Network Science (2025), **13**, e4, 1–23
doi:10.1017/nws.2025.4

RESEARCH ARTICLE

Automated extraction of discourse networks from large volumes of media data

Mario Angst, Neitah Noemi Müller and Viviane Walker

Digital Society Initiative, University of Zürich, Zürich, Switzerland
Corresponding author: Mario Angst; Email: mario.angst@uzh.ch

Abstract

Understanding and tracking societal discourse around essential governance challenges of our times is crucial. One possible heuristic is to conceptualize discourse as a network of actors and policy beliefs.

Here, we present an exemplary and widely applicable automated approach to extract discourse networks from large volumes of media data, as a bipartite graph of organizations and beliefs connected by stance edges. Our approach leverages various natural language processing techniques, alongside qualitative content analysis. We combine named entity recognition, named entity linking, supervised text classification informed by close reading, and a novel stance detection procedure based on large language models.

We demonstrate our approach in an empirical application tracing urban sustainable transport discourse networks in the Swiss urban area of Zürich over 12 years, based on more than one million paragraphs extracted from slightly less than two million newspaper articles.

We test the internal validity of our approach. Based on evaluations against manually automated data, we find support for what we call the window validity hypothesis of automated discourse network data gathering. The internal validity of automated discourse network data gathering increases if inferences are combined over sliding time windows.

Our results show that when leveraging data redundancy and stance inertia through windowed aggregation, automated methods can recover basic structure and higher-level structurally descriptive metrics of discourse networks well. Our results also demonstrate the necessity of creating high-quality test sets and close reading and that efforts invested in automation should be carefully considered.

Keywords: discourse; networks; automated text analysis; urban; sustainability; natural language processing; large language models

1. Introduction

Societies across the world face many challenges at the dawn of the Anthropocene (Lewis and Maslin, 2015; Steffen et al., 2015). Examples include finding ways for broad societal transformations toward sustainability, dealing with the consequences of digitalization or guiding rapid urbanization. Finding answers to these challenges requires governance (Lubell and Morrison, 2021). Governance happens in networks of organizational actors from all societal sectors, including civil society, the private sector, government, and scientific institutions (Rhodes, 1996).

Topics related to governance challenges are often discursively contested. Actors in governance networks constantly need to make choices on how to position themselves in relation to various topics in public discourse. They may want to raise the profile of certain topics and downplay others (agenda-setting). Sometimes actors are forced or sometimes interested in publicly taking a stance concerning key policy issues. Doing so, they can be seen to create steadily evolving *discourse networks*. Discourse networks provide an essential backdrop to material policymaking and governance activity.

From a discourse network perspective, many aspects of discourse can be formally represented as a graph and analyzed using network analysis tools. In this article, we specifically present an exemplary and widely applicable *automated* approach to extract discourse networks from large volumes of media data.

Why would we go to great lengths to develop tools for understanding and keeping track of societal discourse? On the one hand, for the social sciences, the descriptive and theoretical understanding of how and why societal discourse develops over time is a value in itself. It is a way in which the social sciences fulfill their function of providing societal self-reflection. On the other hand, for practice, an understanding of societal discourse is important for actors involved in governance and society at large to understand where possibilities for individual as well as collective action lie.

At the beginning of this article, we already want to stress however that a network perspective is not necessary to understand societal discourse. There are other, not specifically relational ways to understand societal discourse. Most people perform these effortlessly, for example, when reading a newspaper. However, an explicit network perspective on discourse around crucial governance challenges can be a useful *heuristic* tool to highlight and track properties of societal discourse at multiple scales.

The article is structured as follows. First, we outline our theoretical framework. We build on the discourse network analysis (DNA) (Leifeld, 2020) literature and combine it with further insights from policy and governance network theory to arrive at a heuristic to understand discourse as a network with specific, necessary components amenable for automated analysis. We also introduce the central hypothesis of this article, which we call the window validity hypothesis of automated discourse network data gathering, and briefly introduce some exemplary, illustrative network metrics with specific meanings in the context of our graph representation of discourse.

Second, we outline the methods used in an empirical, proof-of-concept application of our automated approach for the specific case of sustainable transport discourses in the urban area of Zürich. We analyze more than one million newspaper articles over a period of twelve years. To do so, we introduce a multi-stage pipeline building on various natural language processing (NLP) tools, alongside qualitative content analysis, which is broadly applicable to many different governance domains. We also introduce the methods used to combine the data created in our empirical study with manually annotated test data to test the window validity hypothesis of automated discourse network data gathering.

Third, we present the results of training and applying our pipeline and present the results of our central hypothesis test. We discuss the limitations of our approach and explore the question of when efforts to create automated data gathering approaches are actually worth it. In the conclusion, we expand on potential use cases and future improvements, which especially include the need for principled mixed method validation approaches.

2. Theoretical framework: discourse as a network

Our starting point for outlining our specific heuristic to understanding societal discourse as a network owes most to the DNA literature started in Leifeld (2009, 2013). Leifeld (2020), in the most recent review of the literature, defines DNA as a combination of network analysis and qualitative content analysis. However, we will assume (in line with Leifeld (2020)) that the heuristic for understanding discourse as developed in the DNA literature is not epistemologically bound to qualitative content analysis. The interesting feature of understanding discourse from a network perspective developed in the DNA literature is breaking societal discourse down into constituent elements, amenable to representation in a graph with a set of fixed components.

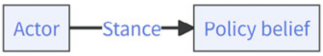


Figure 1. Basic, necessary components of a discourse network as conceptualized in this study.

2.1 Components of a discourse network

When we talk of discourse networks in the following, we require these networks to have three necessary components: actors, beliefs, and stances (see Figure 1).

First, *actors* are the agents in any discourse network. Actors can be individual or organizational actors participating in societal discourse around a specific governance challenge. Analyzing governance more generally (thus, in terms of material and discursive action) centered around an understanding of governance as a network of (organizational) actors follows a governance definition tracing its lineage back at least to Laumann and Knoke (1987), via Rhodes (1996), and has become a predominant paradigm in governance network research (Scott and Ulibarri, 2019).

Second, *beliefs* are sets of normative statements related to salient aspects of a discourse around a governance challenge. We follow existing theory on three types of beliefs in policy systems in the Advocacy Coalition Framework (ACF) (Sabatier, 1988; Weible, et al., 2009), which are so-called deep core beliefs, policy beliefs, and secondary aspects. The three types of beliefs form a hierarchical belief system. Deep core beliefs, on the highest level, encompass the fundamental, general values actors hold and rely on for decision-making. Deep core beliefs are usually quite broad (e.g., general beliefs about the role the state should play in society) and are relevant for an actor beyond a specific policy field. Policy core beliefs translate deep core into a specific policy field and inform actors' preferences about still relatively general directions policymaking in a field should take (e.g., for the governance of sustainable transport systems, this could refer to relative weights given to public versus private means of transport). Secondary aspects are detailed further translations of policy core beliefs. They are about the specific means through which policy core beliefs should be achieved (e.g., for the governance of sustainable transport systems, this could cover the position of an actor toward a specific type of subsidy for electric vehicles). Depending on the specific research question, time frame studied, and case context, discourse network studies might focus on different types of beliefs.

Third, *stances* are expressed, qualified relations of actors to beliefs. For our purposes, we treat stances as purely relational objects, which cannot exist without a subject expressing them (actors) and an object they refer to (belief statements). Stances need a stance qualifier, most likely a member of a fixed set of possible stance qualifiers Q, such as $Q = \{support, opposition, neutral\}$. For every stance qualifier in Q, a set of edges is added to the graph. Stances can occur at different aggregations on a temporal scale. What this means is that in capturing discourse in a network representation, on the one side of the spectrum, stances can represent a more general, broader orientation toward a belief of an actor expressed over a period of time, and there might be only a single stance relation related to each belief present in a discourse. Data for such discourse networks might, for example, be captured through qualitative interviews at a single point in time. Or, on the other side of the spectrum, stance relations can capture observed expressions of an actor at a fine-grained temporal resolution, leading to a dynamic network over time with n_i time-stamped stances for every observed expression of a stance per actor i over an observed time frame. Data for such a network might, for example, be captured by means of participatory observation or, as introduced in this article, through document analysis. Reasonably, for every observation window, an actor can only express one stance out of Q, meaning they cannot, for example, support and oppose a statement at the same time. Figure 2 gives a minimal empirical example for the basic structure described.

https://doi.org/10.1017/nws.2025.4 Published online by Cambridge University Press

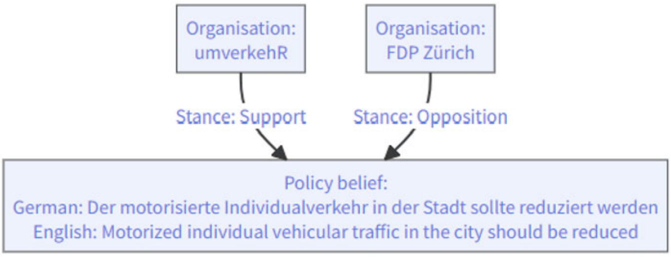

Organisation: umverkehR
Organisation: FDP Zürich
Stance: Support
Stance: Opposition
Policy belief:
German: Der motorisierte Individualverkehr in der Stadt sollte reduziert werden
English: Motorized individual vehicular traffic in the city should be reduced

Figure 2. Example of a discourse network with two actors (organizations), two qualified stance relations, and one policy belief around urban sustainable transport governance.

2.2 Formal graph structure

Formally, given the discussion of discourse network components, a discourse network $G = (A, B, S)$ is an edge-labeled, undirected, bipartite graph, where

- A is a set of actor nodes $A = \{a_1, a_2, \ldots, a_i\}$
- B is a set of belief nodes $B = \{b_1, b_2, \ldots, b_i\}$
- S is a set of stance edges, which are each labeled with a stance classifier out of the set $Q = \{q_1, q_2, \ldots, q_i\}$

2.3 Automating discourse network data gathering

Traditionally, DNA has often relied on small datasets of high-quality, manual annotations of textual documents, such as newspaper articles (Leifeld, 2013) or social media posts (Bossner and Nagel, 2020). Research on developing more automated discourse network data gathering approaches has been increasingly suggested as a crucial component of the overall research discourse network research agenda in both agenda-setting articles (Leifeld, 2020) and empirical applications (Kammerer and Ingold, 2023).

