# herzog_2021_transfer_topic_labeling_with_domain_specific_knowledge_base_an_analysis_of_uk_house_of_commons_speeches_1935_2014

## Year

2021

## Author(s)

Hannah Bechara and Alexander Herzog and Slava Jankin and Peter John

## Title

Transfer Topic Labeling with Domain-Specific Knowledge Base: An Analysis of UK House of Commons Speeches 1935–2014

## Venue

Research & Politics

---

## Topic labeling

Fully automated

## Focus

Primary

## Type of contribution

Novel

## Underlying technique

Transfer topic labeling

## Topic labeling parameters

Considered CAP topic words: 150

Considered DTM topic words: 20

## Label generation

### Introduction

Using the coding instructions of the Comparative Agendas Project (CAP) to label topics.

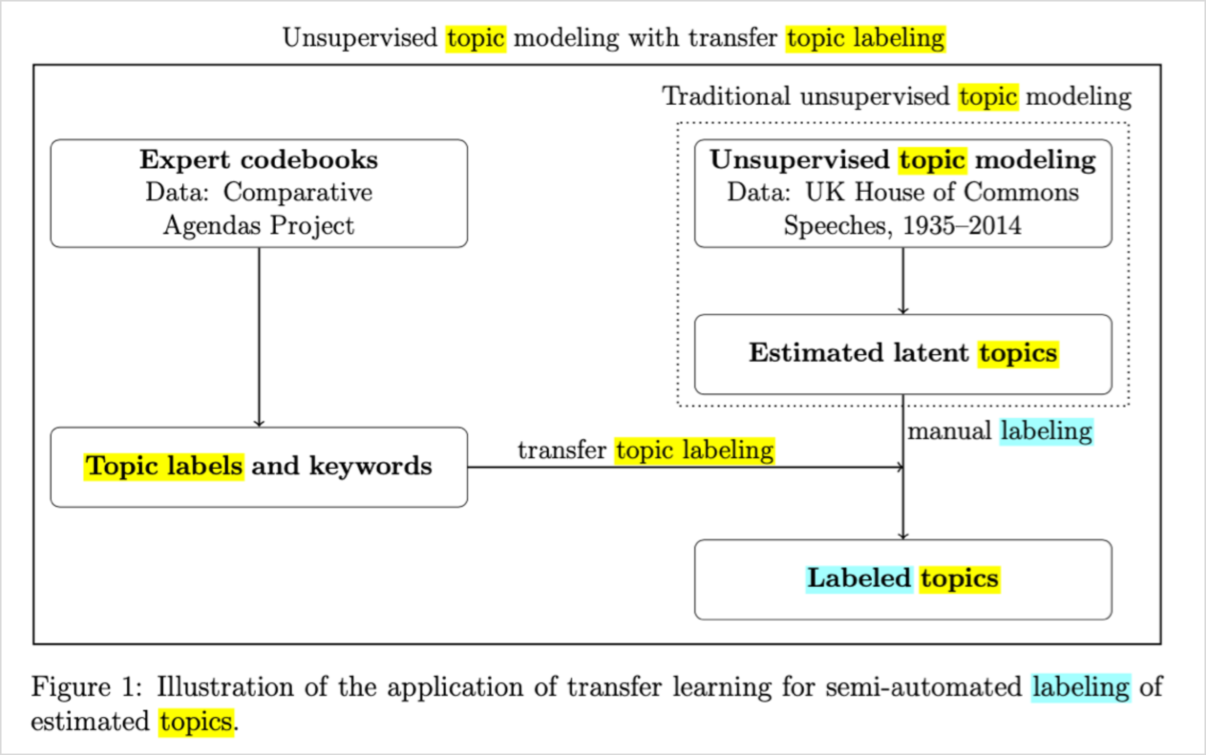

Unsupervised topic modeling with transfer topic labeling
Traditional unsupervised topic modeling
Expert codebooks
Data: Comparative Agendas Project
Unsupervised topic modeling
Data: UK House of Commons Speeches, 1935–2014
Estimated latent topics
manual labeling
Topic labels and keywords
transfer topic labeling
Labeled topics

Figure 1: Illustration of the application of transfer learning for semi-automated labeling of estimated topics.

