Comparing Two Human-Computer Interactive Textual Analyses to Support Policymaking: Analyzing Interview Data for Advancing Educational Equity

Min Sun[0000-0001-5832-1534], Alex Liu[0000-0002-4785-1801], and Katherine Chang

University of Washington, Seattle WA 98195, USA
{misun, aleliux, kachang}@uw.edu

Abstract. Obtaining stakeholders' diverse experiences and opinions about current policy on a timely manner is crucial for policymakers to identify assets and gaps in resource allocations to support policy design and implementation. However, manually coding even moderately sized interview texts or open-ended survey responses from stakeholders can often be labor intensive and time costly. Although automated text analysis has promise to reduce these costs, policymakers may be less inclined to completely rely on an automated approach that is not based on disciplinary theories and policy contexts. Integrating human experts' inputs into automatic textual analysis may mediate policymakers' concerns. In this study, we compare two human-computer interactive learning approaches to analyze stakeholders' interviews about policies that either advance or hinder racial and economic equity in K-12 public schools in one U.S. state. With *computer-human parallel* analysis, human coding guided by a domain-specific theory happens in parallel with unsupervised topic modeling. In the second *computer-human sequential approach*, unsupervised topic modeling occurs first and then human coders use the initial theme discovery to develop a codebook and then code the interview data. While each approach offers nuances and allows human experts to use their domain knowledge to validate, interpret, and supplement the computer analysis results, the *computer-human sequential* approach offers a better integration of the advantages of both computer and human coding to enable a faster generation of evidence for policy decision-making.

Keywords: Human-centered Machine Learning, Educational Equity, Natural Language Processing, Policy.

1 Introduction

Policymakers are seeking reliable, valid, and meaningful evidence to support decision-making in a timely manner. One important source of policy evidence involves stakeholders' lived experiences about the implementation of the current policy and their opinions about how to improve [1, 2]. These stakeholders are individuals and organizations who care about or are otherwise affected by a given policy pertaining to its creation, enactment, and evaluation. Useful insights can be gathered through stakeholders' interviews, open-ended surveys, or texts obtained from their social media posts [3,

4, 5]. These qualitative data about stakeholders' insights, mixed with quantitative causal analysis of policy impact, are often used in policy analysis. In reality when decisions have to be made within a short time frame, the cost of manually analyzing even a moderately sized text may hamper the actual use of stakeholders' voices [6].

Unsupervised machine learning (ML) methods, such as topic modeling, can discover latent themes (or topics) from unstructured texts with speed and at scale. Yet results from automatic topic modeling may not always be accurate, because this approach merely discovers semantic structures in a text body. The process of classifying documents is not particularly guided by domain knowledge related to the policy issue of interest (e.g., racial and economic equity in K-12 public education) or considers the specifics of policy implementations in a local context [6]. Therefore, policymakers may be skeptical to rely on topic modeling results for high-stake, contextualized political negotiations.

Policymakers and researchers commonly concern about two types of validity [7]. First, *construct validity* concerns the degree to which the measures or indicators reflect the underlying concept policymakers aim to capture. Specifically in the context of topic modeling, construct validity can be indicated by the ability to name identified topics with specific construct names chosen from the relevant substantive theory and existing knowledge, guided by operational specifics of the policy program, and situated in local contexts. Another strategy is to assess whether a given topic includes documents that are correlated with each other to confer one common substantive concept (i.e., convergent validity) and at the same time, these documents under the same topic correlate with each other more highly than they correlate with documents under a different topic (i.e., discriminant validity). Second, the *criteria validity* refers to how well topic modeling results perform against a set of standards of "truth" and the utility for policy communication and stakeholders' action-taking. Prior studies of validating natural language analysis (NLP) results used independent human expert coding as the gold standard of "truth." Criteria for the merit of policy evidence are not limited to validity-as-truth. The merit of the evidence can also be indicated by its utility to facilitate stakeholders to interpret, communicate, and use to take actions [8, 9, 10]. Our study aims to illustrate how human-centered ML applications—that involve the iterative interactions between human, computer, and domain knowledge throughout the analysis process—can be used to assess these two types of validity relevant for policy evidence.