For the data domain of textual documents, the obvious candidate for an automated approach is automated text analysis through NLP. Automated text analysis and machine learning approaches in general have recently become much more available to a much broader set of researchers than before, especially in the social sciences (Grimmer, et al., 2021) and recent mainstreaming of advances in NLP technologies such as large language models (LLMs) are likely to continue this trend.

Automation of discourse network data gathering should always be understood as complementary to qualitative approaches. There are three reasons for this. First, automation by design trades some of the quality and in-depth understanding of discourse that can realistically only be gained from close, qualitative engagement with (especially textual) data (Carlsen and Ralund, 2022) for the possibility of larger temporal and thematic scope. Second, the identification of the main empirical components of discourse in empirical practice, in our conceptualization, thus the set of allowed actors, beliefs, and stance qualifiers, needs qualitative understanding and engagement with the data and problem domain. Third, automated approaches need an evaluation against “ground truth,“ which for the domain of societal discourse can only mean the careful, manual annotation of test sets or validating results in exchange with actors active in discourse, as we will argue in this article.

A hybrid, semi-automated approach to discourse network data gathering, combining machine learning methods and manual annotation is presented in Haunss et al. (2020). The approach presented in their study is semi-automated in the sense that machine learning is mainly explored for its usefulness in supporting manual annotation. In such a supporting application, the use of machine learning resulted in slight increases in annotation speed and quality, although at an overall level, time gains may be offset to some degree by the time needed to train the machine learning model. Further, results indicated that the approach was mostly useful to reconstruct the most engaged core of discourse networks but struggled with more peripheral components of the network.

Recently, Ceron et al. (2024) also presented the results of an evaluation of automated text analysis tools for understanding political discourse. They compare automated approaches that (a) create a discourse network from fine-grained, time-stamped newspaper data with (b) automated approaches to understand discourse mainly from the ideological classification of party manifestos. Except for them calling what we call stance claims, their automated discourse network data gathering approach, which is most relevant in the context of this article, follows a discourse network conceptualization basically identical to ours (Ceron et al., 2024, 73). Here, their study innovates significantly in two major challenge areas for automated discourse network data gathering. First, it introduces a well-performing hybrid pipeline for the robust identification of actors (entity linking or canonization), combining a conditional random fields model with an LLM-based procedure (Ceron et al., 2024, 80). Second, evaluating against a gold-standard discourse network dataset based on newspaper articles, (Ceron et al., 2024, 75) show good performance of a sentence-Bidirectional encoder representations from transformers (BERT)-based stance detection classifier, which applies transfer learning techniques, using existing annotation on actor stances in nuclear energy debates to pretrain a model predicting stances on migration debates.

Here, we build on the pioneering works by Haunss et al. (2020) and Ceron et al. (2024) on three fronts. First, we illustrate and conceptualize automated discourse network data gathering at a larger scale, both in terms of breadth of empirical data sources and temporally. Second, due to our focus on a larger scale, we present a slightly more fully automated approach and a full end-to-end processing pipeline. Third, also due to our focus on a larger scale automation, we present a slightly differing approach in terms of the analysis pipeline, staggering qualitative and automated procedures instead of combining them. Altogether, these developments should contribute to broadening the palette of empirical data gathering suggested for applied DNA research on the spectrum from fully manual to fully automated.

2.4 The window validity hypothesis of automated discourse network data gathering

Previous work on automating DNA has made an interesting initial finding suggesting that automated DNA can still provide internally valid data even in the presence of performance issues for individual automated components. For example, even somewhat flawed classifiers were able to recover core network components (Haunss et al., 2020).

We suggest to subsume these findings under what we would call the *window validity hypothesis* of automated discourse network data gathering.

> Window Validity Hypothesis: Given an unbiased automated discourse network data gathering procedure, combining inferences over a time window increases the internal validity of a discourse network representation.

There are two main theoretical mechanisms leading us to formulate this hypothesis, given the data gathering process used to construct a discourse network. First, automated approaches may profit from *redundancy in data points*, for instance, in media reporting on discourse (Haunss et al., 2020). For example, redundancy in data on stances means that media data on relatively salient discourse reports often contains multiple data points about the stance of an actor during a given time window. Or, if social media posts are processed as raw data to construct a discourse network, actors are increasingly likely to express a stance multiple times as an observation window grows. Second, *stance inertia* in discourse networks means that if networks capture stances, especially regarding policy or deep core beliefs in ACF parlance, the nature of these beliefs rules

https://doi.org/10.1017/nws.2025.4 Published online by Cambridge University Press

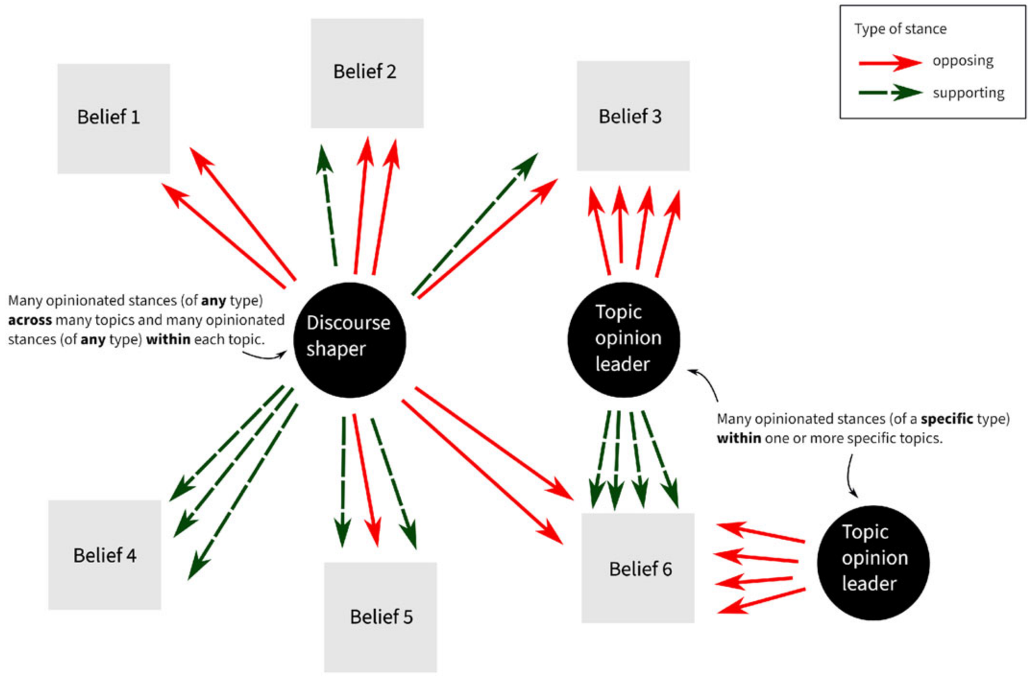

Type of stance
opposing
supporting
Belief 1
Belief 2
Belief 3
Belief 4
Belief 5
Belief 6
Discourse shaper
Topic opinion leader
Topic opinion leader
Many opinionated stances (of **any** type) **across** many topics and many opinionated stances (of **any** type) **within** each topic.
Many opinionated stances (of a **specific** type) **within** one or more specific topics.

Figure 3. Stylized, illustrative example of discourse shaper and topic opinion leader positions within discourse network.

out rapid fluctuations (Weible, et al., 2009). Given such inertia, redundancy in data points will lead an unbiased stance classifier to converge to a valid discourse network representation, even if it faces performance issues within smaller observation windows. As pointed out by a reviewer of this article, this procedure to an extent also resembles the partial pooling that is achieved in Bayesian multi-level models (McElreath, 2016, 408).

2.5 Mapping of structurally descriptive network metrics to properties of discourse

We present three exemplary structurally descriptive metrics in the following, which we suggest to hold specific meanings in our conceptualization of a discourse network. On the actor level, these are discourse shapers and topic opinion leader metrics (see Figure 3). On the network or discourse macro-level, we suggest a measure for ideological alignment across topics over time. These sets of metrics illustrate the possibilities of relational data as conceptualized in this article for understanding discourse, but first and foremost, we will employ these measures as additional comparison metrics for testing the window validity hypothesis, checking if automated discourse network data gathering can recover them.

When we talk of structurally descriptive measures, we follow a helpful conceptualization of governance network analysis modes developed in Scott and Ulibarri (2019) in the following. Generally, three modes of governance network analysis can be distinguished, which are implicit, structurally descriptive, and structurally explicit. Structurally descriptive governance network analysis focuses on the description and comparisons of networks on different levels from characteristics of network embedding of different actors to network-level metrics. Structurally explicit methods focus on the analysis of specific ties in networks involving governance actors and their drivers. DNA as conceptualized in this article allows for both structurally descriptive and structurally explicit analysis. Implicit analysis as the qualitative assessment of the consequences of the presence of a network is a non-applicable mode as the presence of a network is a precondition for a discourse analysis to take place.

https://doi.org/10.1017/nws.2025.4 Published online by Cambridge University Press

2.5.1 Discourse shapers

Actors vary regarding their activity in discourses, both quantitatively and qualitatively. An interesting category of actors in multi-faceted discourses containing a variety of topics are actors that are shaping overall discourse network structure both within and beyond single topics. Such discourse shapers are close analogues to bridging actors or brokers in the governance network literature on more material governance activity (Angst et al., 2018).

We suggest a variant of z- and c-score computations to identify discourse shapers. Z- and c-scores were originally introduced for ecological networks to measure the degree to which species provide connections in ecological networks, differentiating between species that connect within communities (high z-scores), between communities (high c-scores), or do both (so-called hubs) (Olesen et al., 2007). As originally proposed for unipartite networks with community assignments for nodes, the measures are essentially standardized within-community versus between-community degree counts for nodes.

In the case of discourse networks, we suggest to compute z-scores as standardized counts of opinionated stance edges for every policy belief and c-scores as standardized counts of opinionated stance edges across all policy beliefs. By opinionated stance edges, we mean stance edges expressing a non-neutral stance of an actor toward a policy belief. Discourse shapers are then actors scoring both high on z- and c-scores. This captures the intuition of a discourse shaper as a vocal actor expressing stances across different topics *and* shaping discourse also within topics they are engaged in.

