Underneath US Rulemaking:

An Analysis of Legislative Focus Through Public Comments

—

Puran Dou
Yuyan He

Agenda

- Background & Motivation
- Data Collection
- Methodology
- Results
- Lessons Learned & Next Step

Background & Motivation

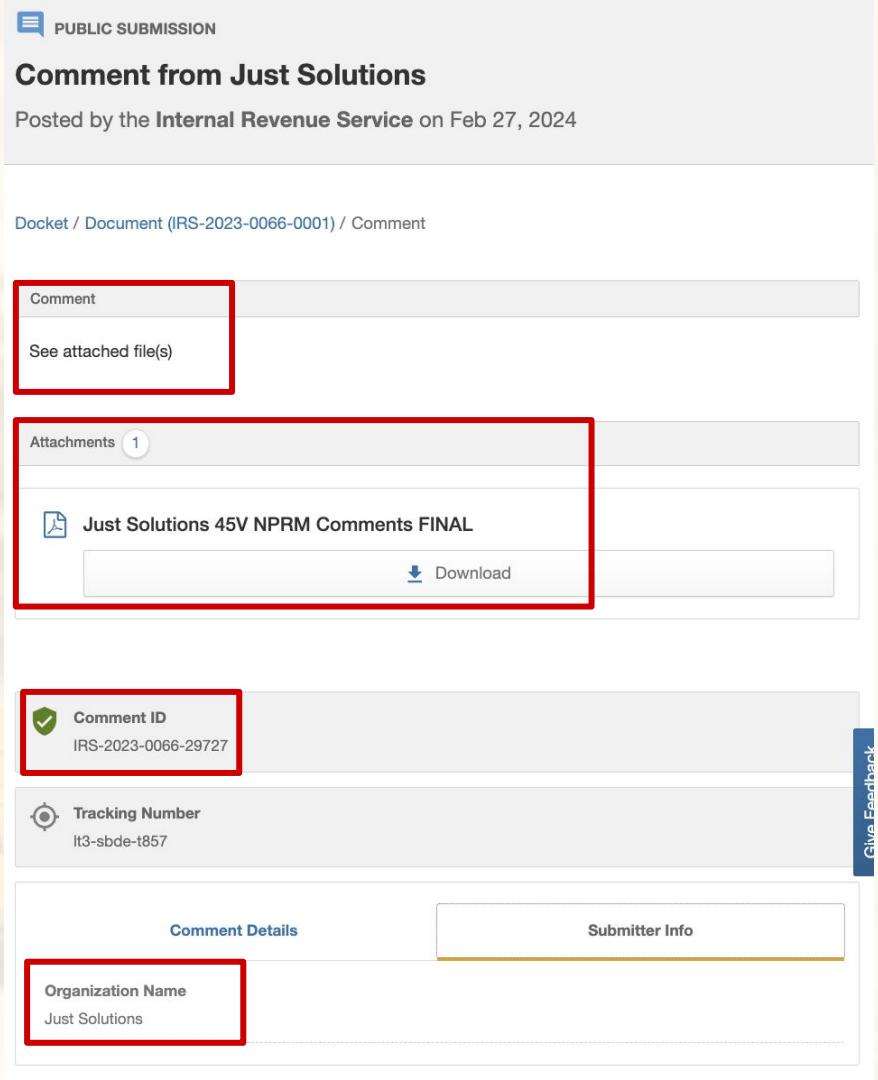
- **Policy Context: Inflation Reduction Act (IRA)**
- **Object of study:** Docket Comments
  - **Why comments?**
    - Intensive ideology expression
    - with extensive commenter detail
    - Thus **great sample for text analysis**

Sub-set of analysis

1. **Energy & Environment**
   - Tax incentive + Prevailing Wage & Registered Apprenticeship
   - Cross-docket
     - ~5,000 from 17 docket
2. **Healthcare & Social Programs**
   - Physician fee schedule & Medicare/Medicaid payment policies
   - Within-docket
     - ~5,000 from 1 docket