This paper is situated in a larger study of identifying policies and programs that either advance or hinder racial and economic equity in one U.S. state—Washington State's— K-12 public school system in 2022. As part of the policy evidence, the project conducted the interviews with a diverse group of stakeholders who were state legislators, other state-level policymakers, school district administrators, teacher union representatives, teachers, policy advocates, and community leaders. To analyze the interview data in a timely manner to support Washington State's educational equity policy redesign, we compared two human-centered ML methods. With *computer-human parallel* analysis, human coding guided by a domain-specific theoretical framework happens in parallel with unsupervised topic modeling. In the second *computer-human sequential approach*, unsupervised topic modeling occurs first and then human coders use the initial theme discovery to develop a codebook and then code the interview data. While

each approach offers nuances and allows human experts to use their domain knowledge to validate, interpret, and supplement the computer analysis results, the *computer-human sequential* approach offers a better integration of the advantages of both computer and human coding to enable a faster and valid evidence generation. The findings offer useful insights into the design and evaluation of computer-human interactive learning to advance evidence-based educational policymaking, an area in which artificial intelligence (AI) has not widely applied to transform its conventional practices.

2 Related Work

Topic Modeling Validation. Automated content methods can make the previously impossible possible to understand stakeholders' narratives about a given policy by systematic analysis of a large text collection without massive labor or time investment. Yet, the complexity of human language implies that automated content analysis cannot simply replace human's careful and close reading of texts. The output of automated text analysis may be incomplete or misleading. Therefore, automatic methods are "best thought of as amplifying and augmenting careful reading and thoughtful analysis" [6]. It is incumbent upon the researchers to validate their use of automated text analysis.

Prior literature has used a variety of strategies to validate unsupervised textual analysis results, often involving (a) comparing results with human expert coding of the same data, (b) comparing the results with alternative data sources about the same phenomena of interest, (c) predicting criteria measures. To illustrate, after applying the unsupervised model to analyze how senators present their work to constituents, Grimmer [11] developed a codebook and asked a research assistant to classify a portion of the documents according to the codebook. The correlation between human and computer coding is 0.96 [6]. Sun and her colleagues [12] applied structural topic modeling to analyze textual data about over two million reform tasks K-12 public schools designed and implemented to turn around the persistently under-performing schools as measured by student achievement on state standardized tests in mathematics and English Language Arts in the state of Washington. The prevalence of identified reform strategies was largely consistent with school leaders' own perceptions of reform priorities via interviews. Several reform strategy measures as indicated by topic proportions were significantly associated with reductions in student chronic absenteeism and improvements in student achievement. Baumer et al. [13] used grounded theory—an interpretive qualitative method widely used in social science [14]—and topic modeling to analyze the same survey data. The results show that the two analyses produce some similar and some complementary insights about the phenomena of interest, in their study, the non-use of social media. This comparison suggests future research to explore mixed computational-interpretive methods that combine human coding and computer textual analysis in novel and compelling ways to advance knowledge discovery in social sciences.

Resource Equity Policy. Drawing on prior research in educational equity policy [15], we conceptualize a Resource Equity framework that includes six essential components

of educational policies that influence equitable student learning experiences and outcomes in schools. The "inner circle" of policy strategies that are most proximate to students and have direct impacts on student learning includes: (1) the diversity and qualifications of *school staff* (teachers and other adults) who have close interactions with students in schools; (2) the *curriculum and instruction* that enable teachers and students to actively engage with rigorous and culturally relevant learning content; and (3) other types of *student support and intervention* programs, such as mental health and social work services, multi-tiered support systems, summer school, and tutoring, which directly support students outside and around the classroom. The "outer circle" of support includes (4) *school finance* that allocates resources to schools to support the offering of educational services, (5) *school governance, accountability, and partnership* that determine school decision-making structure and power dynamics among stakeholders, and (6) the *data and information* that either enable or constrain the design, implementation, and evaluation of all the previous five components. Policy and practices pertaining to each component at each level of the school system and across the hierarchy of schooling systems are embedded within a continuous cycle of improvement. We prioritize these six factors because they pertain to the core practices of public K-12 systems from classroom to state, and because they are malleable from a policy and practice perspective. At the same time, we acknowledge the influences of the *historical conditions and local contexts* on the flow and accumulation of resources, supports, and social conditions in ways that impact student and family experiences in- and outside of the school building. Thus, it is critical to strategically incorporate stakeholder voices from diverse racial backgrounds, professional experiences, and geographic locations in the state.

