:101584, March 2022. ISSN 03064379. doi: 10.1016/j.is.2020.101584. 42

Van Atteveldt, Damian Trilling, and Carlos Arcila Calderon. *Computational Analysis of Communication*. John Wiley & Sons, Hoboken, NJ, March 2022. ISBN 978-1-119-68028-4. Google-Books-ID: 0thjEAAAQBAJ. 29, 45

Christian Baden, Christian Pipal, Martijn Schoonvelde, and Mariken A. C. G Van Der Velden. Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda. *Communication Methods and Measures*, 16(1):1–18, January 2022. ISSN 1931-2458, 1931-2466. doi: 10.1080/19312458.2021.2015574. 27, 29

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep Learning for Hate Speech Detection in Tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 759–760, Republic and Canton of Geneva, CHE, April 2017. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-4914-7. doi: 10.1145/3041021.3054223. 42

Christopher Barrie, Alexis Palmer, and Arthur Spirling. Replication for Language Models Problems, Principles, and Best Practice for Political Science. 2024. URL http://arthurspirling.org/documents/BarriePalmerSpirling_TrustMeBro.pdf. 51

Geoffrey Baym. The Daily Show: Discursive Integration and the Reinvention of Political Journalism. *Political Communication*, 22(3):259–276, July 2005. ISSN 1058-4609, 1091-7675. doi: 10.1080/10584600591006492. 30

Nick Beauchamp. Modeling and Measuring Deliberation Online. In Brooke Foucault Welles and Sandra González-Bailón, editors, *The Oxford Handbook of Networked Communication*, pages 320–349. Oxford University Press, February 2020. ISBN 978-0-19-046051-8. doi: 10.1093/oxfordhb/9780190460518.013.23. 17, 27

Svetlana S. Bodrunova, Ivan Blekanov, Anna Smoliarova, and Anna Litvinenko. Beyond Left and Right: Real-World Political Polarization in Twitter Discussions on Inter-Ethnic Conflicts. *Media and Communication*, 7(3):119–132, August 2019. ISSN 2183-2439. doi: 10.17645/mac.v7i3.1934. 45

Linda Bos and Sophie Minihold. The Ideological Predictors of Moral Appeals by European Political Elites; An Exploration of the Use of Moral Rhetoric in Multiparty Systems. *Political Psychology*, 43(1):45–63, February 2022. ISSN 0162-895X, 1467-9221. doi: 10.1111/pops.12739. 45

Mark Boukes, Bob Van De Velde, Theo Araujo, and Rens Vliegenthart. What's the Tone? Easy Doesn't Do It: Analyzing Performance and Agreement Between Off-the-Shelf Sentiment Analysis Tools. *Communication Methods and Measures*, 14(2):83–104, April 2020. ISSN 1931-2458, 1931-2466. doi: 10.1080/19312458.2019.1671966. 28, 29, 44

Jelle W. Boumans and Damian Trilling. Taking Stock of the Toolkit. *Digital Journalism*, 4(1):8–23, January 2016. ISSN 2167-0811. doi: 10.1080/21670811.2015.1096598. 25, 29

R.L. Boyd, A. Ashokkumar, S. Seraj, and James W. Pennebaker. *The development and psychometric properties of LIWC-22*. University of Texas at Austin, Austin, TX, 2022. URL https://www.liwc.app. 33

André Bächtiger and John Parkinson. *Mapping and Measuring Deliberation: Towards a New Deliberative Quality*. Oxford University Press, 1 edition, January 2019. ISBN 978-0-19-967219-6 978-0-19-187262-4. doi: 10.1093/oso/9780199672196.001.0001. 16, 17, 18, 19, 20, 21, 25, 27, 57, 58

Didier Caluwaerts, Kamil Bernaerts, Rebekka Kesberg, Lien Smets, and Bram Spruyt. Deliberation and polarization: a multi-disciplinary review. *Frontiers in Political Science*, 5, June 2023. ISSN 2673-3145. doi: 10.3389/fpos.2023.1127372. 15

