Do Arguments Migrate? Using NLP for Understanding Academia

Sassan Gholiagha, Jürgen Neyer, Mitja Sienknecht

European New School of Digital Studies
European University Viadrina
gholiagha@europa-uni.de
European New School of Digital Studies
European University Viadrina
neyer@europa-uni.de
European New School of Digital Studies
European University Viadrina
sienknecht@europa-uni.de

Abstract

A crucial question for academia is the relevance of arguments for scientific progress. Are participants in academic debates open to the arguments and insights of other authors even if they are embedded in competing research paradigms? Or is discursive openness limited to intra-paradigmatic debates? What are the conditions under which arguments are migrating inside and across paradigms? The project uses ML and NLP for analysing a large-n data set that combines thousands research articles in International Relations. It sets up the largest text corpus available for international relations and trains an algorithm to recognize and qualify arguments according to their theoretical origin, their supporting evidence and their argumentative structure.

Keywords: NLP, ML, International Relations, Arguing, Academia, Politics

Conference Paper for the 10th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, April 21-23, 2023, Poznań, Poland

1. Do Arguments Matter?[1]

What is the importance of the structure and quality of a scientific argument for its reception? What kind of arguments have a significant probability to change another's opinion, and to what extent can a systematic connection between reception intensity and specific features of scientific arguments be empirically proven? Is there a meaningful nexus between scientific and political communication, and if so, in what direction and under what argumentation-structural conditions do arguments migrate between discursive arenas?

2. Theory

Scientific discourse assumes that argumentative quality matters. Arguments are assumed to be assessed according to the merits of their scientific quality. Relevant standards include different features depending on scientific theoretical provenance. Positivist epistemologies emphasize the empirical verifiability of claims and the repeatability of lines of evidence (King, Keohane and Verba 1994). Constructivist epistemologies reject this claim and instead emphasize the subjectivity of observation and thus the impossibility of objective testing of claims about social facts (Berger and Luckmann 1966; Kratochwil and Ruggie 1986). Therefore, alternative standards of science are emphasized, such as the detailed and plausible reconstruction of meaning with the aim of making them comprehensible and thus understandable (see Jackson 2011 for an overview of different scientific logics for IR). Regardless of the respective scientific theoretical orientation, theoretical reflections are in both cases endowed with additional plausibility when they are supported by empirical evidence. Both perspectives also share the idea that empirical data only become relevant through their explicit integration into a theoretical context. They furthermore both assume that theoretical perspectives gain traction to the degree that they are explained through an explicit exposition of their premises. The idea that quality matters for arguments to be considered seriously

[1] This paper presents the research design and first findings from a four-year research project "Sozialwissenschaftliches KI-Labor für Forschendes Lernen (SKILL)" funded by the German Ministry for Education and Research, the Brandenburg Ministry for Science and Culture and the Thuringian Ministry for Science, Research and Art. It is chaired by Bernd Fröhlich, Katrin Girgensohn, Jürgen Neyer and Benno Stein.

also applies to scientific policy advice. When scientists offer advice to policy-makers, they usually assume that their arguments will be taken the more into account that they comply with scientific standards.

The assumption of a high relevance of argumentation-specific features for their reception by other scientists as well as by policy makers is, however, not undisputed. Receptions within the scientific community are not only influenced by the quality of the arguments presented, but also by their integration into established research networks (Risse, Wemheuer-Vogelaar and Havemann 2020) and sometimes even citation cartels (Teodorescu and Andrei 2013). Intellectually challenging positions that deviate from the majority opinion are easily ignored if they are not backed by particularly strong arguments and evidence, while complying with lower standards is often good enough for arguments that replicate the mainstream. Thomas Kuhn has prominently pointed out that research programs have their own internal logic, selectively receiving content based on whether it fits into dominant paradigms (Kuhn 1962). Arguments would be easily ignored, despite high formal quality, if they ignored dominant understandings of problems and solution strategies (paradigms) and followed unorthodox trajectories. For policy advice, the assumption applies analogously that scientifically sound arguments are only received by policymakers if they can be reconciled with prevailing political calculations, i.e., are politically opportune (Böcher 2022). Luhmann's thesis of different societal functional systems, each with its own language codes and rationality criteria (Luhmann 1984), also suggests that the idea of a search for truth that integrates functional systems and is based on argumentation is at least optimistic: In science, knowledge is generated within the framework of disciplinary concepts and prevailing epistemological interests. It often sits squarely with the logic of politics in which solutions must be negotiated and compromises will often be based on different values and interests. Science also involves a continuous critique and problematization of findings and thus inevitably rejects any conclusive certainty. This irrevocable uncertainty in science is, in turn, difficult to reconcile with the expectation that policymakers are able to make effective decisions that inspire consent and confidence (cf. Böcher 2022).