3 Data and Analysis

3.1 Data Collection and Preprocessing

The data in this study mainly include 24 interviews with a wide range of policy stakeholders. Our purposeful sampling strategies maximize the following criteria [16, 17]: (a) the level of public school systems: classroom, school, district, and state; (b) geographical locations in the state; (c) roles of the interviewees: system actors at different levels of educational systems; three branches of the government at the state level: non-system actors including community organization leaders, advocates and lobbyists, teacher union representatives, philanthropic organizational leaders; and (d) students' and local communities' characteristics in terms of race/ethnicity, socioeconomic status, language and homeless populations. Interviews were conducted through Zoom.com. Each interview lasts about 45-60 minutes.

Moreover, we had deliberately kept the semi-structured interview questions broad to allow a wide range of topics or ideas to naturally emerge (see Online Appendix A1). We asked interviewees to offer a few examples of current state and local policies that they thought most enhance or limit racial and economic equity in Washington state's K-12 public education system. We also probed why and how they thought that way. We then asked them to comment on their access to reliable data and evidence to support

policy development and implementation, as well as their suggestions on state and local policy iterative improvement.

Audios were transcribed to texts. Our research assistants listened to all 24 audio recordings and made necessary corrections to the texts and sentence structures. After initial cleaning, a tidytext-format data contains about 1,700 entries (i.e. documents). Each document captures one complete thought of one interview, which could be one or multiple sentences. Although the text corpus might not be considered as being "big", they provide enough data for computational methods to produce meaningful results but not so much data that iterative human coding becomes intractable. The dataset also includes interviewees' research ID, demographics, job roles, and job location.

3.2 Methods

Drawing on human-centered ML, we designed two approaches: computer-human parallel analysis and computer-human sequential analysis [18]. We started with the approach of *computer-human parallel analysis*, in which three expert coders, led by one co-author (Chang), used a grounded theory approach to develop a codebook based on multiple iterations between their sense-making of samples of interviews and guided by the Resource Equity framework [14]. The three expert coders who have both education policy research knowledge and had extensive experience in working in K-12 schools coded a common set of interview data to establish inter-rater reliability and then they separately coded the rest of the interviews. During the individual coding process, these three coders met weekly to discuss issues and develop consensus to maintain their inter-rater reliability.

In parallel, we used topic modeling, specifically using LDA (latent Dirichlet allocation), to explore themes and patterns based on latent semantic analysis of probability distributions of words and phrases in the interview data [19]. LDA assumes that each document (one complete thought from an interviewee in this study) is a mixture of topics. For each document, π_{ik} represents the proportion of one document i dedicated to topic k. Each task collects the proportions across topics, as $\pi_i = (\pi_{i1}, \pi_{i2}, \dots, \pi_{iK})$. We used an R package (-stm-) to implement the analysis. This mixed membership representation for each document is necessary for analyzing our interview data because when interviewees discussed about school funding, for instance, they could be speaking on other topics like teacher resource distributions or school governance. However, the mixed membership also presents challenges to interpretability [20]. To address this, the other two coauthors of this paper (Sun and Liu) developed rubrics to manually validate themes generated from the topic modeling. Sun and Liu read 20 documents with the highest topic prevalence (the proportion of a document discussing a given topic) and the top 10 most frequent words to interpret the meaning and coherence of each topic. Out of 30 topics, 25 topics are rated as being theoretically coherent and practically relevant to the policy of interest within Washington State's contexts. We then computed the cosine similarities between topic modeling results and qualitative human codes with and without *tf-idf* measures.

The grounded theory approach involves human interpretations and theory-guided inquiries. However, this method is time-consuming. From codebook development to

the competition of all coding, three expert coders spent about 20 hours per week over 3 months (about 720 hours). This process may also be subject to conceptual limitations and subjective decisions of the researchers. While this computer-human parallel process offers independent knowledge discovery and compares topic modeling results with the criteria of human coding, this process did not integrate the "best" of both worlds in one coherent process. We thus developed the second approach.

The second approach of *computer-human sequential analysis* uses computer-assisted text analysis to assist with codebook development and then uses human coders to code the interview data using the developed codebook. Since the three coders at the first stage spent most of three-month's time on codebook development, we explored if we could use the computer-assisted topic modeling to speed up the sense-making of raw interview data, discover themes, then use human experts to code the interviews and supplement topic modeling results drawing on the conceptual framework and disciplinary knowledge. Once a codebook was developed, we trained two doctoral research assistants—who were not involved in the previous stage of coding and who have substantial training in educational policy in Washington state—to code the entire interview data. After establishing inter-rater reliability, these two research assistants coded the most prevalent theme for each document and identified the secondary or third theme if appropriate. These two coders were able to finish the coding using 42 hours combining their time. Adding the codebook development, the second stage took about 60 hours in total, which is a fraction of the 700 hours at stage one. We then compared the themes identified by human coders and the most prevalent topics assigned by the computer for each document.