Lindita Camaj and Arthur D. Santana. Political Deliberation on Facebook during Electoral Campaigns: Exploring the Relevance of Moderator's Technical Role and Political Ideology. *Journal of Information Technology & Politics*, 12(4):325–341, October 2015. doi: 10.1080/19331681.2015.1100224. 32

Simone Chambers. Rhetoric and the Public Sphere: Has Deliberative Democracy Abandoned Mass Democracy? *Political Theory*, 37(3):323–350, June 2009. ISSN 0090-5917, 1552-7476. doi: 10.1177/0090591709332336. 19

Luke Collins and Brigitte Nerlich. Examining User Comments for Deliberative Democracy: A Corpus-driven Analysis of the Climate Change Debate Online. *Environmental Communication*, 9(2):189–207, April 2015. ISSN 1752-4032, 1752-4040. doi: 10.1080/17524032.2014.981560. 28, 32

Dahlberg. The Internet, deliberative democracy, and power: Radicalizing the public sphere. *International Journal of Media & Cultural Politics*, 3(1), 2007. ISSN 17408296. doi: 10.1386/macp.3.1.47/1. 18, 56

Lincoln Dahlberg. The Internet and Democratic Discourse: Exploring The Prospects of Online Deliberative Forums Extending the Public Sphere. *Information, Communication & Society*, 4(4):615–633, January 2001. ISSN 1369-118X, 1468-4462. doi: 10.1080/13691180110097030. 20, 24, 49

Erik de Vries, Rens Vliegenthart, and Stefaan Walgrave. Telling a Different Story: A Longitudinal Investigation of News Diversity in Four Countries. *Journalism Studies*, 23(14):1721–1739, October 2022. ISSN 1461-670X. doi: 10.1080/1461670X.2022.2111323. 60

Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. Generalization or Memorization: Data Contamination and Trustworthy Evaluation for Large Language Models, May 2024. arXiv:2402.15938 [cs]. 47

John Dryzek. *Discursive democracy*. Cambridge University Press, New York, 1990. 20

Jamie Dubs. All the dirty words from Google's "what do you love" project: http://www.wdyl.com/. URL `https://gist.github.com/jamiew/1112488`. 33

Stuart Duncan, Lauren Dwyer, Hanako Smith, Davis Vallesi, Frauke Zeller, and Charles Davis. Toward a computational mixed methods framework to measure online deliberative discourse. *Communication and the Public*, October 2024. ISSN 2057-0473. doi: 10.1177/20570473241284759. 28, 45

Katharina Esau, Dannica Fleuß, and Sarah-Michelle Nienhaus. Different Arenas, Different Deliberative Quality? Using a Systemic Framework to Evaluate Online Deliberation on Immigration Policy in Ger-

many. *Policy & Internet*, 13(1):86–112, March 2021. ISSN 1944-2866, 1944-2866. doi: 10.1002/poi3.232. 20, 49

Miriam Fernandez and Alejandro Bellogin. Recommender Systems and Misinformation: The Problem or the Solution? *OHARS Workshop. 14th ACM Conference on Recommender Systems, 22-26 Sep 2020, [Online]*, 2020. 22, 24

Rudolph Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233, 1948. ISSN 1939-1854. doi: 10.1037/h0057532. 33

Deen Freelon. Analyzing online political discussion using three models of democratic communication. *New Media & Society*, 12(7):1172–1190, November 2010. ISSN 1461-4448, 1461-7315. doi: 10.1177/1461444809357927. 15, 27

Deen Freelon. Discourse architecture, ideology, and democratic norms in online political discussion. *New Media & Society*, 17(5):772–791, May 2015. ISSN 1461-4448, 1461-7315. doi: 10.1177/1461444813513259. 27, 31

