Az ELKH Cloud előnyei a POLTEXT projekt példáján

Sebők Miklós (TK PTI),
Kacsuk Zoltán (HdM IAAI)

Research context

Intersection of two research projects at the Centre for Social Sciences, Institute for Political Science:

- **Text Mining of Political and Legal Texts (POLTEXT) Incubator Project** (Principal Investigator: Miklós Sebők)
- **Hungarian Comparative Agendas Project (CAP) project** (Project leaders: Zsolt Boda & Miklós Sebők)

Research problem

- **Quantitative analysis** of qualitative data
  - Classifying articles according to policy topics
  - Topics: education to defense, Comparative Agendas Project
- **Gold standard**: double-blind human coding by well-trained researchers
- What if this is **unfeasible?**
  - Article counts of over 100.000
  - **Cost and training** of human coders for this scale

An example


2012. november 12., hétfő • LXX. évfolyam • 264. szám, 20 oldal • 220 forint, előfizetőknek 141 forint • www.nol.hu

NÉPSZABADSÁG

Markó Béla szerint nem jó ötlet a külhoni magyarok szavazása
- Sok Fidesz-szimpatizáns voksol az RMDSZ-re Erdélyben.
- A választók pragmatikusak – állítja. Interjú →13

ÚJSÁGOKBÓL KÉSZÜL SZALÁMI NEW YORKBAN
Rivalda →20

Családi gazdálkodás Elégedetlenség Fejér megyében

Földfoglalás Kajászón

Hétszázezer négyzetméter az száz futballpálya FOTÓ: REVICZKY ZSOLT

Rab László

ÖTÖDIK HASÁB

Nem ász

Vélemény

Válogatott keret

250 millió egy állami cégtől a Fradinak aligha nevezhető privát adakozókedvnek – írja Tamás Ervin. →11

Érvek és ellenérdek

Marnitz István szerint az energiaár ugyan nem piaci, de nehezen fizethető. →11

Alaptalan pedofilvádak miatt távozott a BBC vezérigazgatója

George Entwistle csak nyolc hete lépett hivatalba. Azért kellett lemondania, mert a BBC2 egyik műsorában egy tory politikust tévesen vádoltak gyerekmolesztálással. →8

A kormány már lemondott róla, mégsem lesz állami mobilcég?

Lapunk piaci forrásból úgy értesült, hogy a negyedik szolgáltatót birtokló három állami cég már nem szívesen finanszírozná a több százmilliárdos beruházást. →9

EGYRE TÖBB AZ ILLEGÁLIS BELÉPŐ

Magyarország schengeni határain szeptember végéig több mint ötezer, jogellenesen belépő embert fogtak el. A múlt év azonos időszakában még csak 3324-nél tartottak. Afgánok, macedónok, marokkóiak, koszovóiak, szerbek próbálkoznak leginkább. →4

FOTÓ: MÓRICZ SIMON

Complete text:

„The government passed on the option –There will be no state-funded mobile carrier? Our business sources indicate that the three state-owned corporations with a stake in the firm would not want to invest multiple hundred billion HUFs in the fourth carrier.”

Policy code (major topic): 17, which stands for „Space, science, technology, and telecommunications”

The project

- **Hungarian country project** of the Comparative Agendas Project (cap.tk.mta.hu) – Media module – 3 daily newspapers
- For this pilot project:
  - Left-liberal **Népszabadság (NS)**: Over 50 000 front-page articles (1990-2014)
  - Centre-Right **Magyar Nemzet (MN)**: 35 021 articles (2002-2014)
- Hand-coding was unfeasible for our purposes
- Solution: **text mining + machine learning**

A machine learning solution

- **Text as Data** – qualitative data is converted to quantative (matrices)
- How to categorize articles into pre-defined classes: **Dictionary-based** or **supervised learning**
- For the latter a sufficiently **large human-coded *training/test set* is needed**