To assess the predicative validity of the *computer-human sequential* approach, we summarized the themes by stakeholders' job roles. We hypothesized that during the interviews, teachers and teacher mentors might spend more time discussing staffing practices in terms of teacher recruitment, retention, professional development, and mentoring, and/or policies about curriculum and instruction, because they have lived experiences and most relevant knowledge among all stakeholders. In contrast, district and state administrators might focus their attention on topics such as school finance, resources allocation, and politics and bargaining with teacher unions and other interest groups. If the analysis results show the same pattern as we hypothesized based on domain knowledge, this would add further credibility to the *computer-human sequential* approach.

4 Results

4.1 Topic Modeling

The LDA approach requires researchers to specify the number of topics. We used several diagnostic statistics to aid this process, including exclusivity, residuals, semantic coherence, and lower bound as shown in **Fig. 1**. Models with topics between 25 to 35 offer a balance among all four statistics, as indicated by the elbows of more rapid changing slopes of the trajectory lines of these values. To further zoom in and compare models with topics of 25, 30 and 35, we mainly rely on exclusivity and semantic coherence

statistics, which are indicative of discriminant and convergent validity for policy evidence. As shown in
Fig. 2, the 30-topic model has the largest proportion of topics that possess both exclusivity and semantic coherence (towards the upper right corner) and the smallest proportion of topics in the lower left or left size, which indicate a lack of semantic coherence.

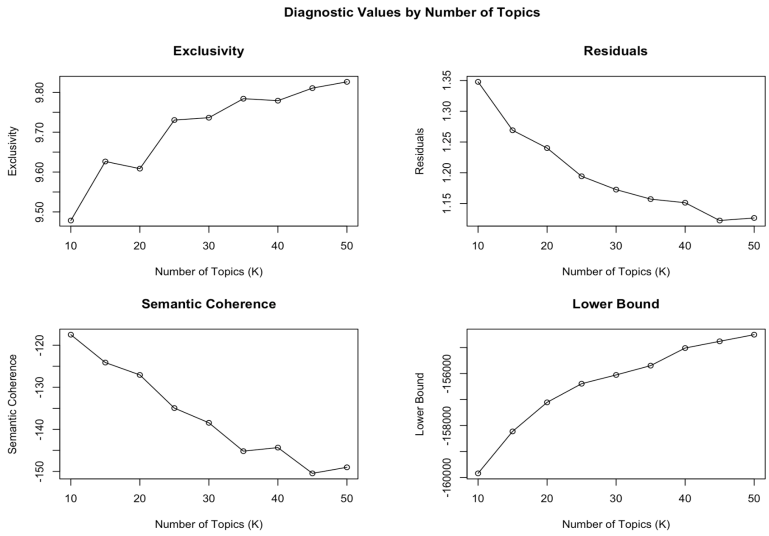

Fig. 1. Diagnostic statistics by number of topics.

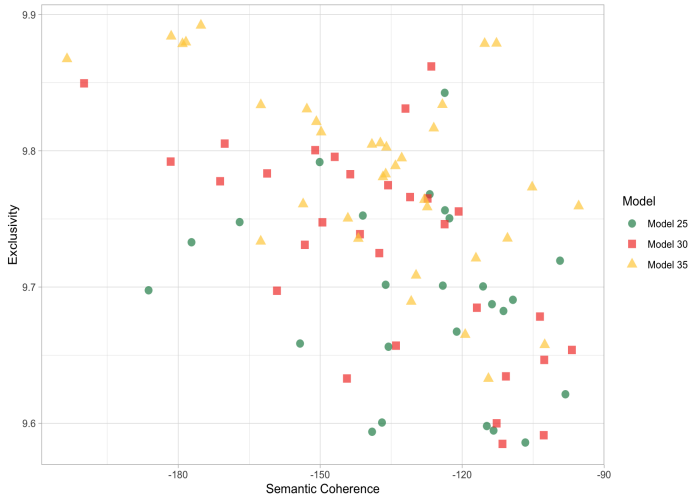

Fig. 2. Comparing the semantic coherence and exclusivity of three topic models

Table 1. Topic descriptions and coherence ratings.