Dennis Friess and Christiane Eilders. A Systematic Review of Online Deliberation Research. *Policy & Internet*, 7(3):319–339, September 2015. ISSN 19442866. doi: 10.1002/poi3.95. 18, 26, 27

Dennis Friess, Marc Ziegele, and Dominique Heinbach. Collective Civic Moderation for Deliberation? Exploring the Links between Citizens' Organized Engagement in Comment Sections and the Deliberative Quality of Online Discussions. *Political Communication*, 38(5):624–646, September 2021. ISSN 1058-4609, 1091-7675. doi: 10.1080/10584609.2020.1830322. 31

Jeremy Frimer. Do liberals and conservatives use different moral languages? Two replications and six extensions of Graham, Haidt, and Nosek's (2009) moral text analysis. *Journal of Research in Personality*, 84:103906, February 2020. ISSN 0092-6566. doi: 10.1016/j.jrp.2019.103906. 33

Jeremy Frimer, R. Boghrati, Jonathan Haidt, Jesse Graham, and Morteza Dehghani. Moral Foundations Dictionary for Linguistic Analyses 2.0, April 2019. 33

Alex Goddard and Alex Gillespie. Textual Indicators of Deliberative Dialogue: A Systematic Review of Methods for Studying the Quality of Online Dialogues. *Social Science Computer Review*, page 089443932311566, February 2023. ISSN 0894-4393, 1552-8286. doi: 10.1177/08944393231156629. 15, 16, 17, 24, 28, 57

Sandra González-Bailón and Georgios Paltoglou. Signals of Public Opinion in Online Communication: A Comparison of Methods and Data Sources. *The ANNALS of the American Academy of Political and Social Science*, 659(1):95–107, May 2015. ISSN 0002-7162, 1552-3349. doi: 10.1177/0002716215569192. 25, 28

Ine Goovaerts. *Destructive or Deliberative? An Investigation of the Evolution, Determinants, and Effects of the Quality of Political Debate*. Faculteit Sociale Wetenschappen - Centrum voor Politicologie (CePo), KU Leuven, Leuven, Belgium, 2021. URL `https://lirias.kuleuven.be/retrieve/633111`. 27

Jesse Graham, Jonathan Haidt, and Brian A. Nosek. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96(5):1029–1046, May 2009. ISSN 1939-1315, 0022-3514. doi: 10.1037/a0015141. 28, 33

Anatoliy Gruzd, Barry Wellman, and Yuri Takhteyev. Imagining Twitter as an Imagined Community. *American Behavioral Scientist*, 55(10):1294–1318, October 2011. ISSN 0002-7642. doi: 10.1177/0002764211409378. Publisher: SAGE Publications Inc. 32

Johannes Grüber. CompText 2024 Keynote, 2024a. URL `https://comptext24.vercel.app/#keynote`. 46

Johannes Grüber. Get up and running with local ChatGPT/gLLMs with Ollama in R, March 2024b. URL `https://www.youtube.com/watch?v=N-k3RZqiSZY`. 46

Lei Guo, Chris J. Vargo, Zixuan Pan, Weicong Ding, and Prakash Ishwar. Big Social Data Analytics in Journalism and Mass Communication: Comparing Dictionary-Based Text Analysis and Unsupervised Topic Modeling. *Journalism & Mass Communication Quarterly*, 93(2):332–359, June 2016. ISSN 1077-6990, 2161-430X. doi: 10.1177/1077699016639231. 28

Jurgen Habermas. *The structural transformation of the public sphere: An inquiry into a category of bourgeois society*. MIT press, 1991. 26, 27

Jürgen Habermas. *Strukturwandel der Öffentlichkeit. Untersuchungen zu einer Kategorie der bürgerlichen Gesellschaft*. Luchterhand, Neuwied and Berlin, 1962. 15

Jürgen Habermas. *The theory of communicative action*. Beacon Press, Boston, 1984. ISBN 0-8070-1506-7. 15, 27