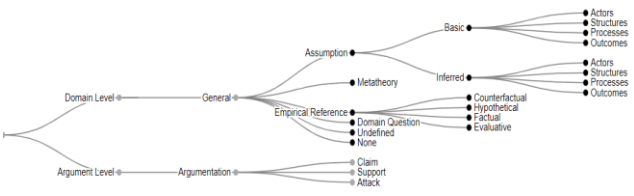
The tension between the thesis of an argumentation-based dynamic of scientific discourse on the one hand and the indications of non-scientific factors influencing the reception of arguments on the other hand gives rise to two interrelated questions: What is the significance of the quality of a scientific argument for its reception and the change of another's opinion, and to what extent can a systematic connection between reception intensity and specific quality features of scientific arguments be empirically proven? Is there a connection between the two spaces of scientific and political communication, and if so, in what direction and under what argumentation-structural conditions do arguments migrate between discursive arenas?

3. Epistemology

The SKILL project addresses these questions by analysing a large-n corpus of academic articles and political speeches in the United Nations General Assembly. It develops an algorithm that can recognize and compare patterns of argumentation structures in both text corpora. The project follows an abductive approach, which is based on a combination of machine learning (ML), artificial intelligence (AI), and natural language processing (NLP). It allows the combination of quantitative and qualitative methods and thus a "methodological twin-move of making big data thick and thick data big." (Adler-Nissen et al 2021: 1, emphasis in original). Abductive approaches to pattern recognition have been quite unusual for the social sciences. They have only recently started to gain some attention in the context of large data sets, and only slowly been taken into account by the social sciences. This immigration into a theory-driven discipline was triggered by the realization that computer-based methods can unveil social patterns which have since long been reflected upon but hardly ever been described empirically. The successes of research driven by big data have underlined that individual decisions often reflect broader social patterns rather that individual reflection (Nassehi 2019, Meyer-Schönberger/ Cukier 2013). Social action is not only shaped by digitalization, but seems to be highly digitally structured itself and shaped by patterns of rule-compliant action. Pattern recognition procedures thus apply a methodology which is very much in line with the an important logic of social action. The seemingly naive question of "what is?", which has often been rejected as unscientific up to now, moves to the center in a recognition-oriented approach. Not the testing of hypotheses or the search for merely subjective meaning inherent in understanding-oriented approaches, but the identification, representation and analysis of regular social phenomena becomes the goal of the research process.

4. Data

4.1. Argumentation model



We use a model of argumentation which uses NLP methodology for allowing an algorithm to identify and classify arguments. The methodology holds that text can be made machine-readable by annotating individual sentences, i.e. using a set of clearly defined categories for attaching meaning to text. The methodology is composed of three main elements:

1. It starts from the assumption that the meaning of sentences can be assessed individually and that they can be understood without reference to the broader context in which they are embedded. Texts are thus decomposed into a set of sentences that are each annotated irrespective of their relationship with provisional or trailing sentences. The decomposition of texts is not unconditional, however. Provisional or trailing sentences are used as an additional resource for annotation if they provide important information without which sentences cannot be properly understood. The process of decomposing texts into sentences is also contextualized by adding relationships between sentences. Sentences which refer to each other and which provide an explicit argumentative context are annotated as hanging together. For example, if sentence 1 contains a claim and sentence 2 lists the supporting evidence, then both sentences are annotated as relating to each other.
2. The annotation process itself works with a category tree that distinguishes between the domain and the argument level. The domain level refers to propositions which make substantive claims about international politics such as "war is wrong", "the US has invaded country xyz" or the like. In order to allow for a more detailed analysis, the model furthermore distinguishes between "assumption", "metatheory", "empirical reference", "domain question", "undefined" and "none". Empirical references are subdivided into counterfactual, hypothetical and evaluative evidences. The model does also distinguish between the categories of "basic assumptions" and "inferred assumptions" and thus allows for the consideration of theoretical contexts such as Realism, Constructivism, etc. Both categories are further subdivided into the subcategories of actors, structures, processes or outcomes. As a result of this category scheme, a detailed analytical framework emerges that allows the algorithm to systematically search for specific arguments and to relate them to theoretical conceptions.
3. Annotation on the argument-level is concerned with the illocutionary aspects of a sentence. Sentences can imply an assertion, a support of some other claim, a contradiction, or an attack on some other position. Annotating these attributes is important for providing guidance to the alghorithm how to present arguments when they are to be set in a discursive context. A Realist debating with a constructivist would, for example, most likely use different concepts on the domain level (emhasizing norms rather than interests) and opt on the argument level for a contradiction or an attack in order to undermine the thrust of a competing argument. Illocutionary annotations are undertaken independently of the material content of a sentence.