2.5.2 Topic opinion leaders

For every policy belief and associated discourse topic, some actors are most vocal in expressing opinionated stances. Such topic-specific opinion leaders may be in the minority or majority regarding a specific policy belief. For discourse networks, we suggest to identify opinion leader scores by calculating z-scores as standardized counts for every opinionated stance qualifier regarding a specific policy belief.

2.5.3 Polarization and ideological alignment across topics

The extent, development, and effects of polarization in public discourse, alongside political polarization in general (Levin, et al., 2021), have been a much discussed topic, both on the level of individual citizens in research on so-called "filter bubbles" (Zuiderveen Borgesius et al., 2016) but also specific to actors in governance networks (Gronow et al., 2020; Angst and Brandenberger, 2022).

A discourse representation, especially using time-stamped data, allows to leverage some relational properties of the data to measure polarization and its evolution over time. A relatively straightforward measure to do so is to track the evolution of bipartite closure for opinionated stances across topics over time. For example, this can be approached through counting network motifs (Bodin and Maria, 2012) such as four cycles including two actors with matching edge sets within a time window. Increasing such closure indicates ideological alignment of actors over topics, which could in turn be an indication of higher-level polarization.

3. Methods

3.1 Case study: urban sustainable transport discourse in Zürich

We demonstrate a proof of concept of our approach to the automated extraction of discourse networks from textual data and test our main hypothesis on the window validity of automated discourse network data gathering in a case study on urban sustainable transport discourse in Zürich. The UN Agenda 2030 and its associated sustainable development goals (SDGs)

https://doi.org/10.1017/nws.2025.4 Published online by Cambridge University Press

(United Nations General Assembly, 2023), the latest in a series of more or less successful, high-level normative documents on how and where to achieve a more sustainable future on this planet, put an unprecedented emphasis on the role of cities in achieving a sustainable future for all (Patel et al., 2017). A key aspect of urban sustainability transformations across cities lies in increasing the sustainability of urban transport systems. At its core, a sustainable transport system combines a decoupling from resource use with considerations for societal equity (Hull, 2008).

For example, in the SDG urban sustainability target 11.2 (SDG targets are slightly more concrete formulations of general sustainable development goals), sustainability in urban transport systems is understood to demand a transport system that combines accessibility, affordability, safety, and a low environmental footprint. While these overall goals are less contested in societal discourse, how to achieve them through concrete policymaking action (thus their implementation) is a different story. Policy action around transportation, mobility, or traffic issues is hotly contested in societal discourse across the world (Shrestha, et al., 2024). A recent, somewhat extreme cases in point are various conspiracy theories surrounding the so-called "15 minute city" concept (Marquet et al., 2024), but there likely is no city in the world where transport policy is not continuously debated in the public sphere to some degree. This is unsurprising as action on transport systems affects the daily life of large proportions of the urban (and suburban) population and shapes cities much beyond the transport system, with knock-on effects on, for example, land valuation, gentrification, or employment opportunities.

The importance of urban transport for achieving sustainable development goals, combined with its salience in societal discourse, makes discourse surrounding urban sustainable transport policy a well-suited domain for demonstrating an application of automated discourse network extraction. In our case study, we focus specifically on the Swiss urban area of Zürich. Zürich is no exception among cities inasmuch policymaking in the transport system has been hotly debated for decades. Beyond this, there is not much that specifically qualifies or disqualifies Zürich as a proof of concept for demonstrating our general approach and testing our main hypothesis in the context of this article. Zürich happens to be the place we are substantively interested in, where our funding for this particular project comes from, and where we live. Our approach, metrics, and central hypothesis test should generalize well enough that we could have applied it to any other city.

3.2 Data

We analyze a dataset of 1,921,681 media articles written in German from the year 2012 to April 14, 2024, from three major Swiss news publications (Tages-Anzeiger, Neue Zürcher Zeitung, 20 Minuten). We chose publications with an emphasis on covering Zürich, which we understood as having dedicated sections for local news about Zürich. In their political orientation, based on Udris (2023), our media sources range from the more right-leaning, economically liberal Neue Zürcher Zeitung (NZZ) to the moderately left-leaning Tages-Anzeiger (see detailed list of media sources in appendix). Our data sources include both offline and print articles. Media data was made available to us via SwissdoxLiRI by the Linguistic Research Infrastructure of the University of Zurich.

3.3 Processing pipeline

Our pipeline starts by identifying Zürich-related articles. For an article to be judged Zürich-related, it either needed to appear in a newspaper-specific local Zürich news section or match one of a list of regular expressions containing place names for Zürich. This initial subsetting led to a reduction of articles to 139,473 articles.

We analyze media content on the paragraph level for most purposes in our processing pipeline. In a second step, we thus split the Zürich-related articles into paragraphs, which results in

https://doi.org/10.1017/nws.2025.4 Published online by Cambridge University Press

1,053,549 paragraphs. To split articles into roughly even-sized paragraphs, paragraph breaks occurring in the article were used, leading to a median paragraph length of 52 words.

The processing pipeline then employs a number of NLP techniques to process the resulting Zürich-related paragraphs.

3.3.1 Sustainable urban transport filter

First, in a relevance filter step, we apply supervised machine learning to classify paragraphs to determine whether they are related or not to the broader discourse on urban sustainable transport (set up as a binary, supervised classification task). To train the machine learning classifier, a team of four annotators first annotated a total of 4320 paragraphs out of the overall corpus iteratively over five batches according to a codebook. The full codebook specification can be seen in the appendix. Shortly, the key criteria for relevance to urban sustainable transport leans on the definition of SDG 11.2 "Sustainable Transport" in the UN Agenda 2030 (United Nations General Assembly, 2015), requiring a direct mention of at least one of safety of transport, environmental impact of transport, accessibility of transport, or affordability of transport.

An initial "batch 0" was used to get annotators acquainted and finalize an initial codebook draft. All annotation examples used in training were annotated independently by at least two annotators, and mismatches in annotations were all reviewed in the larger team during review sessions for each batch. We then set aside a randomly sampled test set of 20% and trained a binary text classifier on the remaining data based on a cased German BERT transformer model[1] in the Python NLP framework spacy (Montani et al., 2022). Our classifier achieves an f1-score of 0.84, a precision of 0.88, and a recall of 0.81 on the test set.

3.3.2 Topic classification

The initial sustainable urban transport filter step, described above, identifies paragraphs that broadly deal with the sustainable transport domain. For an analysis of discourse, this initial filter is still too broad however, as sustainable urban transport, as defined in our approach, is a very multidimensional domain. In a second step, we thus identify specific topics in the sustainable transport policy discourse from a closed set of topics. To do so, we use a rule-based model, mainly relying on regex matches on keywords or a set of keywords tested for high precision, to classify a paragraph as containing zero or more specific urban sustainable transport topics. We chose a rule-based classifier in this step specifically for speed and precision. The model allows for multiple topics to occur in a paragraph.

To identify the overall set of relevant sustainable transport discourse topics in the first place, we relied on qualitative content analysis. Initially, a set of seed articles was chosen based on a very explicit mention of sustainable mobility ("nachhaltige Mobilität") and closely read to develop a feel for language and vocabulary within that topic. An initial set of keywords from close reading was then used to select a set of articles for open coding in the software Atlas.ti. In a first phase, the articles were read, and all possible keywords were assigned so-called in Vivo codes as quotations until, after eight articles and 190 quotations, many in Vivo codes were repeated and no outstanding new ones could be generated. Thematically similar in Vivo codes were then summarized into 29 codes. Finally, further clusters were formed with these codes by reviewing the resulting code list with their quotations and bundling them into seven thematic code groups.

Code groups were named according to their focus—for example, the generated code "Tempo-30" and "Geschwindigkeit reduzieren" (transl.: speed reduction) formed the code group "Fahrgeschwindigkeit“ (transl.: driving speed). The names of these generated code groups resulted in a first set of topics and associated keywords for each topic. In a next step, a team of annotators checked keyword matches on sustainable transport-related paragraphs in our dataset from this

https://doi.org/10.1017/nws.2025.4 Published online by Cambridge University Press

Table 1. Sustainable transport topics and associated main policy belief used in the processing pipeline

	Topic (German)	Topic (English)	Main policy belief (German)
1	Motorisierter Individualverkehr	Motorized private transport	Der motorisierte Individualverkehr (MIV) in der Stadt soll reduziert werden.
2	öffentlicher Verkehr	Public transport	Die Nutzung und der Stellenwert des öffentlichen Verkehrs soll gefördert werden.
3	Flugverkehr	Air travel	Der Flugverkehr soll reduziert werden.
4	Parkplatz	Parking	Das Parkplatzangebot in der Stadt soll reduziert werden.
5	Fahrradinfrastruktur	Cycling infrastructure	Das Fahrrad als Mobilitätsform soll gefördert werden.
6	E-Mobilität	E-mobility	E-Mobilität in Form von E-Autos, E-Bussen, E-Scootern und E-Bikes soll gefördertwerden.
7	Fahrgeschwindigkeit	Driving speed	Zur Minderung von Emissionen soll die Fahrgeschwindigkeit in der Stadt reduziertwerden.

initial set on coherence with topics to create a final set of single- and multi-word keyword rules with improved precision for the rule-based classifier.

Every topic identified was also assigned a single main policy belief, referring to the most salient point of contention per topic, which was also derived during qualitative content analysis. The full set of topics and policy beliefs can be found in Table 1.

3.3.3 Identification of organizational actors

In a third step, we use a named entity recognition (NER) classifier, combined with entity linking, to identify organizational actors occurring in paragraphs. For the NER step, we utilized a general German NER classifier provided by spacy, pretrained on German news articles[2] to identify organizations occurring in text. For entity linking, where named entities in text are associated with unique organizations (for example linking the entities “the Greens“ and “Green Party“ recognized during NER both to the same organization “Green Party“), we created a large set of regex rules, based on working through results from raw NER classification runs, as well as integrating organizations found during close reading of paragraphs, for example, during annotation.

3.3.4 Stance detection

In a fourth step, we use a pre-trainedLLM-based stance detection procedure in a zero-shot approach to classify stances of detected actors regarding the main policy beliefs for topics detected in a paragraph. This means that for every combination of actor and topic detected in a paragraph, we classified the stance of the actor to the main policy belief for the topic, based on the paragraph text. Thus, every paragraph may be processed multiple times, focusing on different actors and beliefs, in the case of multiple topics and actors occurring in a paragraph.