Jürgen Habermas. *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. The MIT Press, Cambridge, MA, 1996. ISBN 0-262-08243-8. 15, 26, 27

Jürgen Habermas. Reflections and Hypotheses on a Further Structural Transformation of the Political Public Sphere. *Theory, Culture & Society*, 39(4):145–171, July 2022. ISSN 0263-2764, 1460-3616. doi: 10.1177/02632764221112341. 16, 19, 20, 21, 22, 27

Jonathan Haidt and Jesse Graham. When Morality Opposes Justice: Conservatives Have Moral Intuitions that Liberals may not Recognize. *Social Justice Research*, 20(1):98–116, June 2007. ISSN 0885-7466, 1573-6725. doi: 10.1007/s11211-007-0034-z. 33

Lucien Heitz, Juliane A. Lischka, Alena Birrer, Bibek Paudel, Suzanne Tolmeijer, Laura Laugwitz, and Abraham Bernstein. Benefits of Diverse News Recommendations for Democracy: A User Study. *Digital Journalism*, 10(10):1710–1730, November 2022. ISSN 2167-0811, 2167-082X. doi: 10.1080/21670811.2021.2021804. 23

Natali Helberger. On the Democratic Role of News Recommenders. *Digital Journalism*, 7(8):993–1012, September 2019. ISSN 2167-0811, 2167-082X. doi: 10.1080/21670811.2019.1623700. 15, 17, 22

Michael Heseltine and Bernhard Clemm von Hohenberg. Large language models as a substitute for human experts in annotating political text. *Research & Politics*, 11(1), January 2024. ISSN 2053-1680. doi: 10.1177/20531680241236239. 30, 44, 49, 50, 51

Francis Heylighen and Jean-Marc Dewaele. Variation in the Contextuality of Language: An Empirical Measure. *Foundations of Science*, 7:293–340, 2002. 33

Frederic R. Hopp, Jacob T. Fisher, Devin Cornell, Richard Huskey, and René Weber. The extended Moral Foundations Dictionary (eMFD): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior Research Methods*, 53(1):232–246, February 2021. ISSN 1554-3528. doi: 10.3758/s13428-020-01433-0. 33

Kokil Jaidka, Alvin Zhou, and Yphtach Lelkes. Brevity is the Soul of Twitter: The Constraint Affordance and Political Discussion. *Journal of Communication*, 69(4):345–372, August 2019. ISSN 0021-9916, 1460-2466. doi: 10.1093/joc/jqz023. 27

Davy Janssen and Raphaël Kies. Online Forums and Deliberative Democracy. *Acta Politica*, 40(3): 317–335, September 2005. ISSN 1741-1416. doi: 10.1057/palgrave.ap.5500115. 27

Glen Joris, Frederik De Grove, Kristin Van Damme, and Lieven De Marez. News Diversity Reconsidered: A Systematic Literature Review Unraveling the Diversity in Conceptualizations. *Journalism Studies*, 21(13):1893–1912, October 2020. ISSN 1461-670X. doi: 10.1080/1461670X.2020.1797527. 57

Anne Kroon, Toni Van Der Meer, and Rens Vliegenthart. Beyond Counting Words: Assessing Performance of Dictionaries, Supervised Machine Learning, and Embeddings in Topic and Frame Classification. *Computational Communication Research*, 4(2):528–570, October 2022. ISSN 2665-9085. doi: 10.5117/CCR2022.2.006.KROO. 44, 45

Anne Kroon, Kasper Welbers, Damian Trilling, and Wouter van Atteveldt. Advancing Automated Content Analysis for a New Era of Media Effects Research: The Key Role of Transfer Learning. *Communication Methods and Measures*, 0(0):1–21, 2023. ISSN 1931-2458. doi: 10.1080/19312458.2023.2261372. 25, 29, 45