Our main idea is illustrated in Figure 1. The dotted box on the right-hand side illustrates traditional unsupervised topic modeling, which stops with estimated latent topics that need manual labeling.

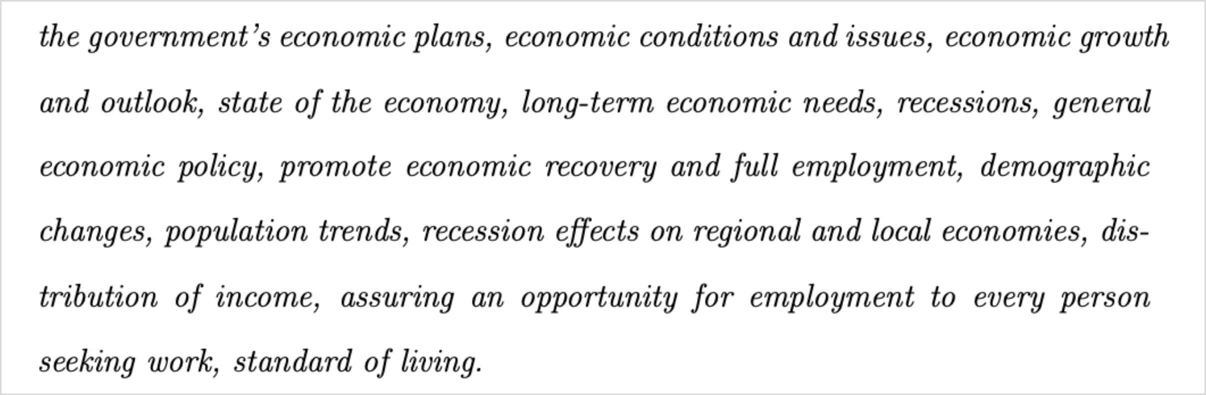
In our approach, we use outside expert codebooks to extract topic labels and associated keywords, which we then use to automatically label the estimated latent topics.
Retaining human-in-the-loop allows for adjustment of the labels for specific domains with sparse coverage in the source knowledge base.
Hence, we use the term semi-automated topic label in this paper.

### Extracting Topic Labels and Keywords from Expert Code- books

Comparative Agendas Project (CAP): What has been called the policy agendas code frame is a fuller articulation of the policy topics ideas with a larger number of major topic codes, which aims at comprehensive coverage of any topic that is likely to appear.

The codebook for the UK Policy Agendas Project includes 19 major topics with subtopics. For each subtopic, the CAP codebook provides written examples of what is being included in each category.

For example, category “1. Macroeconomics – 100: General domestic macroeconomic issues” is described as follows:



*the government’s economic plans, economic conditions and issues, economic growth and outlook, state of the economy, long-term economic needs, recessions, general economic policy, promote economic recovery and full employment, demographic changes, population trends, recession effects on regional and local economies, distribution of income, assuring an opportunity for employment to every person seeking work, standard of living.*

Because the descriptions of CAP subtopics are relatively short, we combine all subtopics under a major topic label into a single document.

We then apply tf-idf weighting to generate 19 weighted word lists (one for each major topic label), where the weight on each word reflects