The full approach for developing the LLM-based stance classifier, including an evaluation of alternative prompting methods to the one described here and comparisons of differently sized and pre-trained LLMs, is described in more detail at https://doi.org/10.5281/zenodo.14795490. The approach is also implemented for generalized usage on German text (not specific to our domain) in an open source Python package (`stance-llm`[3]). Briefly, to develop a reasonably performant prompting method, we follow Lan et al. (2023), who propose prompting multiple LLMs

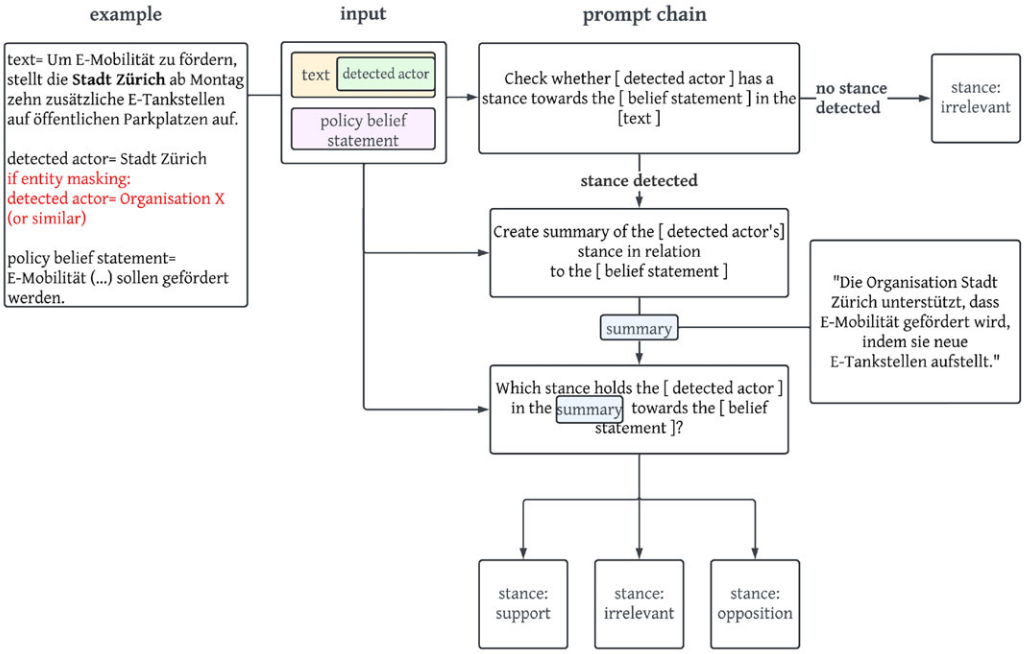

example
input
prompt chain
text= Um E-Mobilität zu fördern, stellt die **Stadt Zürich** ab Montag zehn zusätzliche E-Tankstellen auf öffentlichen Parkplatzen auf.
detected actor= Stadt Zürich
if entity masking:
detected actor= Organisation X
(or similar)
policy belief statement= E-Mobilität (...) sollen gefördert werden.
text
detected actor
policy belief statement
Check whether [detected actor] has a stance towards the [belief statement] in the [text]
no stance detected
stance: irrelevant
stance detected
Create summary of the [detected actor's] stance in relation to the [belief statement]
summary
"Die Organisation Stadt Zürich unterstützt, dass E-Mobilität gefördert wird, indem sie neue E-Tankstellen aufstellt."
Which stance holds the [detected actor] in the summary towards the [belief statement]?
stance: support
stance: irrelevant
stance: opposition

Figure 4. Example of a hierarchical prompt chain structure as employed for automated stance detection. A shortened, stylized English version of the original German prompts is shown. The example text is fictional.

exchanging text generated from different prompts modeled on roles of collaborating actors in a fixed order. We combine this with insights gained by Roy et al. (2022), who applied multiple binary classifiers (CNNs and LSTMs) in a hierarchical structure for stance detection to mitigate problems of class imbalance in stance detection tasks (in our case especially relevant to the likely imbalance between small opposition and large irrelevant class sets).

The prompt chain we utilize is illustrated in Figure 4. It starts with a check on whether there is a stance in an input text by a detected actor at all. If not, the irrelevant class is assigned. If yes, the LLM is further prompted to generate a summary of the input text of the detected actor's position in relation to the policy belief. In the last step of the prompt chain, the LLM is finally prompted to generate text starting a sentence based on this summary with a constrained set of options mapped to support, opposition, and irrelevant classes, which are then assigned as the stance prediction, based on the generated text. Given the nature of LLMs as stochastic text generation models with training data from a variety of sources, one potential bias in applying an LLM for stance detection in this task is the inclusion of well-known entities likely to occur in the training data of the LLM. We guarded against this with entity masking, substituting the string “Organization X” into the input text passed to the classification instead of the actual entity.

To evaluate the LLM-based stance detection component, a team of three annotators annotated 1256 stances of detected entities regarding a policy belief related to a detected topic occurring in sustainable transport-related paragraphs in our dataset iteratively over four batches according to a codebook. The full codebook specification used can be found in the appendix. An initial “batch 0” was used to get annotators acquainted and finalize an initial codebook draft. All annotation examples used in training were annotated independently by at least two annotators and mismatches in annotations were all reviewed in the larger team during review sessions for each batch.

To ensure a degree of reproducibility and due to ethical, copyright, and data protection concerns, we relied on a so-called open source LLM, which was available publicly. The term open source is debatable for many LLMs labeled such, given the often closed nature of training data.

But for research purposes, these models are an improvement over relying on external Application programming interface (API) calls to companies offering so-called AI services, which we would advise against. Crucially, so-called open source models offer a degree of reproducibility, making it possible to point to a publicly accessible model used at the time of analysis. This may be especially important compared to relying on constantly updated closed source models. Among other problems, these expose research to risks regarding hard-to-detect patterns of degradation in output quality (Shumailov et al., 2024).

For actual inference and evaluation runs, we self-hosted the LLM on university infrastructure. For the analysis presented here, we specifically used Kafka 7B,[4] a German LLM based on LeoLM,[5] which is again based on a so-called open source LLM named Mistral 7B originally released by the company Mistral AI. The best scoring prompt chain for Kafka7b with entity masking, which we used in this analysis, achieved a weighted f1-score of 0.60, precision of 0.64 and recall of 0.60, although performance across the stance classes was uneven, with the model generally performing well on the support and irrelevance class, and less well on opposition. Still, these scores are an improvement over the scores reported for claim classification in Haunss et al. (2020), the next most similar, although by no means identical, example for the task we tested for in the literature. Given our evaluations, the individual predictions of our pipeline for *individual stances* should be treated with caution. However, our evaluations also suggest that the results of our pipeline can be considered somewhat trustworthy for stances on average.

3.3.5 Final output

The final output of our processing pipeline is a time-stamped, edge-labeled, undirected, bipartite actor-belief graph as described in the section “Formal Graph Structure“.

Figure 5 gives an overview of the main components of the pipeline.

3.4 Testing the window validity hypothesis

To test the window validity hypothesis, we evaluate if aggregating stances created through our processing timeline over sliding time windows can recover discourse network structure and its evolution based on manual annotations.

We construct our reference network for testing based on the intersection of (1) the discourse network resulting from our automated processing with (2) a discourse network constructed from fully reviewed 1710 manually annotated stances of organizations (annotated using the same procedure and codebook as in training the LLM-based stance detection classifier) regarding main policy beliefs for motorized public transport, air travel, parking, E-mobility, and driving speed topics.

To combine inferences for a time window, we keep the most prevalent stance qualifier for every unique actor-belief edge present in the data. Thus, for an actor a_i, belief b_i and a set of sets of stance edges with k qualifiers $S_{a_i,b_i} = \{S_{q_{a_i,b_{i1}}}, S_{q_{a_i,b_{i2}}}, \ldots, S_{q_{a_i,b_{ik}}}\}$, we apply the function $m = max(\{|S_{q_{a_i,b_{i1}}}|, |S_{q_{a_i,b_{i2}}}|, \ldots, |S_{q_{a_i,b_{ik}}}|\})$ and only keep the edge qualifier k where $m = |S_{q_{a_i,b_{ik}}}|$. We compare network recovery over four different aggregation windows (12, 24, 48, and 96 months).

As test metrics to measure network recovery on the edge level, we evaluate precision, recall, and f1-scores on the predicted versus annotated edge sets. The testing setup here is one of testing a multiclass classifier where the predicted network edges can be one of the three stance qualifiers support, opposition, or irrelevant, and the true classes in the annotated test set are either support or opposition. We calculate both macro- and micro-level averaged test metrics. Macro-level multiclass averaging, as implemented in the R package `yardstick` (Kuhn, et al., 2024) calculates the metric as a combination of simple binary evaluations of the evaluated class against all others and then averages the evaluations with equal weight. Micro-level averaged metrics are calculated in a single evaluation, which may tend toward emphasizing the weight of larger classes.

https://doi.org/10.1017/nws.2025.4 Published online by Cambridge University Press

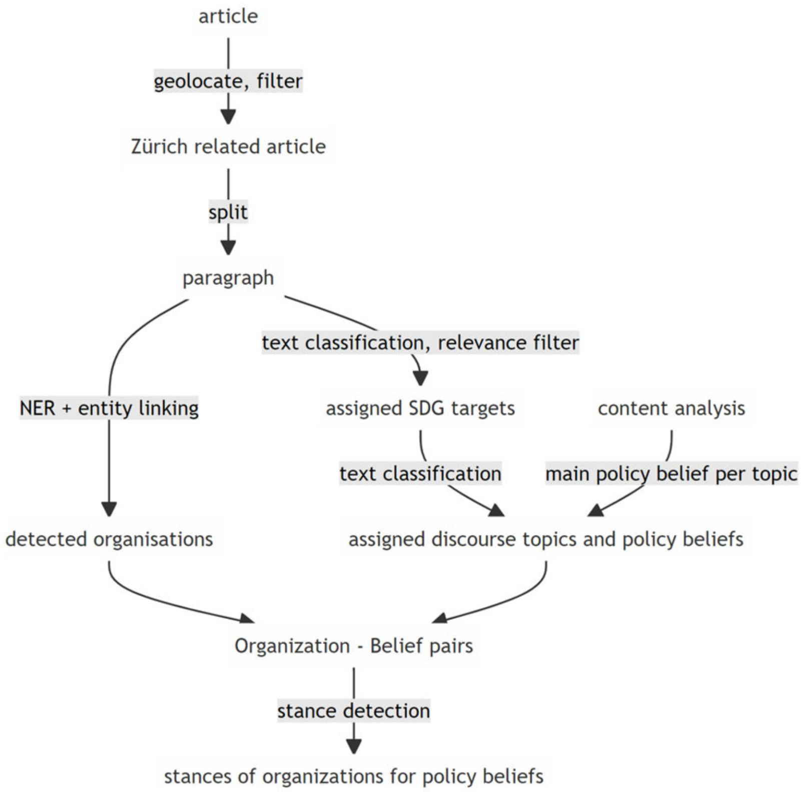

article
geolocate, filter
Zürich related article
split
paragraph
text classification, relevance filter
NER + entity linking
assigned SDG targets
content analysis
text classification
main policy belief per topic
detected organisations
assigned discourse topics and policy beliefs
Organization - Belief pairs
stance detection
stances of organizations for policy beliefs

Figure 5. Components of the processing pipeline for automated extraction of discourse networks.