Thomas B. Ksiazek, Limor Peer, and Andrew Zivic. Discussing the News: Civility and hostility in user comments. *Digital Journalism*, 3(6):850–870, November 2015. ISSN 2167-0811, 2167-082X. doi: 10.1080/21670811.2014.972079. 27, 28, 33

Moritz Laurer. *Language Models as Measurement Tools: Using Instruction-Based Models to Increase Validity, Robustness and Data Efficiency*. PhD-Thesis - Research and graduation internal, October 2024. 30

Thomas Leeper and Rune Slothuus. Deliberation and Framing. In Andre Bächtiger, John S. Dryzek, Jane Mansbridge, and Mark Warren, editors, *The Oxford Handbook of Deliberative Democracy*, pages 555–572. Oxford University Press, September 2018. ISBN 978-0-19-874736-9. doi: 10.1093/oxfordhb/9780198747369.013.37. 56

Felicia Loecherbach, Judith Moeller, Damian Trilling, and Wouter van Atteveldt. The Unified Framework of Media Diversity: A Systematic Literature Review. *Digital Journalism*, 8(5):605–642, May 2020. ISSN 2167-0811. doi: 10.1080/21670811.2020.1764374. 55, 57

Tim Loughran and Bill Mcdonald. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65, 2011. ISSN 1540-6261. doi: 10.1111/j.1540-6261.2010.01625.x. 29

Jane Mansbridge, James Bohman, Simone Chambers, Thomas Christiano, Archon Fung, John Parkinson, Dennis F. Thompson, and Mark E. Warren. A systemic approach to deliberative democracy. In John Parkinson and Jane Mansbridge, editors, *Deliberative Systems*, pages 1–26. Cambridge University Press, 1 edition, July 2012. ISBN 978-1-139-17891-4 978-1-107-02539-4 978-1-107-67891-0. doi: 10.1017/CBO9781139178914.002. 16, 19

Microsoft. Data, privacy, and security for Azure OpenAI Service - Azure AI services, November 2024. URL https://learn.microsoft.com/en-us/legal/cognitive-services/openai/data-privacy. 34

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large Language Models: A Survey, February 2024. arXiv:2402.06196 version: 2. 30

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models, October 2024. arXiv:2410.05229 [cs]. 47

Ashley Muddiman and Natalie Jomini Stroud. News Values, Cognitive Biases, and Partisan Incivility in Comment Sections: Uncivil Comments. *Journal of Communication*, 67(4):586–609, August 2017. ISSN 00219916. doi: 10.1111/jcom.12312. 33, 45

Diana C. Mutz. Is Deliberative Democracy a Falsifiable Theory? *Annual Review of Political Science*, 11(1): 521–538, June 2008. ISSN 1094-2939, 1545-1577. doi: 10.1146/annurev.polisci.11.081306.070308. 16, 17, 18

Diana C. Mutz and Byron Reeves. The New Videomalaise: Effects of Televised Incivility on Political Trust. *American Political Science Review*, 99(1):1–15, February 2005. ISSN 1537-5943, 0003-0554. doi: 10.1017/S0003055405051452. Publisher: Cambridge University Press. 27

Kristinn Már and John Gastil. Do Voters Trust Deliberative Minipublics? Examining the Origins and Impact of Legitimacy Perceptions for the Citizens' Initiative Review. *Political Behavior*, August 2021. ISSN 0190-9320, 1573-6687. doi: 10.1007/s11109-021-09742-6. 16

Matti Nelimarkka, Jean Philippe Rancy, Jennifer Grygiel, and Bryan Semaan. (Re)Design to Mitigate Political Polarization: Reflecting Habermas' ideal communication space in the United States of America and Finland. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–25, November 2019. ISSN 2573-0142. doi: 10.1145/3359243. 16

Rishab Nithyanand, Brian Schaffner, and Phillipa Gill. Measuring Offensive Speech in Online Political Discourse. 2017a. URL https://www.usenix.org/conference/foci17/workshop-program/presentation/nithyanand. 28