its importance to a topic label

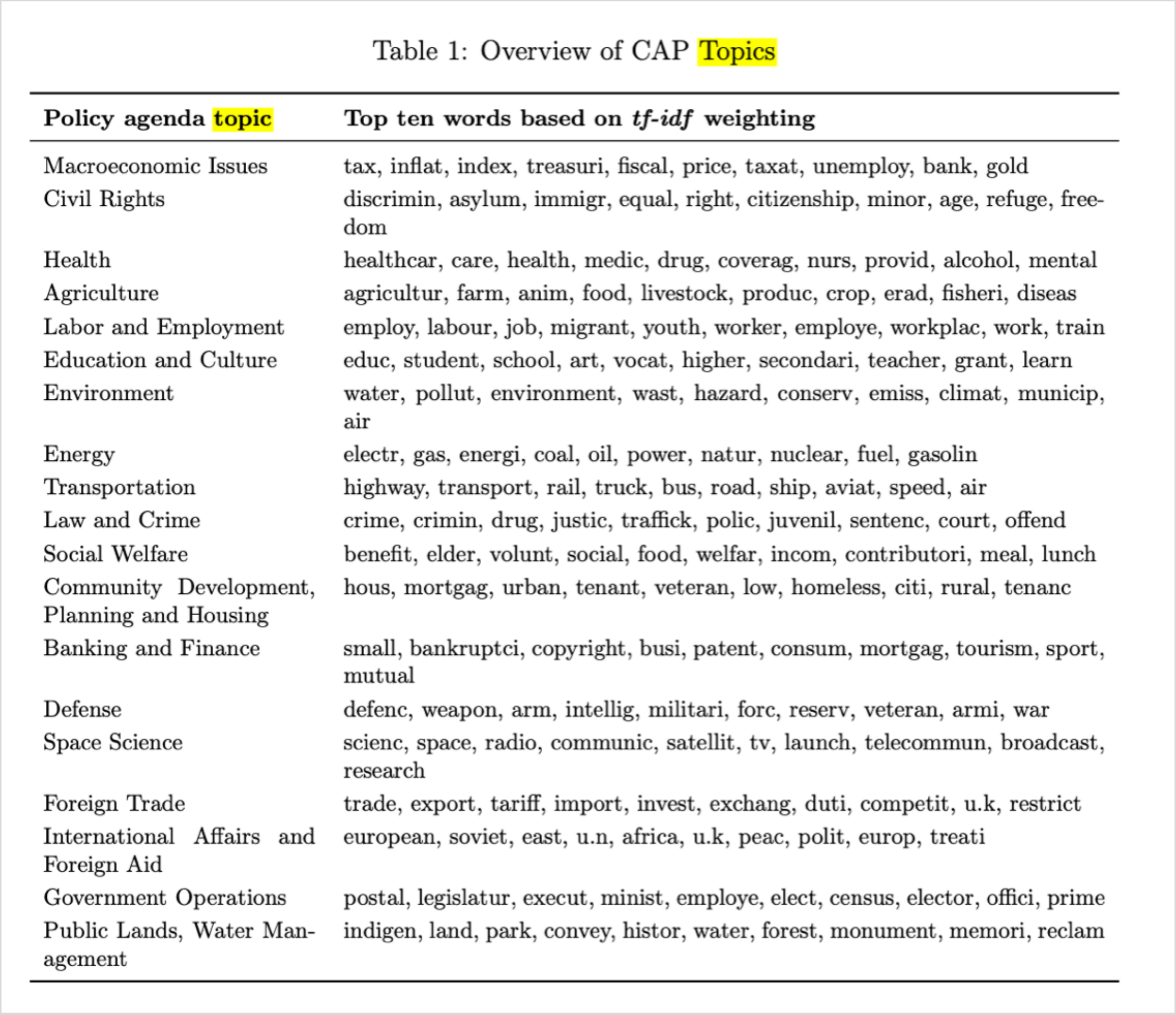


Table 1: Overview of CAP Topics

| **Policy agenda topic** | **Top ten words based on *tf-idf* weighting** |
|---|---|
| Macroeconomic Issues | tax, inflat, index, treasuri, fiscal, price, taxat, unemploy, bank, gold |
| Civil Rights | discrimin, asylum, immigr, equal, right, citizenship, minor, age, refuge, freedom |
| Health | healthcar, care, health, medic, drug, coverag, nurs, provid, alcohol, mental |
| Agriculture | agricultur, farm, anim, food, livestock, produc, crop, erad, fisheri, diseas |
| Labor and Employment | employ, labour, job, migrant, youth, worker, employe, workplac, work, train |
| Education and Culture | educ, student, school, art, vocat, higher, secondari, teacher, grant, learn |
| Environment | water, pollut, environment, wast, hazard, conserv, emiss, climat, municip, air |
| Energy | electr, gas, energi, coal, oil, power, natur, nuclear, fuel, gasolin |
| Transportation | highway, transport, rail, truck, bus, road, ship, aviat, speed, air |
| Law and Crime | crime, crimin, drug, justic, traffick, polic, juvenil, sentenc, court, offend |
| Social Welfare | benefit, elder, volunt, social, food, welfar, incom, contributori, meal, lunch |
| Community Development, Planning and Housing | hous, mortgag, urban, tenant, veteran, low, homeless, citi, rural, tenanc |
| Banking and Finance | small, bankruptci, copyright, busi, patent, consum, mortgag, tourism, sport, mutual |
| Defense | defenc, weapon, arm, intellig, militari, forc, reserv, veteran, armi, war |
| Space Science | scienc, space, radio, communic, satellit, tv, launch, telecommun, broadcast, research |
| Foreign Trade | trade, export, tariff, import, invest, exchang, duti, competit, u.k, restrict |