Topic #	Parent Code	Child Topic Code	Topic Label	Expert Coherenc Rating	Keywords
1			(low coherence)	1	figur, talk, power, people of color, built, month
2	Leadership and governance	Leadership and governance	Relationships between state and local public school systems (decentralization and local control) and relationships between public school systems with non-profit organizations	3.5	posit, question, level, district, local, respons, come, differ, engag, ask, esd, foundat, begin
3	leadership and governance	Local control or school board	School boards' role in decision making (particularly, budgeting) under local control and relevant stakeholders	4	budget, school-board, decis, sometim, make, stakehold, member, director, engag, power, superintend
4	Student interventions	Socio-emotional learning (SEL) and	Resources supporting youth's SEL and mental health	4	mental, health, resourc, young, space, assist, person, social, larg, peopl, academ, best, care
5			(low coherence)	1	attent, need, better, understand, give, hour, help, job
6	System interventions	School system support	Tier support system and school improvement	4	tier, school, high, three, improv, two, year, score, attend, survey, foundat, wsif, elementari
7	System interventions	Judicial systems	Court, judicial systems, and institution's role in racial equity	4	court, state, washington, system, committe, counti, gap, educ, covid-19, repres, justic, institut
8	School finance	Funding formula	Funding formula based on districts' needs, local levies, and allocations of McCleary and Stimulus funds	4	formula, money, fund, spend, financ, district, dollar, levi, school, essa, amount, distribut, mccleari
9	Leadership and governance	Coalition and relationship	Going beyond superficial things (counting and disaggregating numbers) and building relationships and coalition centering the voices of youth and famiies in marganized communities	3	work, tri, hard, nativ, mani, educ, push, tell, center, peopl, excit, voic, talk, famili, count
10	Staffing policy	Teacher union, salary, workforce	Teacher union's politics, local collective bargaining processes, and union influences on teacher salaries and hiring	4	union, teacher, salari, princip, hire, administr, contract, qualiti, local, survey, control, valu, district
11	Data and information	Data access, analysis, use	Data collection, access, analysis and use	4	dashboard, data, inform, assumess, collect, disaggreg, report, websit, access, use, ospi
12			(low coherence)	2	rais, program, number, disciplin, cohort, run, parent
13	Instruction and curriculum	Teaching and learning	Teaching and learning (online and hybrid), teacher training, curriculum, student experience	3	onlin, teach, taught, teacher, high, learn, class, middl, scienc, elementari, experi, student
14	Culture, climate and environment	Anti-racism	Whiteness, barriers for children's success, and anti-racism	3	pull, white, child, success, stay, built, job, indic, keep, whole, four, barrier, stori, middl
15			(low coherence)	1.5	incred, peopl, ask, complic(ated), hold, financ, brown
16	Staffing policy	Teacher mentoring and coaching	New teachers' coaching, mentoring, and professional development in equitable instruction and curriculum (adopt, professional-learn)	4	coach, new, classroom, instruct, role, teacher, practic, equiti, cultur, term, sure, curriculum
17	Leadership and governance	School board	School board diversity and representation, accountability system, and inquiry cycles	4	board, plan, perspect, staff, answer, identifi, action, guess, educ, suppos, one, experi, idea
18	School finance	Targeted funds	Targeted funds for specific activities/programs: funds available for other purposes than teacher salary, funding for teacher prep, and targeted investment in community of color	3	paid, pay, grant, fund, avail, tax, seattl, whether, anoth, obvious, black, spend, one, teacher
19	Data and information	Goals and oucomes	Education and school improvement goals, school courses, educational outcomes, and jobs	3.5	kind, cours, goal, access, outcom, sure, citi, drive, kid, take, pathway, possibl, easi, deeper
20	Leadership and governance	Community	Building school system's capacity to engage with communities	4	capac, term, agenc, build, communiti, depart, collabor, limit, engag, within, understand
21	Student interventions	Differentiated student strategies	Differential and targeted strategies for students of color, particularly American African, Southeast Asian, Hispanic	4	african, group, student, includ, american, focus, achiev, strategi, target, race, specif, goal,
22	Student interventions	Multilingual programs	Access and quality of bilingual, multilingual programs offered to English lanaguage learners	4	dual, languag, program, english, bilingu, learner, student, year, servic, ell, skill, research, access
23	Culture, climate and environment	Trauma at home	Struggling home and family experiences of children of high poverty and of color negativey influence their school learning and graduation pathways post pandemic	4	bad, kid, children, home, school, famili, grade, covid-19, hispan, pandem, third, rate, happen
24	Staffing policy	Diversifying teacher workforce (teacher labor market)	Teacher education, diversifying teacher candidate pool/ pathways, and partnerships between K-12 and higher edu, and with outside organizations	3	field, chang, part, pathway, teacher-educ, local, forc, discuss, invest, partnership, import
25	School finance	Progressive funding	Special education, federal and state funding for districts with low-income, ELL, special-ed, and multicultural students who need more supports	4	special-, student, feder, fund, dollar, district, mention, challeng, addit, popul, identifi
26	Data and information	Tests, standards, measures, and graduation requirements	Tests, standards, graduation requirements, college readiness, and measure and data disaggregation by race and ethnicity	3.5	test, standard, score, tribal, take, math, assumess, let, measur, rate, indic, enrol, consult, colleg
27	Student interventions	Learning opportunities and programs	Learning opportunities in schools, and tribal and native education	3	leav, opportun, year, differ, five, nativ, tribe, type, offic, back, time, last, far, legisl, whole, curriculum
28	Leadership and governance	Leadership diversity	Leaderhsip, diverse community, and leadership in diversity	3	add, support, divers, nativ, trust, leadership, indian, broad, communiti, provid, locat, hispan
29			(low coherence)	2	enhanc, polici, equiti, racial, procedur, econom, law
30	Governance and leadership	Legislation process	Bills and legislation process	3.5	bill, pass, half, one, read, last, educ, way, year, hous, legisl, make, start, actual, basic