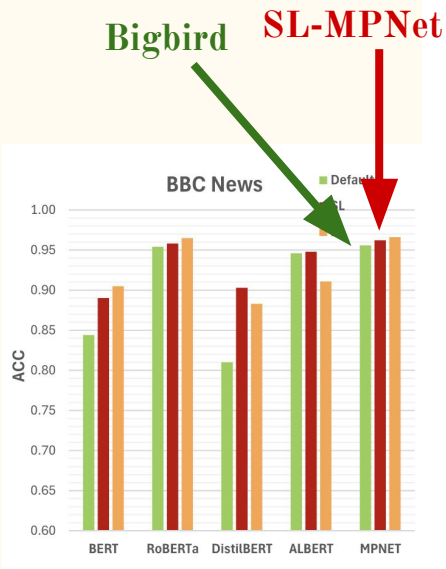
Data Collection

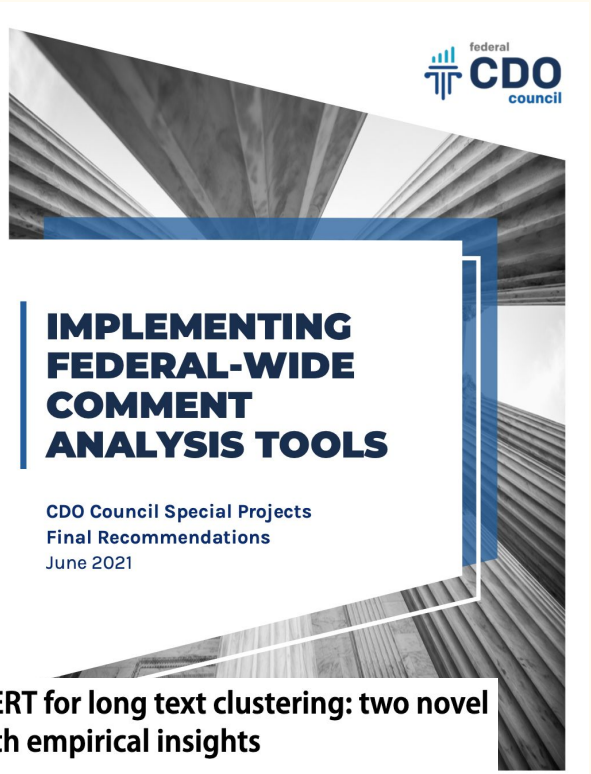
- Regulations.gov
- API
- **Basic info:**
  - Docket Id, agency Id, comment Id
- **Comment info:**
  - Organization Name, first Name, last Name, city, state Province Region
  - Comment, PDF text



Methodology: Embedding & Clustering

- **Embedding**
  - SBERT: MPNet
  - Bigbird: designed for long text processing, but constrained by 4096-token limit
  - Sentence-Level + MPNet (all-mpnet-base-v2)
- **Cosine similarity**
- **DBSCAN clustering**
  - High-density (repetitive) vs. low-density regions (unique)
  - detect “mass campaign”





Optimizing SBERT for long text clustering: two novel approaches with empirical insights

Yasin Ortakci[1] · Burak Borhan[1]

Accepted: 6 May 2025
© The Author(s) 2025

Methodology: Topic Model

Our analysis utilized Latent Dirichlet Allocation (LDA) for topic modeling, running jointly for unique and repetitive comments

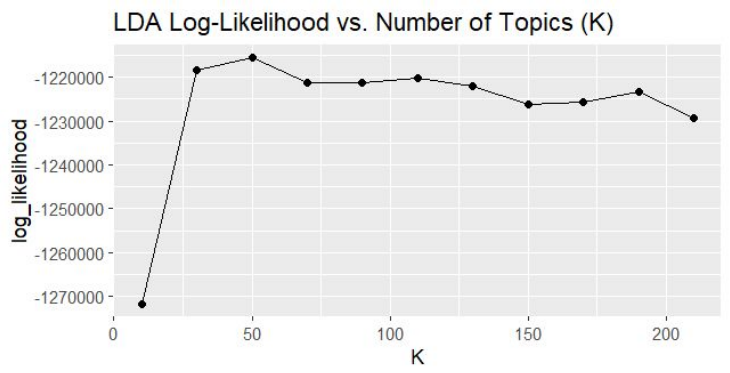
Data Curation for LDA

From repetitive clusters, one representative comment was selected; all unique comments were retained, forming the dataset for LDA application.

Medical comments: n = 516
Energy comments: n = 3,076

Optimal K-Value Determination

Log-likelihood analysis identified K=30 as a robust number of topics for both healthcare and energy comment sets across datasets.



Goals

1. Direct comparison of topic structures between groups.

2. Analyzed organizational information to understand stakeholder contributions to repetitive versus unique content

Methodology: Wordfish

The Wordfish model was employed to explore ideological tendencies within comments

Medical Dockets

For medical dockets, a two-step Wordfish process was used: an initial run to identify extreme documents as anchors, followed by a second run for final scaled positions.