| International Affairs and Foreign Aid | european, soviet, east, u.n, africa, u.k, peac, polit, europ, treati |
| Government Operations | postal, legislatur, execut, minist, employe, elect, census, elector, offici, prime |
| Public Lands, Water Management | indigen, land, park, convey, histor, water, forest, monument, memori, reclam |

**Transfer Topic Labeling**

We transfer topic labels from the CAP to the estimated latent topics through a pair-wise matching procedure that finds the most similar CAP topic word list for each latent dimension.

For the CAP topics, the word lists are the tf-idf -weighted word lists discussed above.

For the dynamic topic model, we construct one word list for each of the 22 estimated latent topics.

To identify the best matching topics, we use the Jaccard index, which is a widely used set-based similarity measure.

For two sets A and B, the Jaccard index is defined as:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

where the numerator is the size of the intersection between A and B, and the denominator is the size of the union of the two sets.

The Jaccard index is bound between 0 and 1, with higher numbers indicating greater overlap between two sets.

We calculate the Jaccard index for each pair of word lists consisting of one CAP topic and one estimated DTM topic.

Using the highest Jaccard value results in 19 unique matches where the CAP label is transferred to the estimated topic.

Table 2 provides an overview of the 22 estimated dynamic topics together with their Jaccard index, the matched CAP topic label, topic label selected by experts, proportion of experts who selected the same topic label as the transfer-learning approach with the first or second choice, Fleiss’ kappa inter-coder agreement, and the top 20 words from each DTM topic.