To test if our automated procedure can recover our proposed discourse shaper, topic opinion leader, and ideological alignment metrics, using windowed aggregation, we calculate these metrics over quarterly computed 48-month aggregation windows as described earlier for both networks created from manual annotation and automated procedures.

As proposed earlier, to identify discourse shapers, we sum normalized z-scores (standardized counts of opinionated edges per policy belief) and normalized c-scores (standardized counts of opinionated edges across all beliefs) per actor, per aggregation window, for an actor's most prevalent stance classifier.

To identify topic opinion leaders, we calculate c-scores per topic for every stance classifier. This, for example, finds the most active actors supporting more cycling infrastructure or the actors indicating most opposition to speed reduction per time window.

To identify ideological alignment, we count the number of closed, balanced four cycles in the network for a time window. Every such four-cycle contains two actors and two beliefs. We consider a cycle balanced if the edge sets per actor match in qualifiers. For example, this could mean that if actor A supports speed reduction and opposes more cycling infrastructure, actor B also does so, leading to four cycles with two support edges and two opposition edges per actor-belief pair.

3.5 Reproducibility

All data and code to enable replication of computations presented here based on the discourse network graph originating from our processing pipeline are available in an open online repository

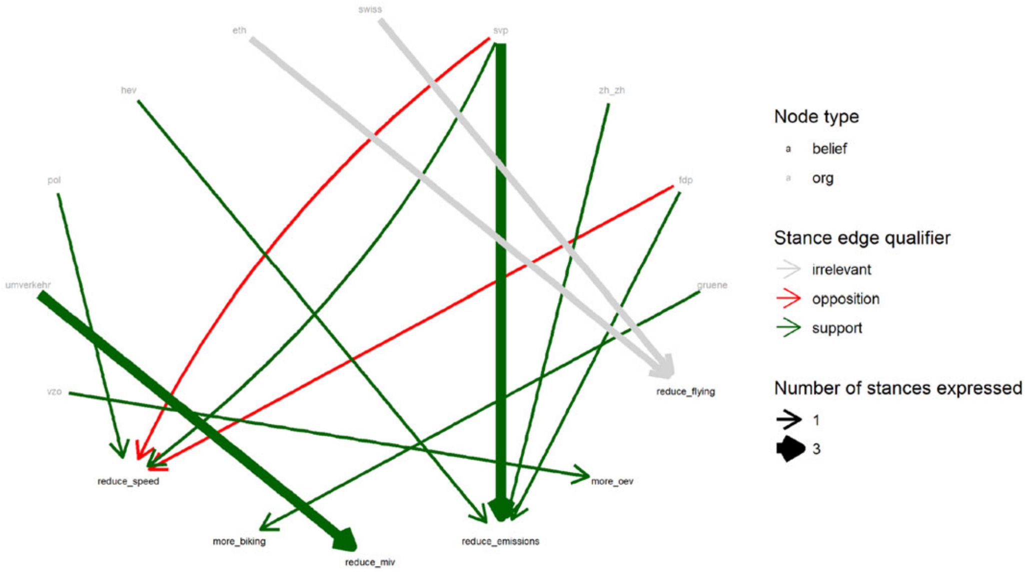


Figure 6. Example of extracted discourse network for March 2023. Stances edges are aggregated over the time frame, and the thickness of edges shows the number of stances expressed during the time frame. Only the most prevalent category/categories after aggregation is/are shown.

at https://doi.org/10.5281/zenodo.14713615. For further computations using a discourse network graph following our conceptualization, we also provide a small R (R Core Team, 2023) package `diskurs`.[6]

For the processing pipeline creating the discourse network graph, given agreements with the publishing companies making the raw media data (thus the roughly 1 million newspaper articles processed) available to us, we cannot provide access to the raw data and thus also no full reproducibility. This also applies to annotated training and testing data for classifiers, which consists of paragraphs extracted from copyrighted media data.

To facilitate investigation of the pipeline and put it under scrutiny, we have made the classification API, which includes all non-LLM-based pipeline components, available at https://zenodo.org/records/14702518. The LLM-based stance classification classifier is made available as a Python package (`stance-llm`[7]), allowing users to supply their own LLM backend model.

4. Results and discussion

4.1 Discourse network

Our processing pipeline results in an eventual set of 7314 stance edges (support: 3968, opposition: 1129, irrelevant: 2217) of 169 unique organizational actors regarding 7 policy beliefs. Stances classified within the irrelevant class (the second-largest category) are also included as stance edges in the graph. On the one hand, we retained stances classified as irrelevant to use in these predictions for tests against annotations in the window validity hypothesis tests and on the other hand to capture information on actors appearing tangentially in media around a discourse but not taking opinionated stances. Figure 6 provides an illustration of a typical network snapshot over a sample (one-month) time window.

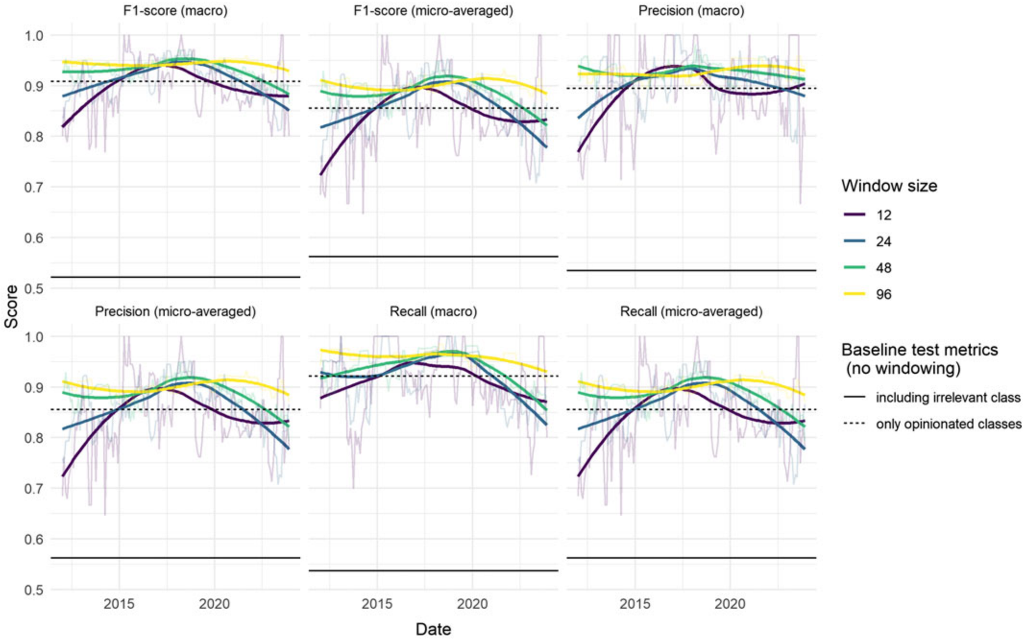


Figure 7. Results of window validity hypothesis test procedure for individual stance edges. The plot panels show scores for macro and micro-averaged precision, recall, and f1-score metrics. Macro-averaged scores average across all classes computed separately, while micro-averaged scores are computed once for the entire dataset. Tests are based on comparing windowed aggregations of edges in the predicted discourse network graph against windowed aggregations of edges based on manual annotations, varying four different window sizes. Windowed aggregations and test statistics are computed for every month within the time frame. Straight horizontal lines show a reference to non-windowed scores.

4.2 Window validity hypothesis: test of edge recovery

Intersecting the discourse network gathered through automated extraction with a network constructed from annotated stances resulted in a reference network for window validity hypothesis testing with 1264 stance edges and 118 organizations. Thus, our randomly sampled test set of stances covers 70% of the 169 actors detected. Figure 7 shows the results of the central window validity hypothesis test for network recovery in terms of individual edges. Windowed aggregations of edges in the predicted discourse network graph are compared to windowed aggregations of edges based on manual annotations. Windowed aggregations and test statistics are computed for every month in the time period covered by the data, with the month splitting the window.

Windowed aggregation of stances greatly improves on the main non-windowed baseline scores (shown as a solid line in 7, with scores varying between 0.5 and 0.6), which show test metrics calculated comparing predicted versus annotated edges without any aggregation. Most importantly, test metrics scores for windowed aggregations are very high, around and above 0.9 for both macro and micro-averaged f1, precision, and recall scores.