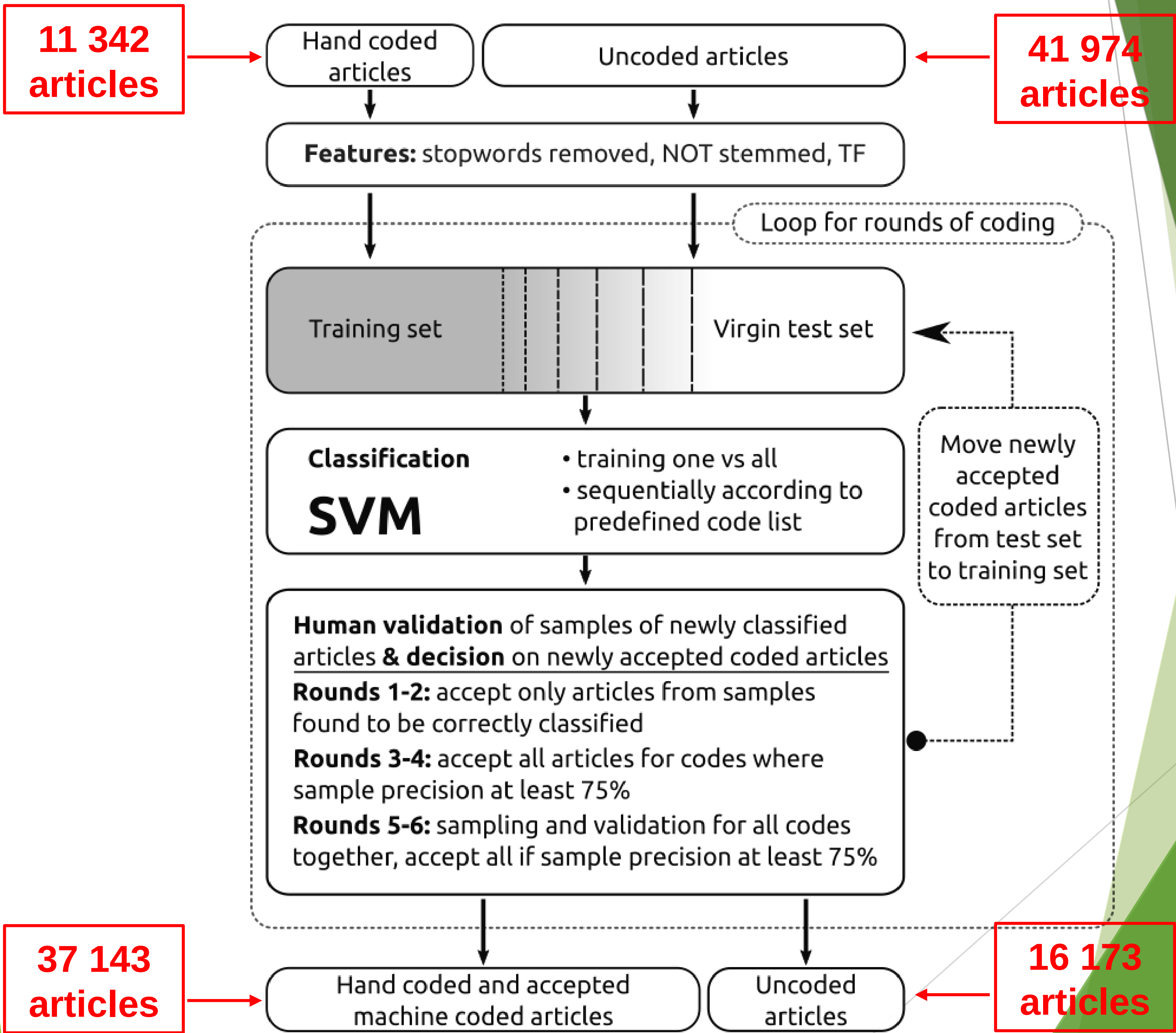
Part 1

CREATING A MACHINE CODED TRAINING SET FOR THE LEFT-WING DAILY NÉPSZABADSÁG (NS)

The Hybrid Binary Snowball (HBS) process

- We need to keep human **coding costs as low as possible**, while extracting the largest possible gain per invested human coding hour
- We simplify multi-class classification by rephrasing it as a **series of pairwise comparisons**
- We apply a snowball method to **augment the training set with machine-classified observations**

Coding NS articles



ELKH Cloud

Infrastructural bottlenecks 1: **Memory**

ELKH Cloud

- Our desktop workstation had **only 32 GB** RAM
- Encountered **problems**:
  - Could not work on the **whole virgin data** set
  - Could not run **certain configurations**, for example: Term frequency - Inverse document frequency (Tf-Idf) weighting
- Even the **solutions were problems**:
  - Virgin data was partitioned up for processing
  - This would impact Tf-Idf weighting significantly
- **Real solution** going forward:
  - Using larger capacity single virtual instances or a **cluster in the cloud**

Infrastructural bottlenecks 2: **Time**

- **Huge numbers of small operations** add up quickly
- If process runtimes become too long, project execution becomes unfeasible
- Solution: **parallelizing** the execution of operations

Part 2

USING THE CODED LEFT-WING DAILY ARTICLES TO TEACH THE ALGORITHM HOW TO CODE THE CONSERVATIVE DAILY MAGYAR NEMZET (MN)

Apache Spark cluster

- With the help of the Laboratory of Parallel and Distributed Systems at the Institute for Computer Science and Control (**SZTAKI LPDS**)
- Apache Spark cluster running on five virtual instances in the **SZTAKI ELKH Cloud**
- All five virtual instances had 8 virtual processors and 32 GBs of RAM each, and were running Ubuntu 16.04.
- Four instances acted as worker nodes and one as the master node of the Spark cluster. Each Spark session was running with 32 VCPUs (but **default parallelism set to 24**) and **96 GBs of RAM** total on the four worker nodes combined.