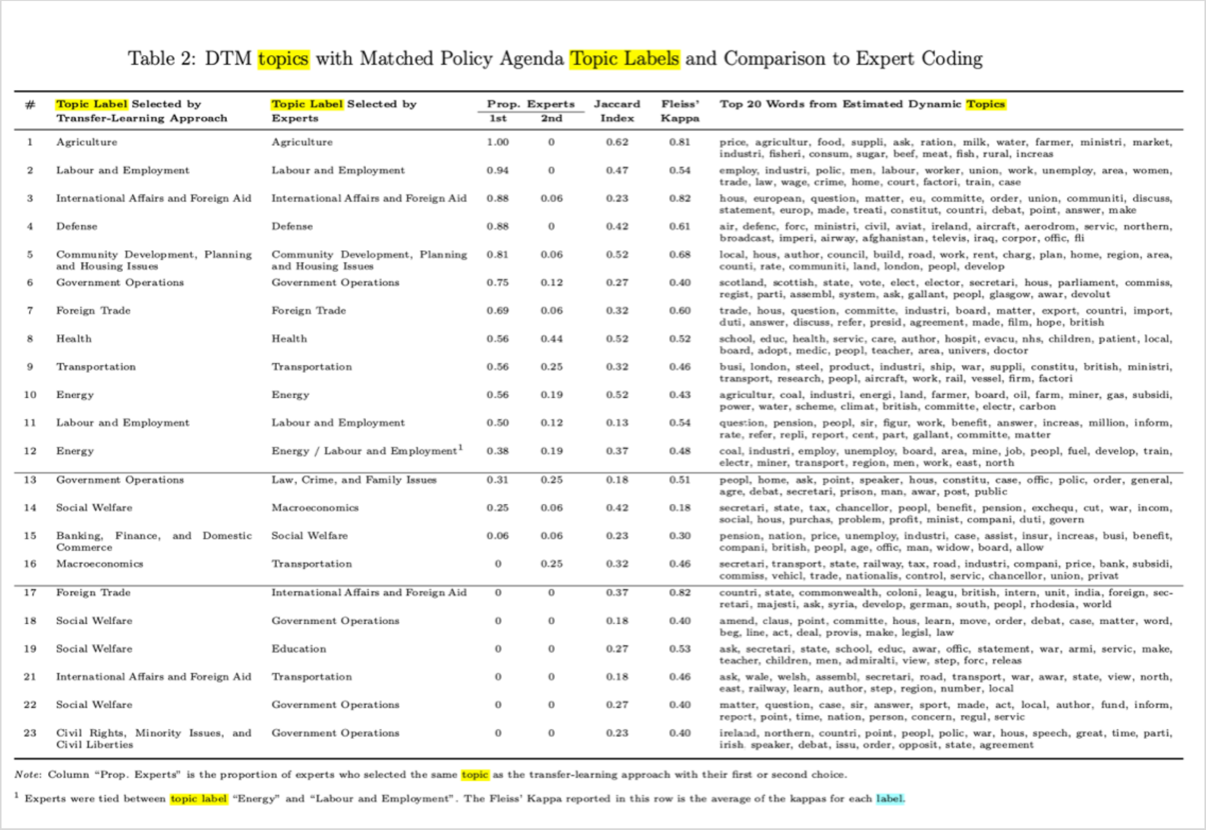


Table 2: DTM topics with Matched Policy Agenda Topic Labels and Comparison to Expert Coding