If the non-windowed baseline score is calculated only for opinionated stances (dashed line in Figure 7), thus comparing manual annotations only against automated predictions that were not in the irrelevant class (the largest), especially the longer 48-month and 96-month windows still consistently outperform the baseline scores. However, evaluating only predictions excluding irrelevant classes, while interesting, is not a fair test for the window validity hypothesis, as an automated stance classifier is in most practical applications likely to need to include an irrelevance class (a way to exclude the very common class of non-expressions of a stance by an entity in the underlying data).

https://doi.org/10.1017/nws.2025.4 Published online by Cambridge University Press

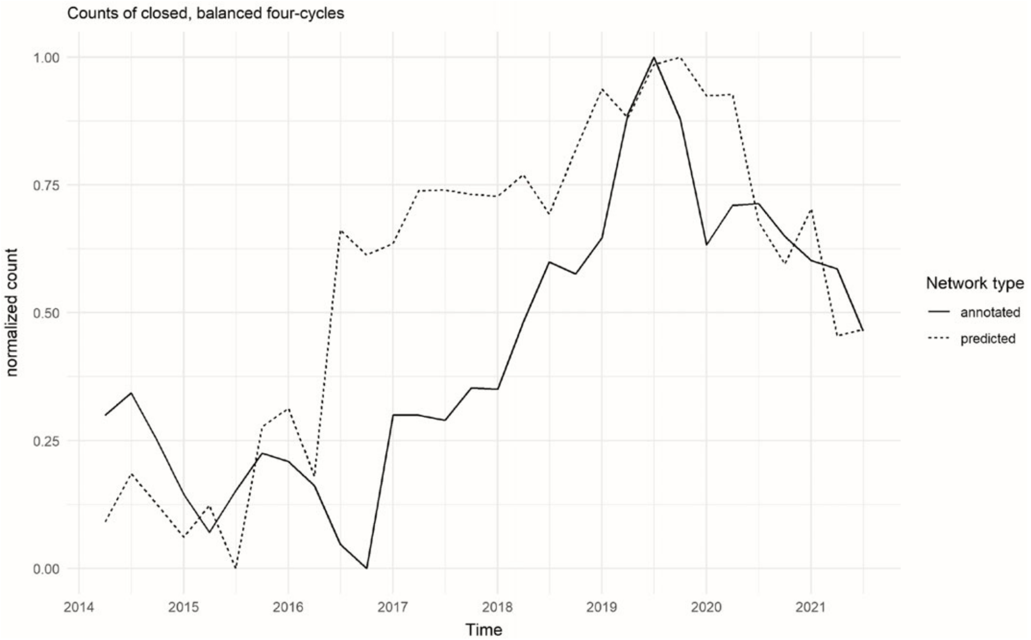


Figure 8. Results of window validity hypothesis test procedure for ideological closure based on normalized counts of closed, balanced four cycles. Comparison of annotated versus predicted networks. 48-month rolling average, computed at quarterly intervals, normalized over time range.

In conclusion, we find strong support for the window validity hypothesis given our results for recovery of individual stance edges. Combining predictions from automated discourse network data gathering over a time period results in almost equivalent results as if the same procedure is carried out based on gold-standard manual annotations. In comparison, individual predictions for edges are much less internally valid.

4.3 Window validity hypothesis: tests of discourse property metrics recovery

Figures 8, 9, and 10 show the results of a test for the window validity hypothesis based on the recovery of our proposed network-level metric for ideological closure over time, as well as node-level metrics for opinion leadership and discourse shaping over time. The tests are conducted starting from the same procedure of intersecting the discourse network gathered through automated extraction with a network constructed from annotated stances used to test network recovery in terms of individual edges. Then, metrics computed on windowed aggregations of the predicted networks from automated extraction are evaluated against metrics computed in the same way from manual annotation.

Figure 8 shows that in terms of normalized counts of closed, balanced four cycles over time, the automated procedure manages to reproduce the time trends apparent in the test set. The general trend of ideological closure among discourse participants being low (relatively speaking) in the first half of the 2010s and then starting to increase, reaching its peak in 2019, is apparent in both annotated and predicted networks.

As introduced earlier, opinion leaders are actors who are especially active in either opposing or supporting a main policy belief shaping a topic. Figure 9 shows that our automated network extraction procedure manages to generally reproduce the identification of opinion leaders fairly well. This is indicated by how close the slope of linear regressions of predicted on annotated scores

https://doi.org/10.1017/nws.2025.4 Published online by Cambridge University Press

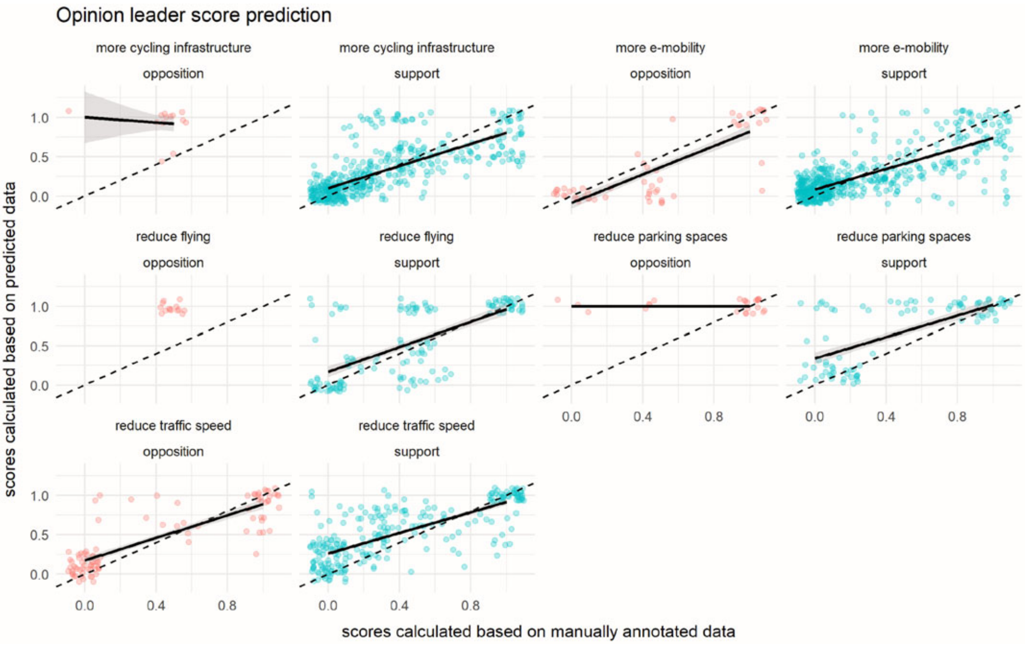


Figure 9. Results of window validity hypothesis test procedure for opinion leader identification per discourse topic based on z-scores of actors including only most prevalent stance classifier per aggregation window. Comparison of individual actor scores for opinion leadership per topic in annotated versus predicted networks. Scores are computed at quarterly intervals with 48-month window aggregations. One point shows scores for one actor within one time window. The dashed reference line shows a perfect linear relationship. Solid lines show linear regression of y-axis scores on x-axis scores with 88% confidence interval.

are to a perfect linear relationship in most categories with larger amounts of data available for testing. For the calculations of opposition opinion leadership regarding more cycling infrastructure, reduction of flying, and reduction of parking spaces, regression lines should probably not be overinterpreted, as the amount of data for evaluation in our test set was quite low. Another indication that the predicted network manages to recover the metric well is that only in a few categories are actors present in quadrants in the off-diagonal (thus left top and right bottom corners), which would indicate the scores being completely off mark.

Figure 10 shows that the automated network extraction procedure further manages to generally reproduce the identification of discourse shapers, our second node-level discourse network metric. We conceptualized discourse shapers as actors who are at the same time active across multiple urban transport topics but also very relevant within specific topics. Our automated procedure does fairly well in recovering these actors, indicated by how close the slope of the linear regression of predicted on annotated scores is to a perfect linear relationship. Further, almost all comparisons are placed along the main diagonal, indicating little to no complete miss-classifications.

In conclusion, we find strong support for the window validity hypothesis given our results for recovery of structurally descriptive discourse network metrics. Computing both node-level and a network-level metric-based predictions from automated discourse network data gathering through windowed aggregations over a time period reproduces the trends and patterns that result if the same procedure is carried out based on gold-standard manual annotations.

We evaluated a set of relatively long aggregation windows to test the window validity hypothesis (ranging from 12 to 48 months). Given the stance inertia present in our data, these represent reasonable time frames in our opinion. However, in other settings of more dynamic or nascent or emerging policy fields (Ingold, et al., 2017), or when investigating stances toward concepts that

https://doi.org/10.1017/nws.2025.4 Published online by Cambridge University Press

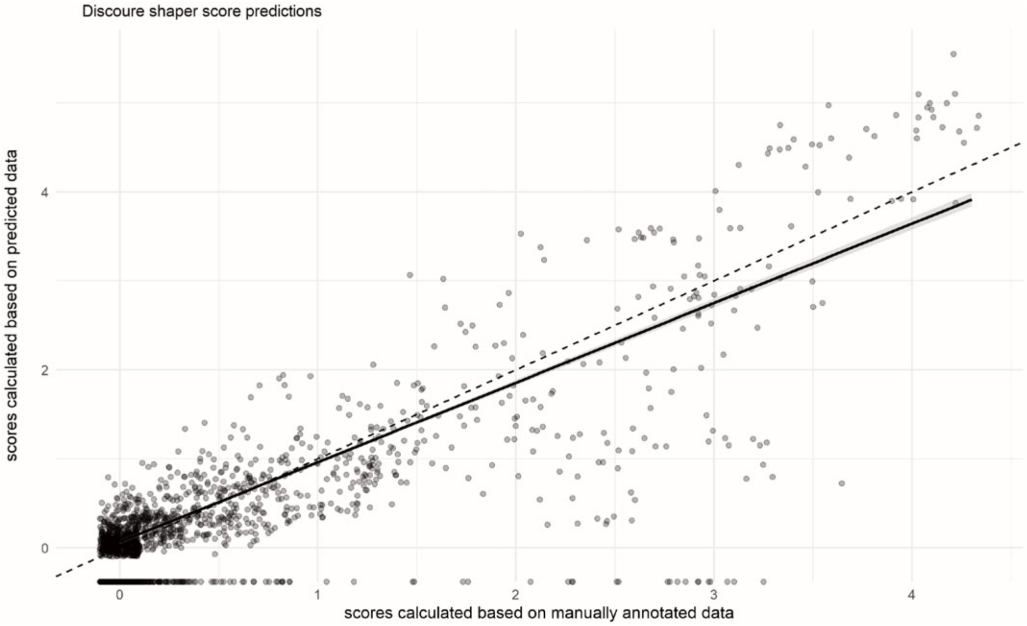


Figure 10. Results of window validity hypothesis test procedure for discourse shaper identification based on summing z-scores and c-scores of actors per aggregation window. Comparison of individual actor scores in annotated versus predicted networks. Scores are computed at quarterly intervals with 48-month window aggregations. One point shows scores for one actor within one time window. The dashed reference line shows a perfect linear relationship. Solid lines show linear regression of y-axis scores on x-axis scores with 88% confidence interval.

are less stable than policy beliefs, such as policy instrument preferences of actors (Weible, et al., 2009), shorter windows for aggregation may be necessary. In turn, this would however require more high-intensity data in order for there to be enough data redundancy for the window validity hypothesis to work out. We see great potential for future research to explore the limits of automated approaches in such settings.