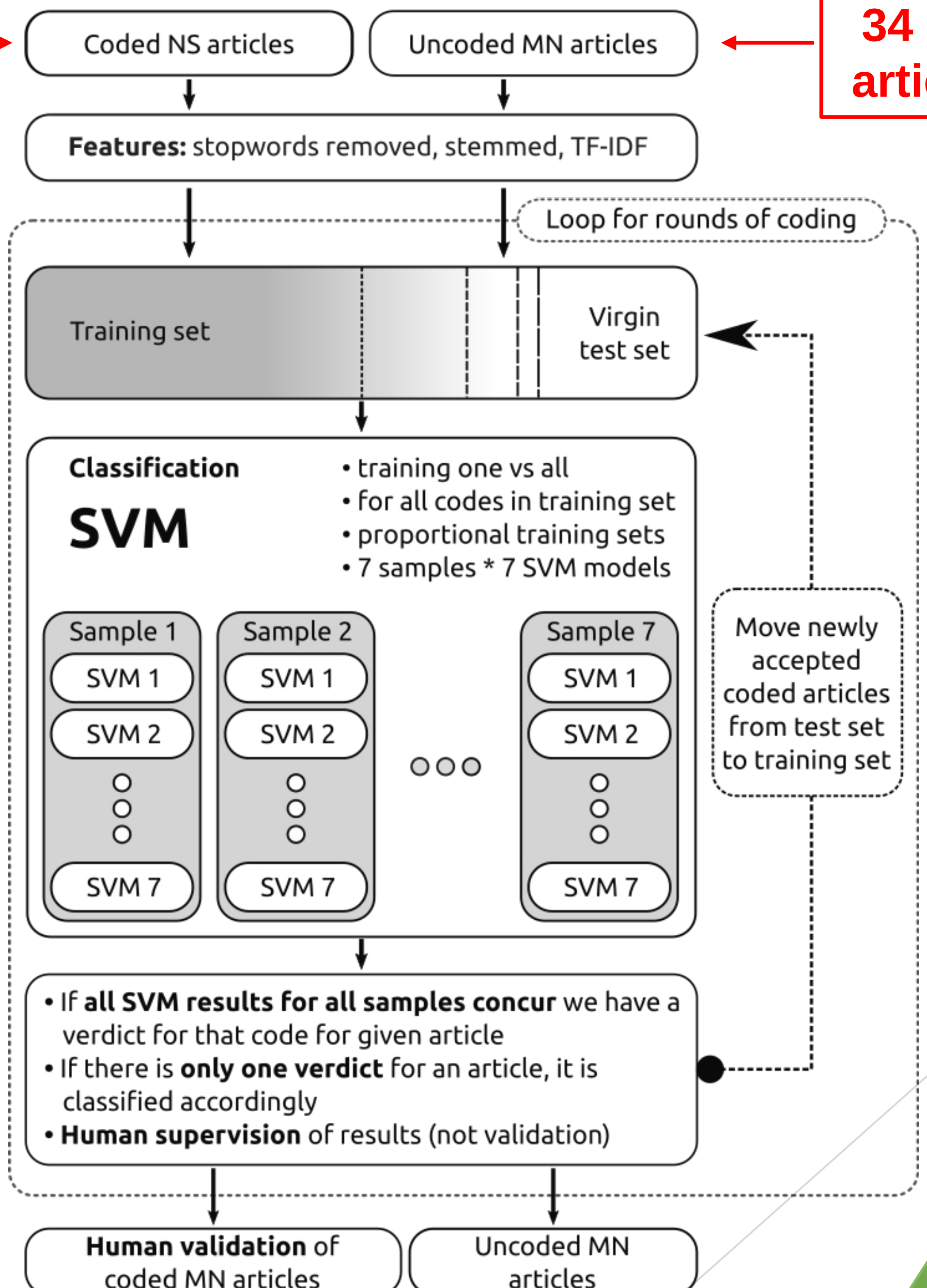
Manifold increase in speed

- **Old desktop setup**: roughly **3 days** for a full round of coding (33 code categories)
- **Spark cluster**: ca. **30 minutes** for a full round of coding
- This increase in speed enabled:
  - **1) Rapid prototyping**
  - **2) Complex classification workflow**