Rishab Nithyanand, Brian Schaffner, and Phillipa Gill. Online Political Discourse in the Trump Era, November 2017b. arXiv:1711.05303. 28, 33

OpenAI. GPT-4o mini: advancing cost-efficient intelligence, 2024a. URL https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. 47

OpenAI. Pricing, 2024b. URL https://openai.com/api/pricing/. 34

OpenAI. Safety & responsibility, 2024c. URL https://openai.com/safety/. 34

Lisa Oswald. Automating the Analysis of Online Deliberation? A comparison of manual and computational measures applied to climate change discussions. *SocArXiv*, 2022. 16, 17

David Owen and Graham Smith. Survey Article: Deliberation, Democracy, and the Systemic Turn. *Journal of Political Philosophy*, 23(2):213–234, June 2015. ISSN 09638016. doi: 10.1111/jopp.12054. 19

Ryan J. Owens and Justin P. Wedeking. Justices and Legal Clarity: Analyzing the Complexity of U.S. Supreme Court Opinions. *Law & Society Review*, 45(4):1027–1061, 2011. ISSN 1540-5893. doi: 10.1111/j.1540-5893.2011.00464.x. 33

Zizi Papacharissi. Democracy online: civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society*, 6(2):259–283, April 2004. ISSN 1461-4448, 1461-7315. doi: 10.1177/1461444804041444. 31, 32

Eli Pariser. *The Filter Bubble: What the Internet is Hiding from You*. Penguin, New York, 2011. 24, 31

James W. Pennebaker, Cindy K. Chung, Molly Ireland, Amy Gonzales, and Roger J. Booth. *The Development and Psychometric Properties of LIWC2007*. LIWC.net, Austin, TX, 2007. URL https://www.liwc.net/LIWC2007LanguageManual.pdf. 33

Timothy Quinn. Introducing Hatebase: the world's largest online database of hate speech, 2020. URL https://thesentinelproject.org/2013/03/25/introducing-hatebase-the-worlds-largest-online-database-of-hate-speech/. 33

Myrthe Reuver, Nicolas Mattis, Marijn Sax, Suzan Verberne, Nava Tintarev, Natali Helberger, Judith Moeller, Sanne Vrijenhoek, Antske Fokkens, and Wouter Van Atteveldt. Are we human, or are we users? The role of natural language processing in human-centric news recommenders that nudge users to diverse content. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 47–59, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.nlp4posimpact-1.6. 17, 22

Filipe N. Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1):23, 2016. doi: 10.1140/epjds/s13688-016-0085-1. 25

Jay Rosen. 1. The People Formerly Known as the Audience. In Michael Mandiberg, editor, *The Social Media Reader*, pages 13–16. New York University Press, March 2012. ISBN 978-0-8147-6302-5. doi: 10.18574/nyu/9780814763025.003.0005. 27

Patrícia Rossini. Beyond Incivility: Understanding Patterns of Uncivil and Intolerant Discourse in Online Political Talk. *Communication Research*, 49(3):399–425, April 2022. ISSN 0093-6502, 1552-3810. doi: 10.1177/0093650220921314. 31

Ian Rowe. Civility 2.0: a comparative analysis of incivility in online political discussion. *Information, Communication & Society*, 18(2):121–138, February 2015a. ISSN 1369-118X, 1468-4462. doi: 10.1080/1369118X.2014.940365. 31

Ian Rowe. Deliberation 2.0: Comparing the Deliberative Quality of Online News User Comments Across Platforms. *Journal of Broadcasting & Electronic Media*, 59(4):539–555, October 2015b. ISSN 0883-8151. doi: 10.1080/08838151.2015.1093482. 31

David M. Ryfe. DOES DELIBERATIVE DEMOCRACY WORK? *Annual Review of Political Science*, 8(1):49–71, June 2005. ISSN 1094-2939, 1545-1577. doi: 10.1146/annurev.polisci.8.032904.154633. 31