| # | Topic Label Selected by Transfer-Learning Approach | Topic Label Selected by Experts | Prop. Experts | | Jaccard Index | Fleiss' Kappa | Top 20 Words from Estimated Dynamic Topics |
|---|---|---|---|---|---|---|---|
| | | | 1st | 2nd | | | |
| 1 | Agriculture | Agriculture | 1.00 | 0 | 0.62 | 0.81 | price, agricultur, food, suppli, ask, ration, milk, water, farmer, ministri, market, industri, fisheri, consum, sugar, beef, meat, fish, rural, increas |
| 2 | Labour and Employment | Labour and Employment | 0.94 | 0 | 0.47 | 0.54 | employ, industri, polic, men, labour, worker, union, work, unemploy, area, women, trade, law, wage, crime, home, court, factori, train, case |
| 3 | International Affairs and Foreign Aid | International Affairs and Foreign Aid | 0.88 | 0.06 | 0.23 | 0.82 | hous, european, question, matter, eu, committe, order, union, communiti, discuss, statement, europ, made, treati, constitut, countri, debat, point, answer, make |
| 4 | Defense | Defense | 0.88 | 0 | 0.42 | 0.61 | air, defenc, forc, ministri, civil, aviat, ireland, aircraft, aerodrom, servic, northern, broadcast, imperi, airway, afghanistan, televis, iraq, corpor, offic, fli |
| 5 | Community Development, Planning and Housing Issues | Community Development, Planning and Housing Issues | 0.81 | 0.06 | 0.52 | 0.68 | local, hous, author, council, build, road, work, rent, charg, plan, home, region, area, counti, rate, communiti, land, london, peopl, develop |
| 6 | Government Operations | Government Operations | 0.75 | 0.12 | 0.27 | 0.40 | scotland, scottish, state, vote, elect, elector, secretari, hous, parliament, commiss, regist, parti, assembl, system, ask, gallant, peopl, glasgow, awar, devolut |
| 7 | Foreign Trade | Foreign Trade | 0.69 | 0.06 | 0.32 | 0.60 | trade, hous, question, committe, industri, board, matter, export, countri, import, duti, answer, discuss, refer, presid, agreement, made, film, hope, british |
| 8 | Health | Health | 0.56 | 0.44 | 0.52 | 0.52 | school, educ, health, servic, care, author, hospit, evacu, nhs, children, patient, local, board, adopt, medic, peopl, teacher, area, univers, doctor |
| 9 | Transportation | Transportation | 0.56 | 0.25 | 0.32 | 0.46 | busi, london, steel, product, industri, ship, war, suppli, constitu, british, ministri, transport, research, peopl, aircraft, work, rail, vessel, firm, factori |
| 10 | Energy | Energy | 0.56 | 0.19 | 0.52 | 0.43 | agricultur, coal, industri, energi, land, farmer, board, oil, farm, miner, gas, subsidi, power, water, scheme, climat, british, committe, electr, carbon |
| 11 | Labour and Employment | Labour and Employment | 0.50 | 0.12 | 0.13 | 0.54 | question, pension, peopl, sir, figur, work, benefit, answer, increas, million, inform, rate, refer, repli, report, cent, part, gallant, committe, matter |
| 12 | Energy | Energy / Labour and Employment[1] | 0.38 | 0.19 | 0.37 | 0.48 | coal, industri, employ, unemploy, board, area, mine, job, peopl, fuel, develop, train, electr, miner, transport, region, men, work, east, north |
| 13 | Government Operations | Law, Crime, and Family Issues | 0.31 | 0.25 | 0.18 | 0.51 | peopl, home, ask, point, speaker, hous, constitu, case, offic, polic, order, general, agre, debat, secretari, prison, man, awar, post, public |
| 14 | Social Welfare | Macroeconomics | 0.25 | 0.06 | 0.42 | 0.18 | secretari, state, tax, chancellor, peopl, benefit, pension, exchequ, cut, war, incom, social, hous, purchas, problem, profit, minist, compani, duti, govern |
| 15 | Banking, Finance, and Domestic Commerce | Social Welfare | 0.06 | 0.06 | 0.23 | 0.30 | pension, nation, price, unemploy, industri, case, assist, insur, increas, busi, benefit, compani, british, peopl, age, offic, man, widow, board, allow |
| 16 | Macroeconomics | Transportation | 0 | 0.25 | 0.32 | 0.46 | secretari, transport, state, railway, tax, road, industri, compani, price, bank, subsidi, commiss, vehicl, trade, nationalis, control, servic, chancellor, union, privat |
| 17 | Foreign Trade | International Affairs and Foreign Aid | 0 | 0 | 0.37 | 0.82 | countri, state, commonwealth, coloni, leagu, british, intern, unit, india, foreign, secretari, majesti, ask, syria, develop, german, south, peopl, rhodesia, world |
| 18 | Social Welfare | Government Operations | 0 | 0 | 0.18 | 0.40 | amend, claus, point, committe, hous, learn, move, order, debat, case, matter, word, beg, line, act, deal, provis, make, legisl, law |
| 19 | Social Welfare | Education | 0 | 0 | 0.27 | 0.53 | ask, secretari, state, school, educ, awar, offic, statement, war, armi, servic, make, teacher, children, men, admiralti, view, step, forc, releas |
| 21 | International Affairs and Foreign Aid | Transportation | 0 | 0 | 0.18 | 0.46 | ask, wale, welsh, assembl, secretari, road, transport, war, awar, state, view, north, east, railway, learn, author, step, region, number, local |
| 22 | Social Welfare | Government Operations | 0 | 0 | 0.27 | 0.40 | matter, question, case, sir, answer, sport, made, act, local, author, fund, inform, report, point, time, nation, person, concern, regul, servic |
| 23 | Civil Rights, Minority Issues, and Civil Liberties | Government Operations | 0 | 0 | 0.23 | 0.40 | ireland, northern, countri, point, peopl, polic, war, hous, speech, great, time, parti, irish, speaker, debat, issu, order, opposit, state, agreement |

*Note*: Column "Prop. Experts" is the proportion of experts who selected the same topic as the transfer-learning approach with their first or second choice.