5. Conclusion

Our analysis suggests that automated discourse network data gathering can result in internally valid representations of broad trends in discourse networks over time, even though classification tasks are challenging. Where to go from here?

We see our results as indicating great potential for studies on discourse networks to leverage automated data gathering approaches to contribute insights into societal discourse around central governance challenges of our time, at temporal, spatial, and topical scales previously almost impossible. We hope our conceptualization and tools provide a blueprint to do so, especially if approaches can make use of data redundancy and stance inertia in data.

Still, our results also suggest that there is a limit to insights gained from automated network data gathering in comparison to qualitative, close reading approaches, and the latter are needed more than ever if the goal is actual understanding of discourse.

First, the internal validity of our automated procedure, which we demonstrate in this article, does not imply the external validity of our suggested metrics. To evaluate the external validity of results from an automated procedure such as ours, mixed method approaches, integrating qualitative validation techniques, for example, through semi-structured interviews, where actors in discourse are confronted with predictions about their position in a discourse, are likely needed.

Second, the upfront cost in terms of effort to establish an internally valid automated processing pipeline, including adequate testing, is likely inefficient for many research applications where measurements of discourse are not repeated over a period of time. The efforts of the research community around automated approaches should thus be focused on reducing these upfront costs by sharing test sets, tooling, and models, while recognizing that the temporal and material conditions of discourse set a natural limit to re-use as well. Further, there is likely a clear delta in terms of choosing or combining more manual or automated approaches. Most one-off discourse network analyses at a given point in time are likely better served by robust manual annotation. Automated components seem only worth introducing if re-usable components can be generated and utilized or if the spatial or temporal scope of the research project is such that manual annotation cannot generate valid inferences.

Third, general-purpose, zero-shot models such as the LLMs we employed in our analysis pipeline can be helpful in overcoming some problems in automation, such as the stance detection component of our pipeline, but they are no panacea and may offer only limited internal validity. More than ever, the deployment of such models needs evaluations against high-quality, situated test sets, as general-purpose models meet highly specific situations. This is where close reading and careful annotation are still needed.

Acknowledgments. We want to thank Myriam Pham-Truffert for her help in annotating and reviewing test and development data for methods and tests presented in this article.

Competing interests. Mario Angst is president of the board of a local bike messenger cooperative in Zürich committed to offering sustainable transport services. Neitah Müller is a member of and was a candidate for parliament for the political party "Grünliberale." Both activities are likely to influence their subjective perception of local urban sustainable transport discourse, but they have tried to mitigate this as much as possible during work on this project.

Funding statement. This research was supported by the DIZH, Digitalisierungsinitiative der Zürcher Hochschulen.

Notes

[1] Based on https://huggingface.co/google-bert/bert-base-german-cased
[2] In version 3.7.0, see https://github.com/explosion/spacy-models/releases/tag/de_core_news_lg-3.7.0 and https://spacy.io/models/de#de_core_news_lg
[3] See https://pypi.org/project/stance-llm/
[4] Specifically the variant published to the model hosting platform Hugging Face at https://huggingface.co/seedboxai/KafkaLM-7B-DARE_TIES-LaserRMT-QLoRA-DPO-v0.5
[5] https://laion.ai/blog/leo-lm/, accessed 22.4.2024
[6] Available at https://urban-sustainability-lab-zurich.r-universe.dev/diskurs
[7] See https://pypi.org/project/stance-llm/

References

Angst, M., & Brandenberger, L. (2022). Information exchange in governance networks–who brokers across political divides? [In en]. *Governance*, *35*(2), 585–608. doi: 10.1111/gove.12601.

Angst, M., Widmer, A., Fischer, M., & Ingold, K. (2018). Connectors and coordinators in natural resource governance: insights from Swiss water supply. *Ecology and Society*, *23*(2), 1.

Bodin, Ö., & Tengö, M. (2012). Disentangling intangible social–ecological systems. *Global environmental change: human and policy dimensions*, *22*(2), 430–439. doi: 10.1016/j.gloenvcha.2012.01.005.

Bossner, F., & Nagel, M. (2020). Discourse networks and dual screening: analyzing roles, content and motivations in political twitter conversations. *Politics and governance*, *8*(2), 311–325. doi: 10.17645/pag.v8i2.2573.

Carlsen, H. B., & Ralund, S. (2022). Computational grounded theory revisited: from computer-led to computer-assisted text analysis. *Big Data & Society*, *9*(1), doi: 10.1177/20539517221080146.

Ceron, T., Baric, A., Blessing, A.É, Haunss, S., Kuhn, J., Lapesa, G. . . . Zauchner, P. F. (2024). Automatic analysis of political debates and manifestos: successes and challenges [in en]. In *Robust argumentation machines, 71-88. Lecture* *notes in computer science*. Cham: Springer Nature Switzerland. isbn: 9783031635359,9783031635366. doi: 10.1007/978-3-031-63536-6_5.

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2021). Machine learning for social science: an agnostic approach. *Annual Review of Political Science*, *24*(1), 395–419. doi: 10.1146/annurev-polisci-053119-015921.

Gronow, A., Wagner, P., & Ylä-Anttila, T. (2020). Explaining collaboration in consensual and conflictual governance networks [in en]. *Public administration*, *98*(3), 730–745. doi: 10.1111/padm.12641.

Haunss, S., Kuhn, J., Padó, S., Blessing, A., Blokker, N., Dayanik, E., & Lapesa, G. (2020). Integrating manual and automatic annotation for the creation of discourse network data sets. *Politics and Governance*, *8*(2), 326–339. doi: 10.17645/pag.v8i2.2591.

Hull, A. (2008). Policy integration: what will it take to achieve more sustainable transport solutions in cities? *Transport Policy*, *15*(2), 94–103. doi: 10.1016/j.tranpol.2007.10.004.

Ingold, K., Fischer, M., & Cairney, P. (2017). Drivers for policy agreement in nascent subsystems: an application of the advocacy coalition framework to fracking policy in Switzerland and the UK. *Policy Studies Journal*, *45*(3), 442–463. doi: 10.1111/psj.12173.

Kammerer, M., & Ingold, K. (2023). Actors and issues in climate change policy: the maturation of a policy discourse in the national and international context. *Social Networks*, *75*, 65–77. doi: 10.1016/j.socnet.2021.08.005.

Kuhn, M., Vaughan, D., & Hvitfeldt, E. (2024). Yardstick: tidy characterizations of model performance. R package version 1.3.1. https://yardstick.tidymodels.org, https://github.com/tidymodels/yardstick.

Lan, X., Gao, C., Jin, D., & Li, Y. (2023). Stance detection with collaborative role-infused llm-based agents. https://arxiv.org/abs/2310.10467.

Laumann, E. O., & Knoke, D. (1987). *The organizational state: social choice in national policy domains*. Madison: University of Wisconsin Press.

Leifeld, P. (2009). Die untersuchung von diskursnetzwerken mit dem discourse network analyzer (DNA). In: Schneider, V., Janning, F., & Leifeld, P., ed. *Politiknetzwerke. modelle, anwendungen und visualisierungen*. Thomas Malang. Wiesbaden: Springer VS.

Leifeld, P. (2013). Reconceptualizing major policy change in the advocacy coalition framework: a discourse network analysis of German pension politics. *Policy studies journal: the journal of the Policy Studies Organization*, *41*(1), 169–198. doi: 10.1111/psj.12007.

Leifeld, P. (2020). Policy debates and discourse network analysis: a research agenda. *Politics and Governance*, *8*(2), 180. doi: 10.17645/pag.v8i2.3249.

Levin, S. A., Milner, H. V., & Perrings, C. (2021). The dynamics of political polarization [in en]. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(50), e2116950118. doi: 10.1073/pnas.2116950118.

Lewis, S. L., & Maslin, M. A. (2015). Defining the anthropocene [in en]. *Nature*, *519*(7542), 171–180. doi: 10.1038/nature14258.

Lubell, M., & Morrison, T. H. (2021). Institutional navigation for polycentric sustainability governance. *Nature Sustainability*, *4*(8), 664–671. doi: 10.1038/s41893-021-00707-5.

Marquet, O., Mojica, L., Fernández-Núñez, M.-B., & Maciejewska, M. (2024). Pathways to 15-minute city adoption: can our understanding of climate policies' acceptability explain the backlash towards x-minute city programs? *Cities*, *148*, 0264–2751. doi: 10.1016/j.cities.2024.104878.

McElreath, R. (2016). *Statistical rethinking: a Bayesian course with examples in R and Stan*, 1st edition. Chapman and Hall/CRC. doi: 10.1201/9781315372495.

Montani, I., Honnibal, M., Honnibal, M., Van Landeghem, S., Boyd, A., Peters, H., McCann, P. O. et al. (2022). explosion/spaCy: v3.2.2: Improved NER and parser speeds, bug fixes and more, February. doi: 10.5281/zenodo.6045547.

Olesen, J. M., Bascompte, J., Dupont, Y. L., & Jordano, P. (2007). The modularity of pollination networks. *Proceedings of the National Academy of Sciences*, *104*(50), 19891–19896. doi: 10.1073/pnas.0706375104.

Patel, Z., Greyling, S., Simon, D., Arfvidsson, H., Moodley, N., Primo, N. & Wright, C. (2017). Local responses to global sustainability agendas: learning from experimenting with the urban sustainable development goal in cape town. *Sustainability Science*, *12*(5), 785–797. doi: 10.1007/s11625-017-0500-y.

R Core Team (2023). *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rhodes, R. A. W. (1996). The new governance: governing without government. *Political studies*, *44*(4), 652–667. doi: 10.1111/j.1467-9248.1996.tb01747.x.

Roy, A., Fafalios, P., Ekbal, A., Zhu, X., & Dietze, S. (2022). Exploiting stance hierarchies for cost-sensitive stance detection of web documents. *Journal of Intelligent Information Systems*, *58*(1), 1–19. doi: 10.1007/s10844-021-00642-z.

Sabatier, P. A. (1988). An advocacy coalition framework of policy change and the role of policy-oriented learning therein. *Policy sciences*, *21*(2), 129–168. doi: 10.1007/BF00136406.

