Subject: Fwd: Høringer klima og miljø
Date: Sunday, 17 March 2024 at 19:51:33 Central European Standard Time
From: Karl Henrik Sivesind
To: Nicolai Sivesind

Hei,
Her ser du opplegget fra et PhD prosjekt. Det er kanskje litt for omfattende for dere, men der kan dere se hva som kan gjøres.

Karl Henrik

Fra: Vibeke Wøien Hansen <v.w.hansen@samfunnsforskning.no>
Sendt: Sunday, March 17, 2024 5:01:47 PM
Til: Karl Henrik Sivesind <k.h.sivesind@samfunnsforskning.no>
Emne: Re: Høringer klima og miljø

Hei Karl Henrik,

Så bra! Klima og miljø høres lurt ut. Ja, de må da skrape høringer fra nettsidene. Man kan starte med oversikt-datasettet over høringene til departementene (se lenke under) og bruke den infoen til å systematisere skrapingen.

I PhD-en sin som snart skal leveres og som ser på norske og EU-høringer skriver Idunn Nørbech en del om metode, klipper inn noen utdrag her tilfelle nyttig for Nicolai og co. Til deres oppgave blir det jo et mindre omfang enn hva som er gjort under og fokus på input i tekst fremfor hvem som har deltatt: men tenkte det var greit å gi info om fremgangsmåte brukt for tror det er mulighet for noen tips fra denne. Hun bruker ikke høringene til direktoratene fordi disse er mindre standardiserte og skriver: "In constrast, we noticed a lot more variation in how Norwegian executive agencies provide online information about consultations held and received stakeholder feedback. This creates important empirically challenges to mapping empirically the patterns of consultations and stakeholder participation across agencies, and informed our decision to focus our empirical analysis on consultations organised by the Norwegian ministries only." Men om Nicolai og co kun skal se på et direktorat funker det jo likevel, men de/vi må da være klar over at man ikke kan overføre fremgangsmåte til et annet direktorat.

Idunn og co tok utgangpunkt i denne nettsiden og datasettet som ligger der da skulle skrape:

https://www.regjeringen.no/no/dokument/hoyringar/oversyn-over-hoyringssaker/id546535/

Hun skriver:
"The final dataset, for Article 4, was the result of an ambitious data collection effort. The actual data collection was fully automatised using first a dataset provided online by the Government containing titles, dates, and links of the Norwegian government consultations. Using web-scraping we could then collect the names (and stakeholder types when provided) of 251,153 participants in the 4,062 consultations. We then

later decided to also collect the list of 325,332 invitations that the government sent out to stakeholders nested within 3,952 consultations .A challenge working with the Norwegian government webpages compared to the EC,is that the web pages for each consultation do not always use the same format. For the Ministry invitations, there was seemingly random variation in how the Ministries presented this information. This required accommodating different paths in the data collection script and using additional tools like Rselenium to gather information from JavaScript tables. After collecting the data, there were also challenges in categorizing the participating stakeholders into stakeholder types because the dataset is so large, and only around 100,000 observations were already categorized. Several steps were therefore taken in order to categorize the participants. First, because many stakeholders participate several times, these could be matched by name and then if they at one point had self-selected into a category, this category could then be applied to all observations with the same name. However, because stakeholders self-selected, they did not always select the same category across consultations, therefore the category they had most consistently chosen across consultations was chosen. The stakeholder names also contained a lot of typos, and different spellings (e.g. “LO”,“Landsorganisasjonen LO”, “Landsorganisasjonen”, “Landsorganisasjonen (LO)”etc.) between consultations, which made the application of automated methods more challenging. Next, we characterized stakeholders based on terms in their names (like “Kommune”, “Direktorat”, “Skole”, etc. along with common first names for citizens). In the end, this left about 20,000 stakeholders which were manually coded into categories (see appendix for Article 4 for codebook). In the end, the stakeholder categories were used to calculate a stakeholder diversity index at the consultation level, so the full potential of the large dataset remains open for future exploration. Finally, it is worth noting that the stakeholder type categorization provided by the Norwegian government is not as detailed as the EC categories, making comparisons between the participation of different stakeholder types between the two polities difficult. The most obvious example of this is the Norwegian government’s stakeholder category of “Consumer – or interest group” which is broad and includes everything from professional associations to patient groups. In the future, a more detailed categorization of the stakeholders would be beneficial, however, it would be a time-consuming endeavour.

Om EU-konsultasjoner og en variabel som måler støtte for et forslag gjennom tekstanalyse/gruppering av støtte (for/mot):

"We use a semi-supervised machine-learning model called Latent semantic scaling (LSS) (Watanabe 2021) to measure stakeholders’ support for proposals expressed in online open-text feedback and/or pdf. attachments.

Before applying the LSS model we pre-processed the texts, by removing URLs, punctuation, numbers, and common English stop words. LSS requires a small set of ‘seed words’ to calculate document-level polarity scores (Watanabe 2020). We decided on which ‘seed words’ to use based on the actual feedback texts and considerations linked to our dependent variable: support for legislative proposals. We chose the following positive seed words: “support”, “agree”, “welcome”, “ambitious”, and “encourage”, as well as their variations. The negative seed words were: “do not support”, “failure”, “disagree”, “bad”, “difficult”, “against”, “reject”, ”challenges”, ” amend”, ”adjust”, ”clarify” ,and “recommend” (the last based on the observation that it was most often used in the context of feedback texts in which the stakeholder recommended changes or amendments to the proposal thus implicitly expressing less support for the current version of the proposal). We further specified that these words should be used in the context of the following terms: “proposal”, “initiative”, “measure”, “article”, “regulation”, “directive”, “decision” and “definition”. This ensure that the model captures better stakeholder stance towards the proposal. We also exclude some words that are often mentioned in this context but don’t carry any substantial meaning for the polarity scores like “the Commission”, “EC”, “European Commission”. "

Dette er veldig spennende! Håper noe av infoen over kan være nyttig. Det hadde vært supert om de f. eks. kunne lage grupperinger av aktørene: koalisjoner/de som er enige med hverandre. Dette kan også brukes til å bestemme vinnere og tapere av politikk som bli vedtatt.

Fortsatt god søndag,

Vibeke

From: Karl Henrik Sivesind
Sent: 16 March 2024 11:57:37
To: Vibeke Wøien Hansen
Subject: Høringer klima og miljø

Hei,
Nicolai mfl leter etter store tekstbaser til å trene KI på for "argument mining", og de trenger det raskt til en studentoppgave. Jeg nevnte at klima og miljø er bra fordi det er relativt klare motsetninger som er greit for informatikk studenter som kan koding men ikke så mye om interessegrupper og lobbying. Må man scrape dataene fra nettsidene til departementet og direktoratet?

https://www.regjeringen.no/no/dokument/hoyringar/id1763/

https://www.miljodirektoratet.no/hoeringer/avsluttede-hoeringer/

Har du noen forslag? Det kan jo være interessant og se hvordan de griper det an med utgangspunkt i KI-verktøy.

Fortsatt god helg!

Karl Henrik