Coding MN articles

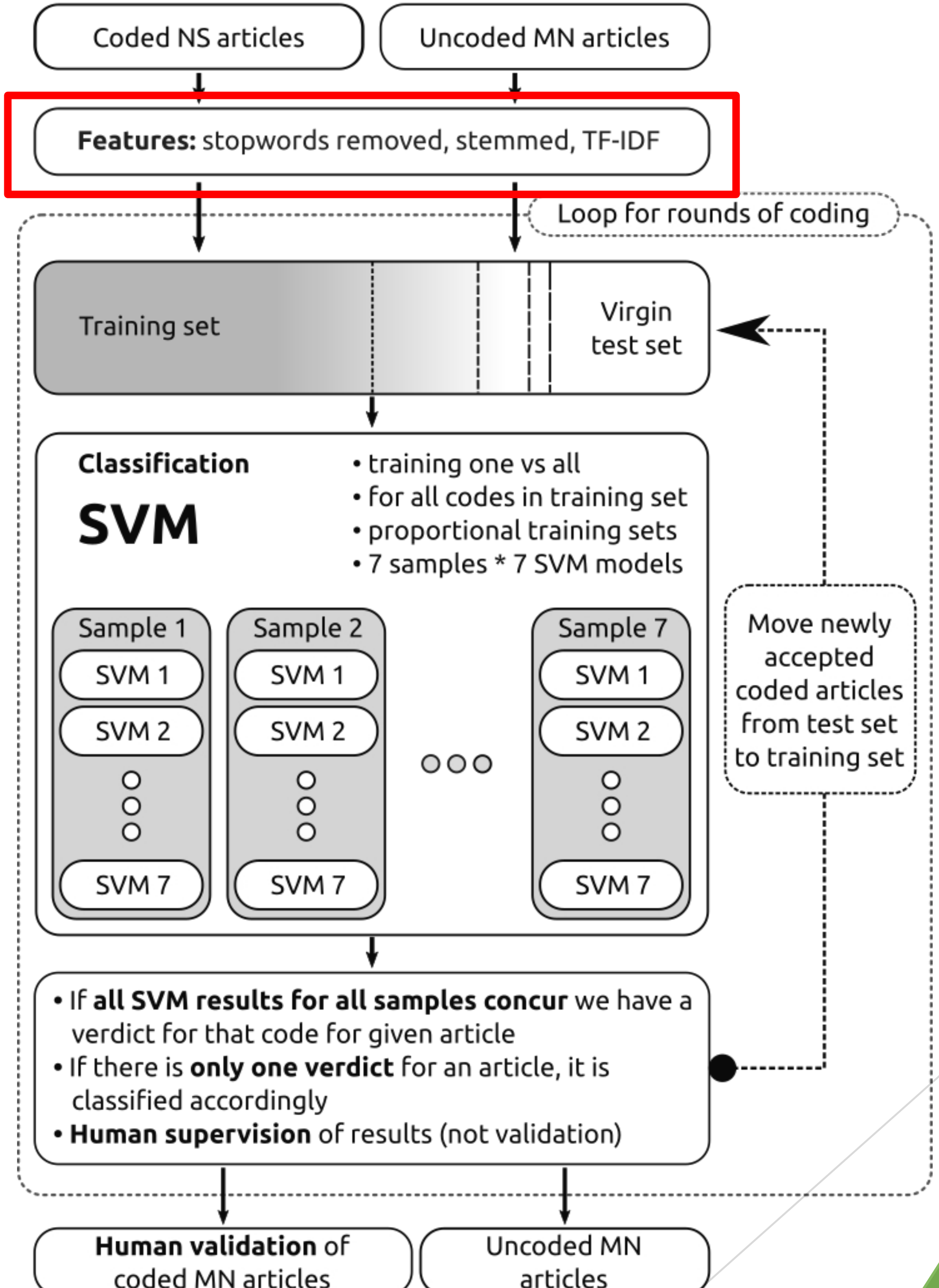
! 34 650 articles

34 670 articles



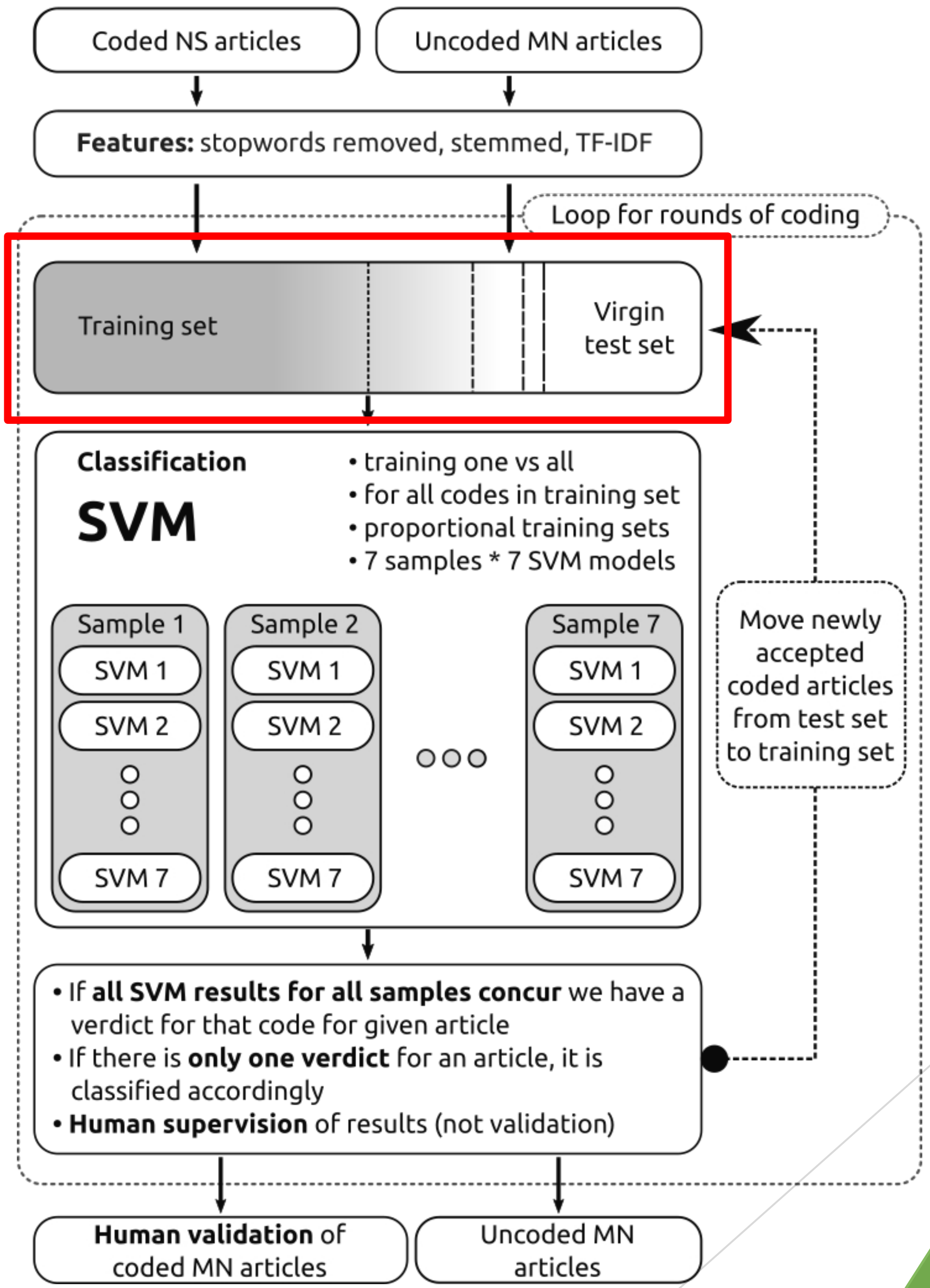
ELKH Cloud

Coding MN articles

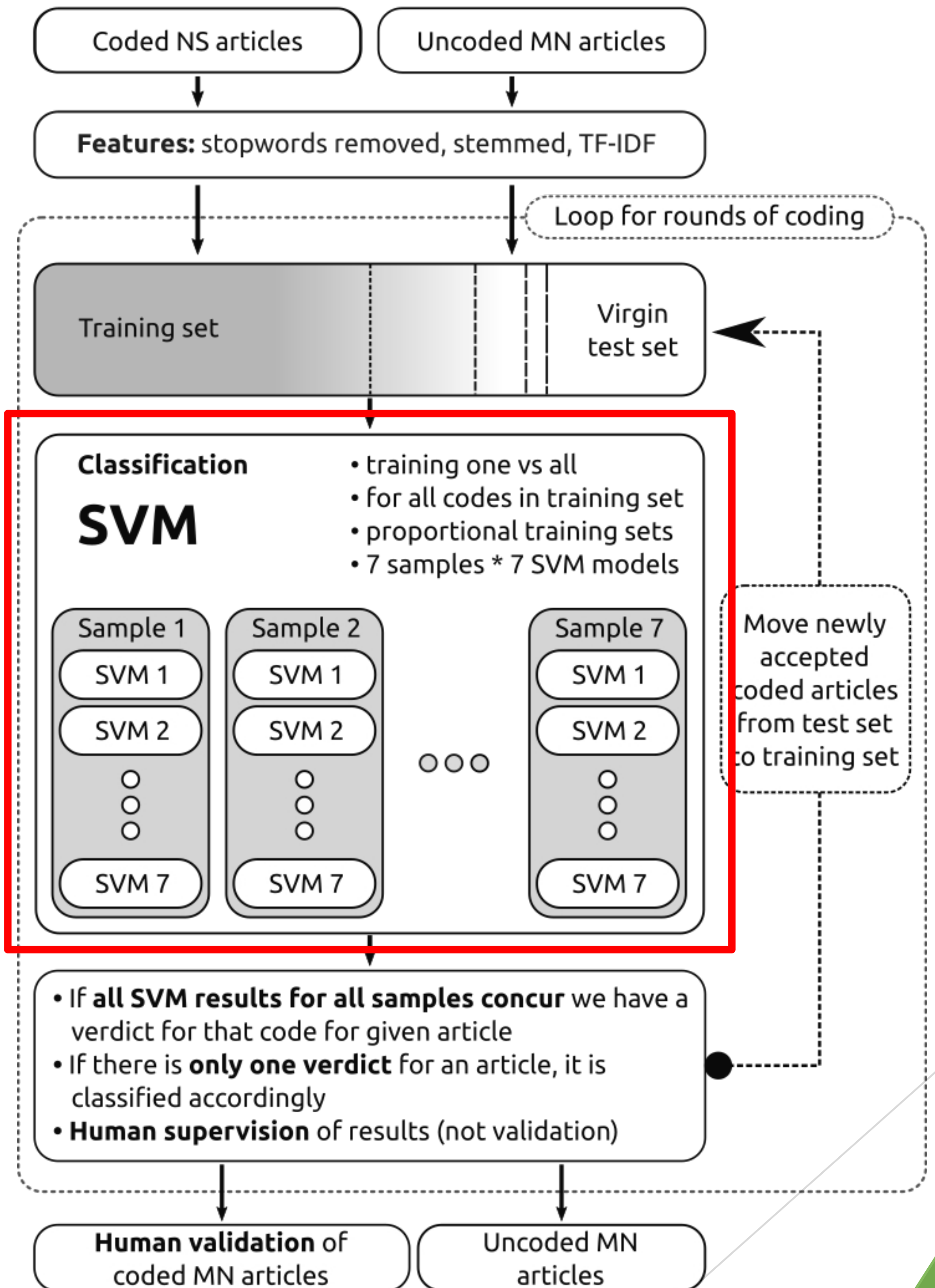


ELKH Cloud

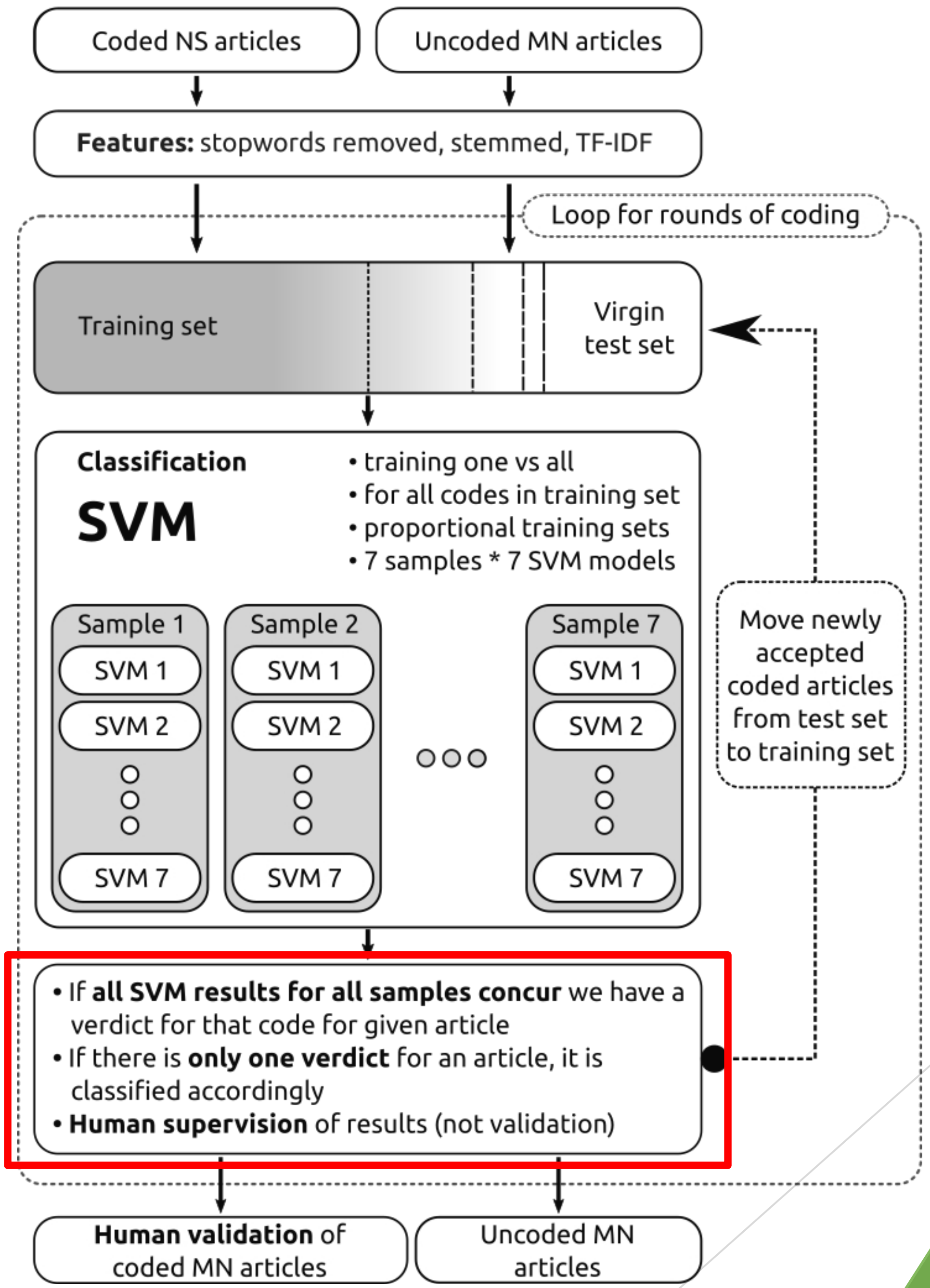
Coding MN articles



Coding MN articles