Scott, T. A., & Ulibarri, N. (2019). Taking network analysis seriously: methodological improvements for governance network scholarship. *Perspectives on Public Management and Governance*, *2*(2), 89–101. doi: 10.1093/ppmgov/gvy011.

Shrestha, S., Haarstad, Håvard, & Rosales, R. (2024). Power in urban logistics: a comparative analysis of networks and policymaking in logistics sustainability governance. *Environmental Innovation and Societal Transitions*, *51*, 100845. doi: 10.1016/j.eist.2024.100845.

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data [in en]. *Nature*, *631*(8022), 755–759. doi: 10.1038/s41586-024-07566-y.

Steffen, W., Richardson, K., Rockström, J., Cornell, S. E., Fetzer, I., Bennett, E. M., Biggs, R. et al. (2015). Planetary boundaries: guiding human development on a changing planet. *Science*, *347*(6223). doi: 10.1126/science.1259855. eprint: https://science.sciencemag.org/content/347/6223/1259855.full.pdf

Udris, L. (2023). Unabhängigkeit und politische positionierung der medien bei volksabstimmungen: jahrbuch qualität der medien studie 3 / 2023 [in de]. Jahrbuch Qualität der Medien (Zürich), Jahrbuch Qualität der Medien, 2023/3 (September): 17. https://doi.org/10.5167/uzh-236162

United Nations General Assembly. (2023, November 4). Transforming our world: the 2030 Agenda for Sustainable Development. Refworld; Refworld - UNHCR's Global Law and Policy Database. https://www.refworld.org/docid/57b6e3e44.html (accessed on 20 January 2021).

Weible, C. M., & Sabatier, P. A. (2009). Coalitions, science, and belief change: comparing adversarial and collaborative policy subsystems. *Policy studies journal: the journal of the Policy Studies Organization*, *37*(2), 195–212. doi: 10.1111/j.1541-0072.2009.00310.x.

Weible, C. M., Sabatier, P. A., & McQueen, K. (2009). Themes and variations: taking stock of the advocacy coalition framework. *Policy studies journal: the journal of the Policy Studies Organization*, *37*(1), 121–140. doi: 10.1111/j.1541-0072.2008.00299.x.

Zuiderveen Borgesius, F. J., Trilling, D., Müller, J., Bodó, B., de Vreese, C.H., & Helberger, N. (2016). Should we worry about filter bubbles? *Internet Policy Review*, *5*(1). doi: 10.14763/2016.1.401.

Appendix 1. Data documentation

Appendix 1.1 List of media sources

Media names: Tagesanzeiger (TA), nzz.ch (NZZO), Neue Zürcher Zeitung Folio (NZZF), NZZ Campus (CAMP), Neue Zürcher Zeitung am Sonntag (NZZS), Neue Zürcher Zeitung (NZZ), Neue Zürcher Zeitung am Sonntag Magazin (NZZM), bellevue.nzz.ch (NZZB), 20 Minutes (ZWAS), 20 Minuten (ZWA), 20 Minuten Friday (ZWAF), 20 Minuten Online (ZWAO), züritipp (Tages-Anzeiger) (TAZT), Zürich Express (ZUE)

Appendix 2. Codebook for relevance classifier annotation

Overall goal of annotation.

Is paragraph related to sustainable transport?

Flag accept (label: 11_2_TRANSPORT) marks paragraphs that are related to SDG target 11.2, sustainable transport systems for all.

From the UN Agenda 2030 definition of SDG 11.2:

By 2030, provide access to safe, affordable, accessible, and sustainable transport systems for all, improving road safety, notably by expanding public transport, with special attention to the needs of those in vulnerable situations, women, children, persons with disabilities, and older persons.

Four key criteria. Text paragraph contains a direct mention of:

- Safety of transport
- Environmental impact of transport
- Accessibility of transport
- Affordability of transport

Location criteria. While we annotate data mostly within urban contexts, we do **not** consider the location of a paragraph in the decision of whether or not to accept or reject it.

Direct versus indirect relation to criteria. In many cases, paragraphs do not contain a direct mention of the four key criteria but refer to a concept that is indirectly linked to these. **Paragraphs are to be accepted if the link is proximate to a transport-related topic**. See examples below.

Accept choices and edge cases log.

- Tempo 30 and other Geschwindigkeitsreduktion related topics are generally accepted and considered to be related to at least impact in terms of noise
- Autofrei is generally accepted and considered related to impact in terms of modal split
- Modal split accepts discussions of shifts in modes of transport
- Rolltreppe, if considered in the context of accessibility for pedestrians
- Infrastructure expansion (such as new rail lines or roads) is accepted only if a sustainable transport context can be assumed. For public transport and active transport infrastructure, this is assumed a priori
- Parking (e.g., Parkplätze/ParkhausBlaue Zonen), if linked to broader discussions of car traffic impact or accessibility
- The mention of an organization or other entity clearly connected to transport (e.g., Verkehrs-Club der Schweiz (VCS)) if a reference to, for example, an potential impact of transport (e.g., Lärm), is made, even if the paragraph itself would otherwise not refer to transport
- Car bypasses (Umfahrungen) are accepted if a reference to the four criteria is made
- Drones are only accepted if discussed for transport purposes and a criteria (e.g., noise) is mentioned
- Safety of transport includes measures for protection against disease. As such measures for reducing the risk of transmission of the coronavirus are accepted
- Affordability of transportation covers both costs for individuals and society
- Discussions of flying as a form of transport: accepted if environmental impacts and noise emissions are mentioned
- Reporting on outcomes of popular initiatives related to sustainable transport
- incentives to use public transport
- Fahrplanwechsel is only accepted if a connection to one of the four criteria is made
- Mention of electric vehicles—only accept if some connection to the four criteria

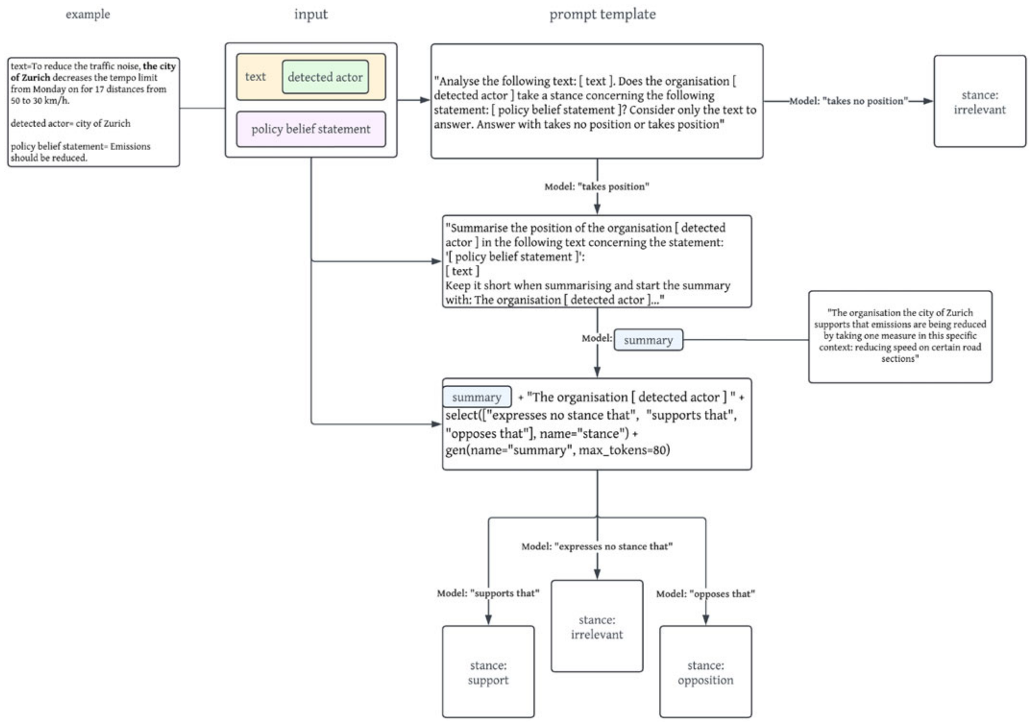
Reject choice log.

- Sports events (e.g., bike races, car races) *except* if they are put into context of the four criteria, for example, if done to promote a sustainable transport form
- Accident reports in general *except* if put into a broader context of, for example, safety of transport beyond the specific case
- Autofrei if solely byproduct of, for example, construction/*clearly* not independent
- Autoraser (excessive speeding) topics
- Descriptive statistics/traffic statistics *except* if they are connected to a broader context of sustainable transport
- Historical transport events that are not recent. Recent is generally 90s+
- Pure mentions of traffic jams *except* if put into broader context of a sustainable transport topic

https://doi.org/10.1017/nws.2025.4 Published online by Cambridge University Press

- Infrastructure expansion of transport modes without mention of reason because not every expansion is for safety/impact/affordability/accessibility
- Gig economy (e.g., Uber) working conditions
- short term construction emissions (e.g., road upkeep/building)
- Choice of tram model—tend to reject (because not equivalent to increase public transport)
- reject pure mentions of *Parkplatzkompromiss* except if it is connected to a broader context of sustainable transport

Appendix 3. Prompt Illustration—English version



text=To reduce the traffic noise, **the city of Zurich** decreases the tempo limit from Monday on for 17 distances from 50 to 30 km/h.

detected actor= city of Zurich

policy belief statement= Emissions should be reduced.

"Analyse the following text: [text]. Does the organisation [detected actor] take a stance concerning the following statement: [policy belief statement]? Consider only the text to answer. Answer with takes no position or takes position"

"Summarise the position of the organisation [detected actor] in the following text concerning the statement: '[policy belief statement]':
[text]
Keep it short when summarising and start the summary with: The organisation [detected actor]..."

"The organisation the city of Zurich supports that emissions are being reduced by taking one measure in this specific context: reducing speed on certain road sections"

summary + "The organisation [detected actor] " + select(["expresses no stance that", "supports that", "opposes that"], name="stance") + gen(name="summary", max_tokens=80)

Figure 11. Example of the structure of the prompt chain ”is2,” translated to English.

Cite this article: Angst M., Müller N. N. and Walker V. (2025). Automated extraction of discourse networks from large volumes of media data. *Network Science* **13**, 1–23. https://doi.org/10.1017/nws.2025.4