Energy Dockets

For energy dockets, Wordfish was run on organization-specific comments, aligning with "Bootlegger-and-Baptist" style advocacy positions.

Data Curation for Wordfish

Combining unique and representative repetitive comments

Wordfish analysis included
n = 516 *medical comments*

n = 41 & 57 *energy comments*
of specific organizations across
2 dockets

Results

Results: LDA

For comments from the **medical dockets**, while both unique and repetitive comments address the CMS-1807-P docket, our LDA results reveal distinct topical emphases and stakeholder origins.

Repetitive Comments (15 after de-duplicated)

- Focus on physician fee schedules and payment concerns, primarily from organized interest groups.
  Clear template-style texts detected.

Unique Comments (501 comments)

- Emphasize insurance clause changes, patient impact, treatment options, and overall costs. Submitted by medical research companies, laboratory groups, and clinician-researchers.

For comments from the **energy/env dockets**, LDA result successfully captures (i) the underlying direction of ruling (ii) people's points of interest.

Docket ID:	IRS-2023-0066	Topic Numbers (?/30):	8

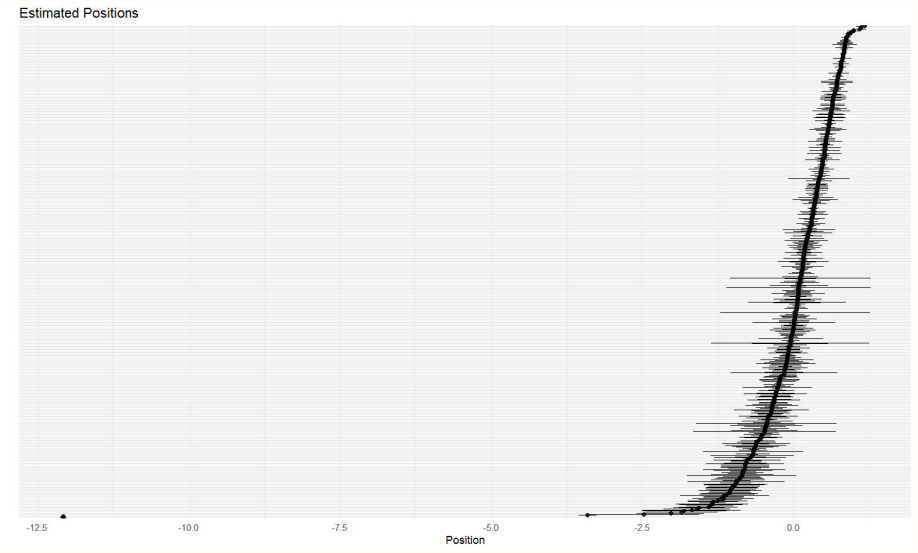
Topics:

[1. Different power sources for hydrogen production:]
hydropower, nuclear, gas, facility, epa, waste, ng, ons, sec, power
[2. 45V clean hydrogen + LCA/GREET model + natural gas route:]
hydrogen, production, clean, 45v, emissions, gas, **greet**, carbon, **model**, natural
[3. semi-nonsense:]
comments, energy, revenue, internal, irs, notice, service, washington, request, credit
[4. Matching clean electricity with clean hydrogen:]
hydrogen, clean, electricity, production, energy, power, grid, renewable, matching, 45v
[5. Macro/Industry -- IRA serves as a clean energy industry policy/tax policy tool:]
energy, clean, tax, industry, projects, u.s, treasury, new, support, development
[6. LCA GHG Quant + Methane Leakage + Data Disclosure:]
emissions, carbon, gas, tax, ghg, greenhouse, data, credits, methane, lifecycle
[7. extrem nonsense:]
e, t, o, s, n, r, c, d, l, p
[8. Biomass/Forestry + Bioenergy + Carbon Accounting:]
biomass, forest, carbon, emissions, **bioenergy, wood, et, al**, energy, production

Results: Wordfish for medical comments

Despite obvious anchors we found in the comments set, we did not generate clear position scaling using the Wordfish model

- Left-leaning anchor: Comment advocating for canceling copays to protect patients
- Right-leaning anchor: Comment opposing government proposal