4.2. Data Set

The project sets up a dataset that comprises all open access articles of the most important English-speaking political science journals dealing with international relations[2] as well as records of relevant debates in the United Nations General Assembly.
All sentences together will build on a dataset consisting of approximately 800,000 annotated sentences, each with a specific domain meaning and a syntactic (illocutionary) meaning. In this process, subjective meaning is quasi-reified by being assigned an objectified meaning. An annotated sentence is no longer merely an author's subjective opinion or a recipient's interpretation, but becomes a datum with objective domain meaning, syntactic meaning, and a relation to another datum also with objectified domain and syntactic meaning. This basic sum of annotated sentences represents the raw mass by means of which the algorithm begins to search for specific arguments and patterns of domain and argumentation attributes. With each additional analytic category added to its repertoire, its sensitivity to additional patterns increases, and with each additional text, its ability to process additional statements grows. The resulting dataset allows the algorithm to be trained to identify argumentative patterns from assumptions, processes, and outcomes of different theoretical provenance, and to discriminate according to whether and with what kind of structure and evidence they are provided. The end result is an instrument that can be used to interrogate texts across theories and time with respect to their argumentative structures and to generate statements about the conditions of their reception or rejection.

4.3. Training

The project invests much effort in training. Here, the training of the algorithm (step 1) must be distinguished from its subsequent independent learning and further data processing (step 2). In step 1 of the training process, the algorithm is taught a basic repertoire of examples for the syntactic and domain-specific categories. It learns to identify arguments relating to theory-specific propositions, to tell, for example, an assumption from an empirical reference and to distinguish between different types of empirical references. For training the algorithm, we start with a set of four scientific articles, each representative of a certain theoretical perspective (Realism, Constructivism,

[2] The currently used text corpus comprises a total of 25 different scientific journals with a total of 1980 OpenAccess texts, which are available independently of institutional accesses. These are the American Journal of Political Science, British Journal of Politics and International Relations, Cooperation and Conflict, Ethics & International Affairs, European Journal of International Relations, European Journal of International Security, Foreign Affairs, Global Constitutionalism, Global Society, International Organization, International Security, International Studies Quarterly, International Theory, Journal of Common Market Studies, Journal of Conflict Resolution, Journal of Peace Research, Millennium: Journal of International Studies, Political Research Exchange, Politics and Governance, Politics & Society, Review of International Studies, Security Dialogue, Third World Quartely, West European Politics, World Politics

Liberalism, Feminism). Each of the articles covers 20-30 pages text and encompasses roughly 500 sentences. All of the sentences (a total of approx. 2,000 sentences) are annotated by ten individual annotators and discussed until they (or almost all of them) agree on a specific annotation. A high degree of inter-annotator reliability is to be achieved in order to provide the algorithm with an unambiguous guide line for how to proceed in step 2.
The second step grants the algorithm access to the corpus of open access articles of the most important English-speaking political science journals dealing with international relations. In this second annotation step, the algorithm is set up for (semi-)autonomous annotation and machine learning. In this process, it is closely guided by the original annotators and monitored to see if the annotations made match the understanding that was developed in step 1. This second step leads to a large argumentative repertoire of the algorithm, and thus a significant usability. The repertoire should allow both the systematic search for arguments by users as well as allow to infer statements about correlations of domain-level features and illocutionary arguments. This opens a promising way for answering the research question about the relevance of successful, i.e persuasive arguments and their domain- and illocutionary features.

5. Findings

The approach taken here to researching the relevance of arguments in scientific debates goes a qualitative step further than most previous social science projects. It does not count missiles, wars, or cash flows, but looks for argumentative patterns in complex communicative acts. Not material reality, but scientific exchange and thus communication about reality is made the object of knowledge. Such a combination of AI/ML and NPL for social scientific reflection as well as its relevance for political reality has not yet been attempted in this way and to this extent.