Abel Salinas and Fred Morstatter. The Butterfly Effect of Altering Prompts: How Small Changes and Jailbreaks Affect Large Language Model Performance, April 2024. URL `http://arxiv.org/abs/2401.03729`. arXiv:2401.03729. 44

Lynn M. Sanders. Against Deliberation. *Political Theory*, 25(3):347–376, June 1997. ISSN 0090-5917, 1552-7476. doi: 10.1177/0090591797025003002. 18

Rüdiger Schmitt-Beck and Christiane Grill. From the Living Room to the Meeting Hall? Citizens' Political Talk in the Deliberative System. *Political Communication*, 37(6):832–851, November 2020. ISSN 1058-4609. doi: 10.1080/10584609.2020.1760974. 26

Stuart Soroka, Lori Young, and Meital Balmas. Bad News or Mad News? Sentiment Scoring of Negativity, Fear, and Anger in News Content. *The ANNALS of the American Academy of Political and Social Science*, 659(1):108–121, May 2015. ISSN 0002-7162. doi: 10.1177/0002716215569217. 25

Rosalynd Southern and Emily Harmer. Twitter, Incivility and "Everyday" Gendered Othering: An Analysis of Tweets Sent to UK Members of Parliament. *Social Science Computer Review*, 39(2):259–275, April 2021. ISSN 0894-4393. doi: 10.1177/0894439319865519. 31, 32

Jennifer Stromer-Galley. Measuring Deliberation's Content: A Coding Scheme. *Journal of Deliberative Democracy*, 3(1), July 2007. ISSN 2634-0488. doi: 10.16997/jdd.50. 27, 58

Jennifer Stromer-Galley, Magdalena Wojcieszak, Nicholas John, and Adrienne L Massanari. Introduction to the special issue of social media: the good, the bad, and the ugly. *Journal of Communication*, 73(3):193–197, June 2023. ISSN 0021-9916. doi: 10.1093/joc/jqad016. 26

Natalie Jomini Stroud. Polarization and Partisan Selective Exposure. *Journal of Communication*, 60(3): 556–576, September 2010. ISSN 0021-9916. doi: 10.1111/j.1460-2466.2010.01497.x. 31

Jesper Strömbäck. In Search of a Standard: four models of democracy and their normative implications for journalism. *Journalism Studies*, 6(3):331–345, August 2005. ISSN 1461-670X, 1469-9699. doi: 10.1080/14616700500131950. 15, 26

Cass Sunstein. *Republic.Com*. Princeton University Press, Princeton, NJ, 2001. 24, 31

Dennis F. Thompson. Deliberative Democratic Theory and Empirical Political Science. *Annual Review of Political Science*, 11(1):497–520, June 2008. ISSN 1094-2939, 1545-1577. doi: 10.1146/annurev.polisci.11.081306.070555. 16, 17, 19, 21

Petter Törnberg. Best Practices for Text Annotation with Large Language Models. *Sociologica*, 18(2): 67–85, October 2024a. ISSN 1971-8853. doi: 10.6092/issn.1971-8853/19461. 34, 35, 77

Petter Törnberg. Large Language Models Outperform Expert Coders and Supervised Classifiers at Annotating Political Social Media Messages. *Social Science Computer Review*, September 2024b. ISSN 0894-4393. doi: 10.1177/08944393241286471. 30

Wouter Van Atteveldt, Mariken A. C. G. Van Der Velden, and Mark Boukes. The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms. *Communication Methods and Measures*, 15(2):121–140, April 2021. ISSN 1931-2458, 1931-2466. doi: 10.1080/19312458.2020.1869198. 42, 43, 44