Next, two authors (Sun & Liu) developed rubrics to manually rate the extent to which the identified topics were practically meaningful, consistent with the literature, and semantically coherent (see Online Appendix Table 2 for the rubric and coding procedure)

[21]. Using a scale of 1–4, we first independently labeled each topic and rated its coherence by reading a sample of tasks with the highest loadings on a given topic. We then compared and discussed our ratings. Interrater reliability is high, as measured by Krippendorff's alpha=0.7. Most of the differences were only a 1-point difference (e.g., one coder rated a topic's coherence as a 2, while the other rated it a 3). If the coders could reach agreement, we adjusted our individual scores to the score we agreed to. If we could not, we preserved the original ratings and took the averages of them. Low-rating topics (<3) were not included in the next stage of human coding. Table 1 illustrated the topic descriptions and coherence ratings: the column of Parent Code column includes the aggregated categories that align with the conceptual framework in section 2.2. Child Topic Code includes the short description of the topic label. Topic Label includes the labels agreed by the two authors of this paper. The column of Expert Coherence Rating includes the agreed rating by these two coders and the Keywords include the most frequent words for each topic.

4.2 Computer-Human Parallel Analysis

We then compared topic model labels with the first round of parallel human qualitative codes. Overall, we observed human and computer produced similar results, with average cosine similarity of 0.65, ranging from 0.46 to 0.80. This finding is consistent with prior studies in other fields [13]. 80% of the 25 coherent computer-generated topics are also labeled as the same at either the parent code level or child code level by human coders. Noticeably, computer topics with high expert coherence ratings are more likely to align with human qualitative codes. 12 out of 14 topics with expert rating of 4 match the same human qualitative codes at either parent or child level, of which 10 topics match at both parent and child code levels. Table 2 illustrates a few topics that have matched computer codes and human codes at both parent and child code levels. Online Appendix Table 3 includes the full list of matching between computer and human codes.

Table 2. Computer-human parallel text similarities.

Topic #	Expert Coherence Ratings	Computer Parent code	Computer Child Topic code	Human Parent Code	Human Child Code	Cosine similarity
2	3.5	leadership and governance	leadership and governance	leadership, governance, and structures	governance structures	0.55
11	4	data and information	data access, analysis, use	data, measurement, and information (DMI)	data access	0.8
16	4	staffing policy	teacher mentoring and coaching	staffing resources	teachers training support	0.76
20	4	leadership and governance	community	leadership, governance, and structures	community level strategies	0.61

4.3 Computer-Human Sequential Analysis

In this *computer-human sequential* approach, human experts manually labeled the most prevalent themes of all documents based on the codebook developed from the topic modeling results. Human coders labeled almost all documents with at least one code and up to three codes, of which 98% were labeled with child topic codes and the remaining 2% were labeled with parent codes only. In addition, 33% of documents received a secondary code in addition to the most prevalent code and only 3% of document received a tertiary code.