Even though SKILL is still in an early phase, first substantial findings can already be reported. The training of the annotators and the implementation of the first annotation exercises on texts from International Relations have underlined the need for, and difficulty of, assigning subjectively meaningful interpretations to an objectifiable schema. This difficulty is first expressed in the definition of separable categories at the domain level. On the one hand, the categories have to be specific enough to allow for a high degree of inter-annotator reliability. At the same time, they have to be sufficiently general to be applicable to different theories. What becomes clear in this process is that the structure of arguments in scientific texts is far more complex than in other text genres such as debate articles. The difficulty of objectifying subjective meanings is also evident in the fact that annotators and domain experts each work with subjective understandings about IR theories. Establishing an intersubjectively shared understanding thus requires not only mutual explanation but also a high degree of external understanding (Schütze et al 1973). This presents one of the greatest challenges: Is it possible to develop a sufficiently intersubjectively shared understanding of theory without one of the existing interpretations claiming hegemonic status and thus marginalizing equally valid interpretations? Or is it the case that the method of pattern recognition by necessity implies the setting of an exclusionary "gold standard"? Is ML and NLP thus necessarily establishing an algorithmic entity with a quasi-scientific "personality" which relies on specific interpretations of reality and which will hardly ever be more objective than its annotators?

A final remark relates to the status of theory in a data-driven approach: Social science has since many years been dominated by theory. Good scientific work was only too often expected to start with theoretical reflections and to use data only for illustrating its findings. Big data, ML and NLP reverse this methodological bias. The seemingly naïve question of "what is?", hitherto often rejected as unscientific, moves to the center in a pattern-oriented approach. An approach oriented at recognizing patterns must not be misunderstood as an analytical or theoretical tabula rasa, however. Exaggerated and misguided misunderstandings of pattern recognition circulate in the literature. Anderson, for example, fears that in the future digital data analysis will be able to do without researchers, since machines could also independently develop the necessary expertise that would be needed in the algorithmic research process (Anderson 2008, Müller and Ritschel 2016: 5). Such fears are based on a misunderstanding of how algorithm-based pattern recognition works. Algorithms can only recognize meaningfully at all, i.e. distinguish relevant from irrelevant, if they have criteria that allow them to make this distinction. An unguided search for patterns may allow the description of reality but will hardly allow any focused statements about scientifically relevant questions. Meaningful recognition therefore requires cognition-structuring analytical criteria. These criteria, in turn, cannot be drawn from a conceptual vacuum, but must be anchored in theoretical discourses. Just like any other social science question, a pattern recognition approach therefore requires a thorough connection to theoretical discourses.

References

Adler-Nissen, R., Eggeling, K.A. and Wangen, P. (2021): Machine Anthropology: A View from International Relations. In *Big Data & Society* 8 (2), 1-6

Anderson, C. (2008): The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Available online at https://www.wired.com/2008/06/pb-theory/

Berger, P.L. and Luckmann, Thomas (1967 [1966]): *The Social Construction of Reality*. First Anchor Books Edition. New York: Anchor Books.

Böcher, Michael (2022): Wie funktioniert wissenschaftliche Politikberatung? in *Forschung und Lehre*, 02.06.2022, https://www.forschung-und-

lehre.de/politik/wie-funktioniertwissenschaftliche-politikberatung-4759.

Jackson, P. T. (2011): *The conduct of inquiry in international relations. Philosophy of science and its implications for the study of world politics*. London, New York: Routledge

King, G., Keohane, R.O. and Verba, Sidney (1994): *Designing social inquiry*. Princeton, N.J., Chichester: Princeton University Press.

Kuhn, T.S. (1962): *The structure of scientific revolutions*. Chicago: Chicago University Press.

Luhmann, N. (1984): *Soziale Systeme: Grundriss einer allgemeinen Theorie.* Frankfurt am Main: Suhrkamp.

Mayer-Schönberger, V. and Cukier (2013): *Big Data: Die Revolution, die unser Leben verändern wird.* 3. Auflage. München: Redline Verlag.

Müller, T. and Ritschel, G (2016): Big Data als Theorieersatz? In T. Müller, G. Ritschel, A. Amberger, S. Böschen, R. Broemel, U. Busch et al. (eds.): Big Data als Theorieersatz. *Berliner Debatte Initia*l 4/2016. (2016) 4), 1–8.

Nassehi, A. (2019): *Muster: Theorie der digitalen Gesellschaft.* München: C.H. Beck.

Schütze, F., Meinefeld, W., Springer, W. and Weymann, A. (1973): Grundlagentheoretische Voraussetzungen methodisch kontrollierten Fremdverstehens. In Arbeitsgruppe Bielefelder Soziologen (Hrsg.): *Alltagswissen, Interaktion und gesellschaftliche Wirklichkeit* - Band 2. Reinbek bei Hamburg: Rowohlt, 433–495.

Teodorescu, D. and Andrei, T. (2014): An examination of “citation circles” for social sciences journals in Eastern European countries. *Scientometrics* 99 (2), 209–231.