Marieke Van Hoof, Damian Trilling, Corine Meppelink, Judith Möller, and Felicia Loecherbach. Googling Politics? Comparing Five Computational Methods to Identify Political and News-related Searches from Web Browser Histories. *Communication Methods and Measures*, 0(0):1–27, June 2024. ISSN 1931-2458. doi: 10.1080/19312458.2024.2363776. 44

Bertie Vidgen and Leon Derczynski. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):e0243300, December 2020. ISSN 1932-6203. doi: 10.1371/journal.pone.0243300. 29

Paul S. Voakes, Jack Kapfer, David Kurpius, and David Shano-yeon Chern. Diversity in the news: A conceptual and methodological framework. *Journalism & Mass Communication Quarterly*, 73(3): 582–593, September 1996. ISSN 10776990. doi: 10.1177/107769909607300306. 58

Sanne Vrijenhoek, Mesut Kaya, Nadia Metoui, Judith Möller, Daan Odijk, and Natali Helberger. Recommenders with a Mission: Assessing Diversity in News Recommendations. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, pages 173–183, Canberra ACT Australia, March 2021. ACM. ISBN 978-1-4503-8055-3. doi: 10.1145/3406522.3446019. 15, 16, 17

Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. Argumentation Quality Assessment: Theory vs. Practice. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 250–255, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2039. 58

Maximilian Weber and Merle Reichardt. Evaluation is all you need. Prompting Generative Large Language Models for Annotation Tasks in the Social Sciences. A Primer using Open Models, December 2023. arXiv:2401.00284 [cs]. 54

René Weber, J. Michael Mangus, Richard Huskey, Frederic R. Hopp, Ori Amir, Reid Swanson, Andrew Gordon, Peter Khooshabeh, Lindsay Hahn, and Ron Tamborini. Extracting Latent Moral Information from Text Narratives: Relevance, Challenges, and Solutions. *Communication Methods and Measures*, 12(2-3):119–139, April 2018. ISSN 1931-2458, 1931-2466. doi: 10.1080/19312458.2018.1447656. 27

Kasper Welbers, Wouter van Atteveldt, Jan Kleinnijenhuis, and Nel Ruigrok. A Gatekeeper among Gatekeepers: News agency influence in print and online newspapers in the Netherlands. *Journalism Studies*, 19(3):315–333, February 2018. ISSN 1461-670X. doi: 10.1080/1461670X.2016.1190663. 58

Hartmut Wessler. *Habermas and the Media*. Polity Press, Cambridge, UK, 2018. ISBN 978-0-7456-5134-7. 20, 21, 22, 24, 49

Alvin Zhou, Wenlin Liu, Hye Min Kim, Eugene Lee, Jieun Shin, Yafei Zhang, Ke M. Huang-Isherwood, Chuqing Dong, and Aimei Yang. Moral Foundations, Ideological Divide, and Public Engagement with U.S. Government Agencies' COVID-19 Vaccine Communication on Social Media. *Mass Communication and Society*, 27(4):739–764, July 2024. ISSN 1520-5436. doi: 10.1080/15205436.2022.2151919. 32

Marc Ziegele, Oliver Quiring, Katharina Esau, and Dennis Friess. Linking News Value Theory With Online Deliberation: How News Factors and Illustration Factors in News Articles Affect the Deliberative Qual-

ity of User Discussions in SNS' Comment Sections. *Communication Research*, 47(6):860–890, August 2020. ISSN 0093-6502, 1552-3810. doi: 10.1177/0093650218797884. 31, 32

Frederik J. Zuiderveen Borgesius, Damian Trilling, Judith Möller, Balázs Bodó, Claes H. De Vreese, and Natali Helberger. Should we worry about filter bubbles? *Internet Policy Review*, 5(1), March 2016. ISSN 2197-6775. doi: 10.14763/2016.1.401. 19

twon

twin of online social networks

Contact us

Damian Trilling
Project Coordinator

+31 62 782 7904

d.c.trilling@uva.nl

University of Amsterdam
Postbus 15791
1001 NG Amsterdam

Funded by
the European Union