We then compare those human qualitative codes for each document with the top three computer-generated topics based on the highest topic proportions. The result shows 83% of the documents have at least one pair of matching computer topic and human-assigned code at the parent-code level, and 49% matched at the child code level (at least one of human-labeled themes matched to one of the topic modeling labels). This finding suggests again that human and computer agreed on the most prevalent themes of some documents, while they also complement each other. It is worth to note that the second stage similarity comparison is stricter than the first stage of *computer-human parallel* analysis in that the first stage compares overlapping texts while the second stage examines similarity in topic assignment of the top three most prevalent themes and at the individual sentence/paragraph level.

4.4 Thematic Patterns by Interviewees' Roles

As shown in **Fig. 3** on next page, human-and computer-coding results largely suggest the same patterns of topic distribution among these three types of stakeholders' job roles: administrators and policymakers, educators, non-profit and advocacy organizations. Compared to the other two job categories, educators including teachers and teacher mentors were more concerned about teacher mentoring and coaching, teacher recruitment and retention, teacher union and salary, and instruction and curriculum. In contrast, administrators and policymakers had the knowledge and purviews related to school finance, such as distributing targeted funds, progressive funding practices, and funding formula revisions. It is also not surprising to see that the topics of "governance, leadership, and community" and "data access, analysis, and use" were prevalent among all interviews since all stakeholders had relevant experiences and knowledge to discuss these topics. This finding suggests the predictive validity and practical utility of *computer-human sequential* analysis results to explain social phenomena, in our study understanding different stakeholders' policy interests and foci.

5 Conclusion

This paper presents two interactive approaches between computer and domain experts to analyze a corpus of interview data about policy stakeholders' opinions and practices to advance educational equity. The second approach of *computer-human sequential* approach is much more time efficient, while from a knowledge discovery approach, each approach offers nuances. The first approach of human parallel coding using a grounded

theory method offers details that cannot be easily captured by automated textual analysis, while the second approach allows human experts to code interview data in a much more efficient way, and at the same time, allows human coders to deviate from the computer results by incorporating their own domain knowledge and interpretation of the data. The sequential approach offers a better integration of the advantages of both computer and human coding by using the machine's computing power to boost human coders' analytic capabilities, while human experts bring domain knowledge to validate, interpret, and supplement the computer-analysis results. This interactive learning enables a faster and valid evidence generation to support policy analysis.

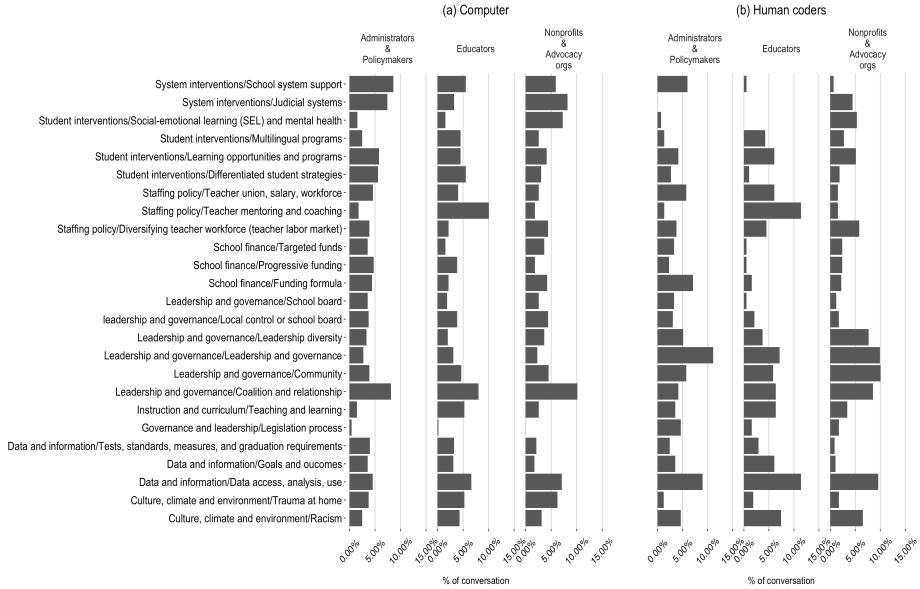

Fig. 3. Fraction of topics discussed by interviewees' job roles. Topics were assigned by computer (left) and human-coding (right).