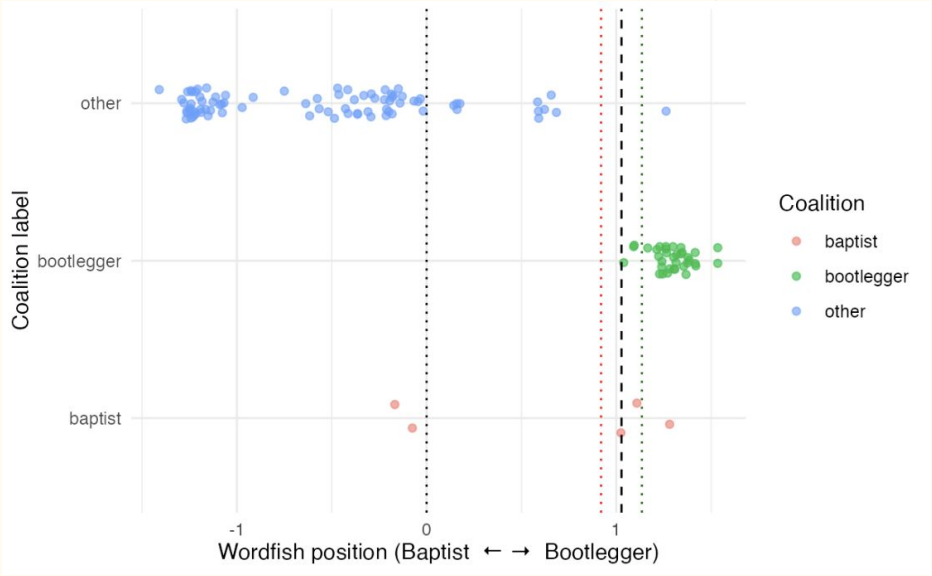
`win win, remove copay. having a copay on care management programs is hurting patients…`

`i am writing to oppose the current work values and practice expenses assigned to…`

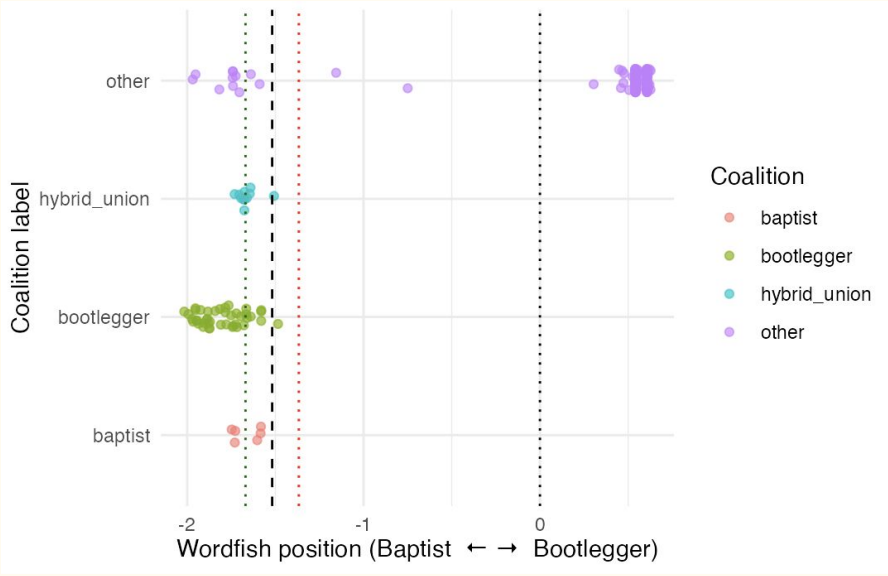
Results: Wordfish for energy comments

Pre-label → WF → Anchor: [minimum θ Baptist, maximum θ Bootlegger] → Re-run WF → Map

IRS-2023-0066: 45V Clean Hydrogen



IRS-2023-0042: Tax Credit + PWA



- **Useful: when there is real ideological conflict and variation**
- **Not useful: more harmonized comments**

Lessons Learned & Next Steps

Lessons Learned

Wordfish Application

Wordfish requires significant document variation for meaningful results. Using comments from a single docket or comments that are homogenous limited ideological separation.

Crucial Role of Preprocessing

- Irregularities (e.g., fullwidth letters, lumped words) can compromise results for LDA and Wordfish.
- Specific preprocessing, including removing rare words and stemming, was essential for Wordfish convergence and credible outputs.
- LDA naturally picks up boilerplate terms (e.g., energy, revenue, internal, IRS, notice, service, Washington, request, credit). By removing these generic words, LDA can better capture the substantive topics of the comments.

Next Steps & Enhancements

- Apply Wordfish across multi-year comment data for broader variation and potential ideological patterns.
- Further test whether the docket level document condensing diminish or reserve the representativeness/distinctiveness of each docket.
- Explore cross-referencing between energy and medical datasets to uncover deeper insights.

Thank you!