[1] Experts were tied between topic label "Energy" and "Labour and Employment". The Fleiss' Kappa reported in this row is the average of the kappas for each label.

## Motivation

The agenda of the legislature and the content of debates shifts across topics over time, according to functional pressures and agenda-setting in the media and from public opinion (This is known as concept drift).
Dynamic topic models are often used to capture the evolution of content structure over time.

Human coders are intuitively better placed to pick up the change in meaning in political texts, while machine coding is often faulted with failures to detect semantic change.

We present a method that automatically transfers existing domain-specific knowledge base (Comparative Agendas Project) for topic labeling.
Our method applies more generally and can be easily extended to other areas with existing domain-specific knowledge base.

## Topic modeling

Dynamic topic model (DTM) Blei and Lafferty (2006)

## Topic modeling parameters

Nr of topics: 22

## Nr. of topics

22

---

## Label

One of 19 single or multi-word labels extracted from the codebook for the UK Policy Agendas Project

## Label selection

\

## Label quality evaluation

**Jaccard’s index**
Our method for matching topic labels to estimated latent topics produces goodness-of-fit measures (the Jaccard’s index) for each match, which allows to evaluate how well the topics derived from a codebook capture the estimated latent dimensions.

**Comparison with manual labeling (Fleiss’ kappa)**
As a validation exercise we recruited a group of CAP experts to label the word lists for each topic according to the CAP categorization.
Seventeen experts who participated in this exercise could submit two choices of the labels (most appropriate and second most appropriate).
Experts were ask to select a topic most appropriate to the topic word list as the best fit and the second best fit, and indicate their confidence in the selection

We assess the quality of expert labeling using Fleiss’ kappa measure of inter-coder agreement.
We also calculate proportion of experts who agree with the automatically selected topic label as their first or second choice.

A majority of experts agreed with the automatic approach on 12 topic labels .

## Assessors

Seventeen CAP experts

---

## Domain

Paper: Topic labeling
Dataset: Policy

## Problem statement

The process of manual labeling is not scalable and suffers from human bias.
We present a semi-automatic transfer topic labeling method that seeks to remedy these problems.
Domain-specific codebooks form the knowledge-base for automated topic labeling.

## Corpus

Origin: TheyWorkForYou
Nr. of documents: 4.3M (47,524 after pre-processing)
Details:

- Complete corpus of UK House of Commons speeches 1935-2014

## Document

Combined text of a given MP's contributions.

## Pre-processing

- Excluding contributions that concern the rules of procedure or the business of the House, such as the reading of the parliamentary agenda or formal announcements
- Removed the traditional prayer at the beginning of each sitting and all contributions and announcements by the Speaker.
- Stemming
- Removed words that appeared less than 50 times and in fewer than 5 documents
- Removed punctuation, numbers, symbols, stopwords, hyphens, singleletters, and a custom list of high-frequency terms

```
@article{bechara_2021transfer_learning_for_topic_labeling_analysis_of_the_uk_house_of_commons_speeches_1935_2014,
  author = {Hannah B{\'{e}}chara and Alexander Herzog and Slava Jankin and Peter John},
  date-added = {2023-04-08 18:04:21 +0200},
  date-modified = {2023-04-08 18:04:21 +0200},
  doi = {10.1177/20531680211022206},
  journal = {Research {\&}amp$\mathsemicolon$ Politics},
  month = {apr},
  number = {2},
  pages = {205316802110222},
  publisher = {{SAGE} Publications},
  title = {Transfer learning for topic labeling: Analysis of the {UK} House of Commons speeches 1935{\textendash}2014},
  url = {https://doi.org/10.1177%2F20531680211022206},
  volume = {8},
  year = 2021}
```

#Thesis/Papers/BS