Online Appendix

https://github.com/AlexLiuxx/Comparing-Two-Human-Computer-Interactive-Textual-Analysis-to-Identify-Evidence-for-Policymaking

References

1. Davidson, P., Arndt-Bascle, C., Liedekerke, M.-G. de, & Reyes, R.: Improving stakeholder engagement and evidence-based policy making. https://www.theregreview.org/2022/12/07/davidson-improving-stakeholder-engagement/, last accessed 2023/01/14.
2. Fedorowicz, M., & Aron, L. Y.: Improving evidence-based policymaking: a review. https://www.urban.org/sites/default/files/publication/104159/improving-evidence-based-policymaking.pdf, last accessed 2023/01/14.

3. Berry, K. S., & Herrington, C. D.: Tensions across federalism, localism, and professional autonomy: social media and stakeholder response to increased accountability. Educational Policy, 27(2), 390–409 (2013).
4. Rosenberg, J. M., Borchers, C., Dyer, E. B., Anderson, D., & Fischer, C.: Understanding public sentiment about educational reforms: the next generation science standards on twitter. AERA Open, 7 (2021):
5. Wallner, J.: Legitimacy and public policy: seeing beyond effectiveness, efficiency, and performance. Policy Studies Journal, 36: 421-443 (2008).
6. Grimmer, J., & Stewart, B.: Text as data: the promise and pitfalls of automatic content analysis methods for political texts. Political Analysis, 21(3), 267-297 (2013).
7. Shadish, W. R., Cook, T. D., & Campbell, D. T.: Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton Mifflin, Boston, MA (2002).
8. Kvale, S.: The social construction of validity. Qualitative Inquiry, 1(1), 19–40 (1995).
9. Patton, M. Q.: Utilization-Focused Evaluation: The New Century Text. 3rd edn. Sage Publications, Thousand Oaks, CA (1997).
10. Peck, L. R., Kim, Y., & Lucio, J.: An empirical examination of validity in evaluation. American Journal of Evaluation, 33(3), 350–365 (2012).
11. Grimmer, J.: Appropriators not position takers: the distorting effects of electoral incentives on congressional representation. American Journal of Political Science, 57: 624-642 (2013).
12. Sun, M., Liu, J., Zhu, J., & LeClair, Z.: Using a text-as-data approach to understand reform processes: a deep exploration of school improvement strategies. Educational Evaluation and Policy Analysis, 41(4), 510–536 (2019).
13. Baumer, E.P.S., Mimno, D., Guha, S., Quan, E. and Gay, G.K.: Comparing grounded theory and topic modeling: extreme divergence or unlikely convergence?. Journal of the Association for Information Science and Technology, 68: 1397-1410 (2017).
14. Glaser, B., & Strauss, A.: The Discovery of Grounded Theory: Strategies for Qualitative Research. Aldine Publishing, Chicago, IL (1967).
15. Dimensions of Equity, https://www.educationresourceequity.org/dimensions, last accessed 2022/01/16.
16. Maxwell, J. A.: Causal explanation, qualitative research, and scientific inquiry in education. Educational Researcher, 33(2), 3–11 (2004).
17. Patton, M. Q.: Qualitative Evaluation and Research Methods. 2nd edn. Sage Publications, Thousand Oaks, CA (1990).
18. Aragon, C., Guha, S., Kogan, M., Muller, M., & Neff, G.: Human-Centered Data Science: An Introduction. MIT Press, Cambridge, MA (2022).
19. Blei, D.M., Ng, A.Y. and Jordan, M.I.: Latent dirichlet allocation. The Journal of Machine Learning Research, 3, 993-1022 (2003).
20. Grimmer, J., Roberts, M. E., & M., S. B.: Text as Data: A New Framework for Machine Learning and the Social Sciences. Princeton University Press, Princeton, NJ (2022).
21. Mimno, D., Wallach, H.M., Talley, E.M., Leenders, M., & McCallum, A.: Optimizing semantic coherence in topic models. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 262-272. Association for Computational Linguistics, Edinburg, Scotland, UK, (2011).