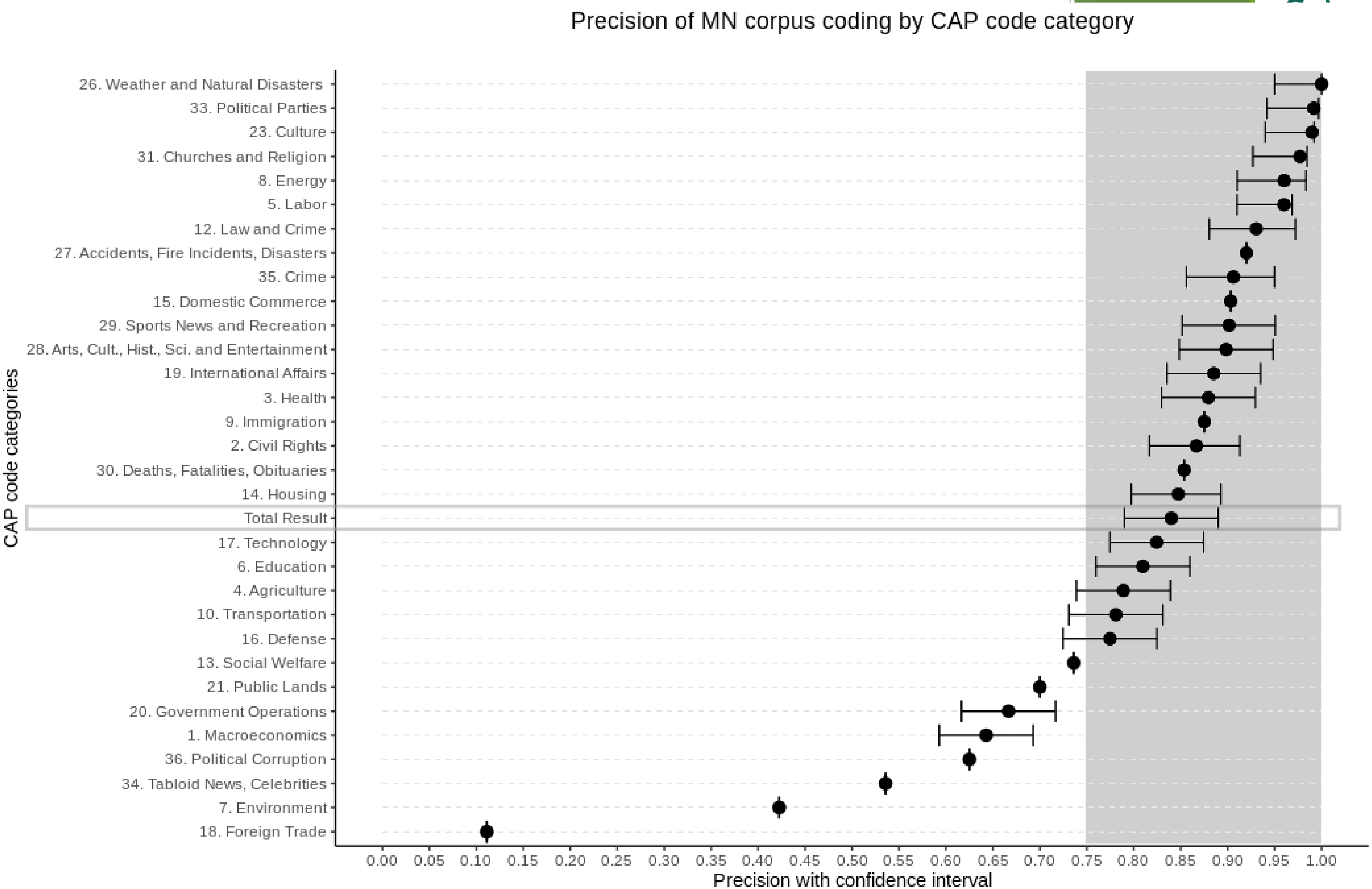


Coding MN articles

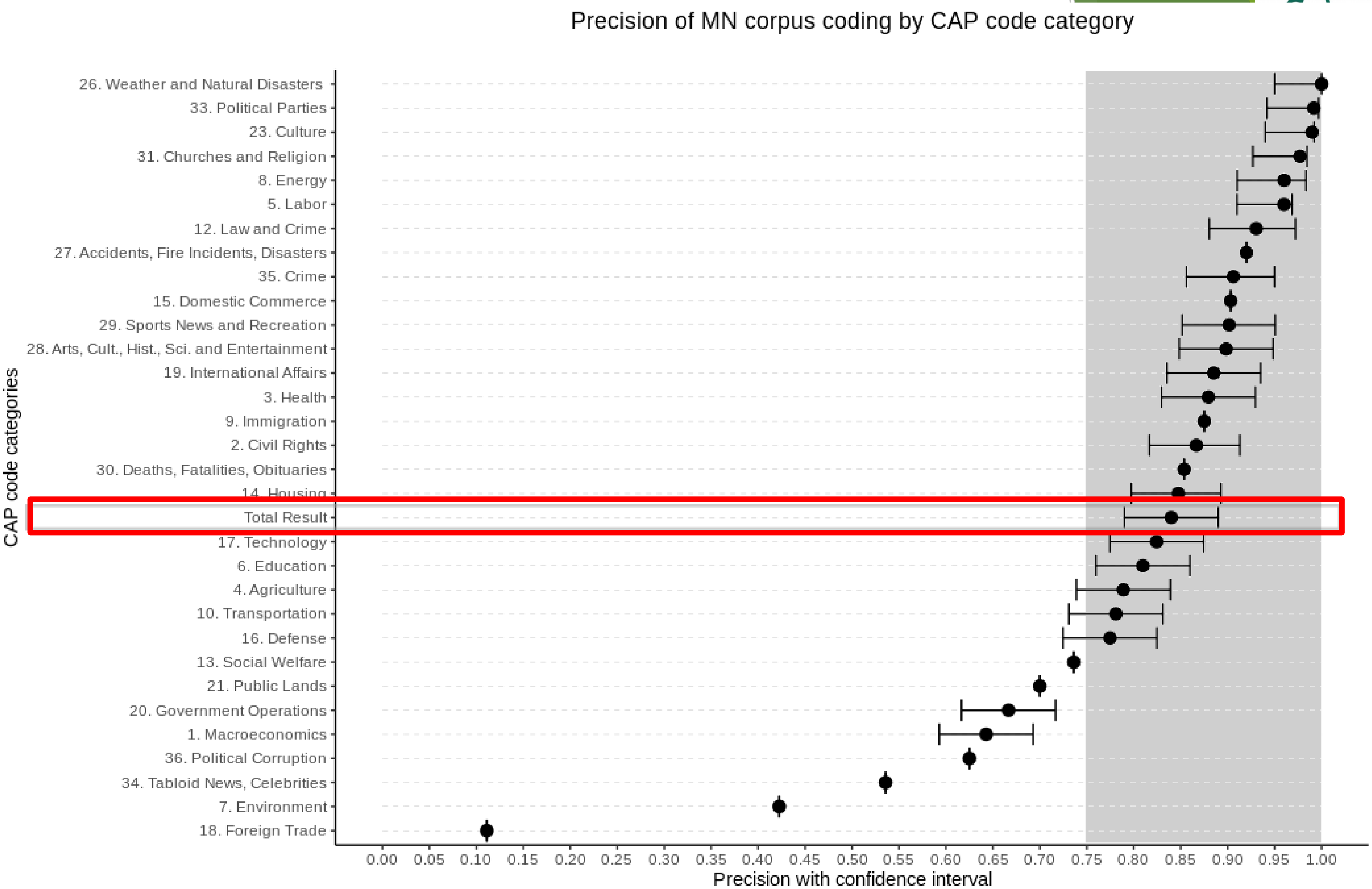


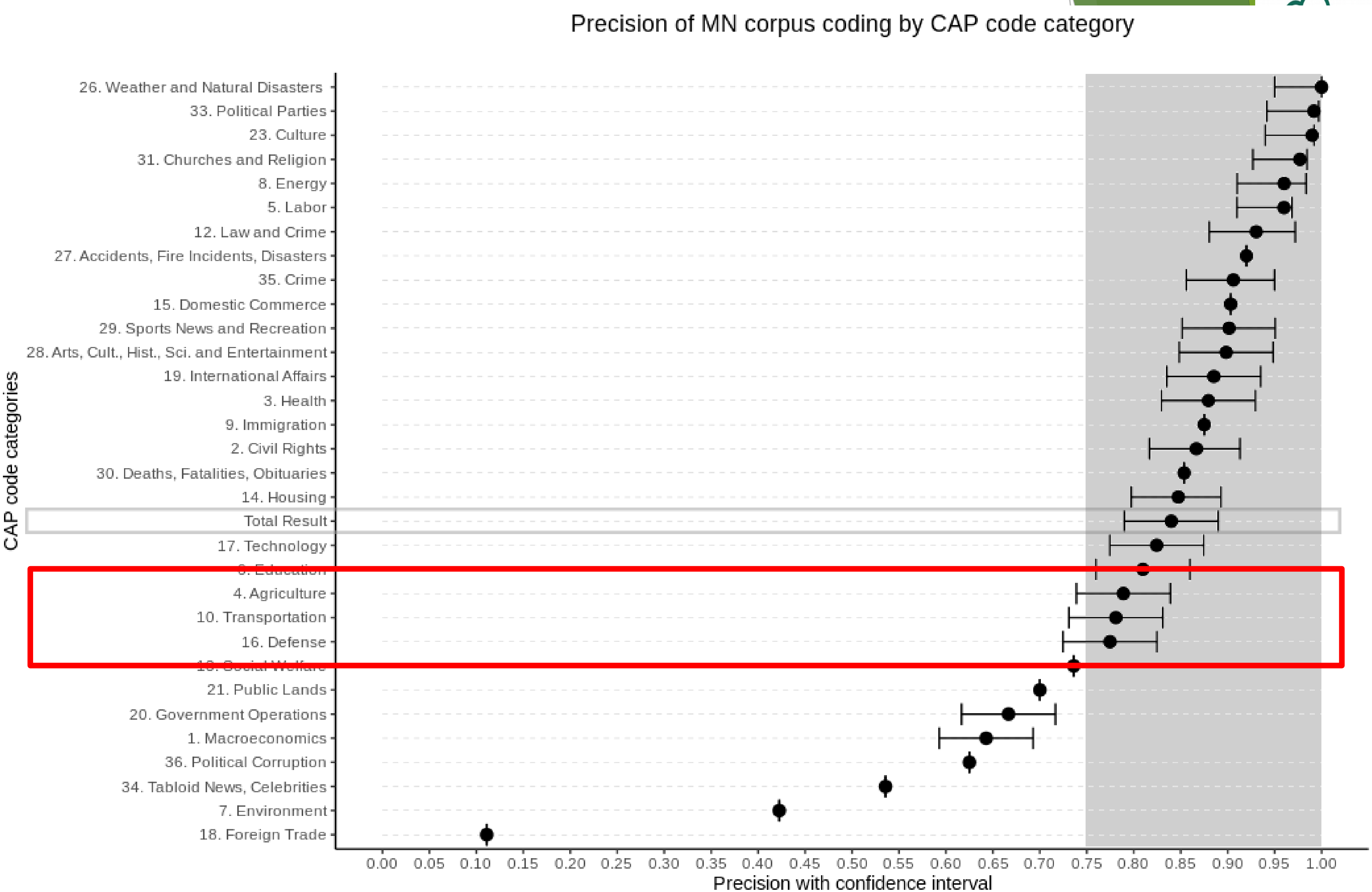
ELKH Cloud

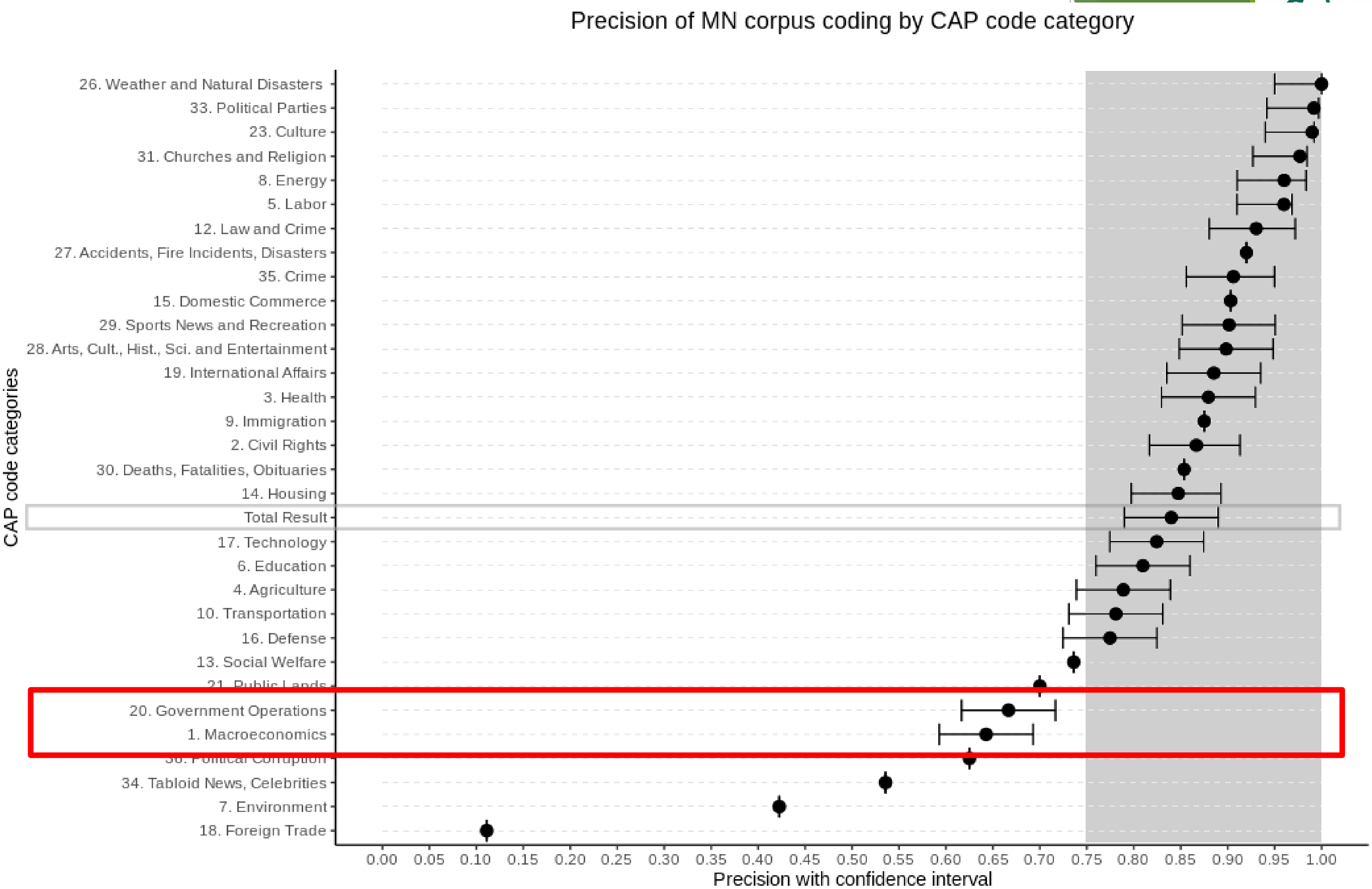
Major Topic	Coded Articles	Sample Size	Precision
1. Macroeconomics	3833	350	0.64
2. Civil Rights	207	135	0.87
3. Health	700	249	0.88
4. Agriculture	408	199	0.79
5. Labor	127	100	0.96
6. Education	436	205	0.81
7. Environment	71	71	0.42
8. Energy	542	226	0.96
9. Immigration	8	8	0.88
10. Transportation	459	210	0.78
12. Law and Crime	570	230	0.93
13. Social Welfare	72	72	0.74
14. Housing	168	118	0.85
15. Domestic Commerce	93	93	0.90
16. Defense	342	182	0.77
17. Technology	196	131	0.82
18. Foreign Trade	27	27	0.11
19. International Affairs	7617	366	0.89
20. Government Operations	1247	294	0.67
21. Public Lands	10	10	0.70
23. Culture	124	100	0.99
26. Weather and Natural Disasters	201	133	1.00
27. Accidents, Fire Incidents, Disasters	50	50	0.92
28. Arts, Cult., Hist., Sci. and Entertainment	677	246	0.90
29. Sports News and Recreation	385	193	0.90
30. Deaths, Fatalities, Obituaries	41	41	0.85
31. Churches and Religion	194	130	0.98
33. Political Parties	661	244	0.99
34. Tabloid News, Celebrities	28	28	0.54
35. Crime	339	181	0.91
36. Political Corruption	8	8	0.63
Total Result	19841	4630	0.84

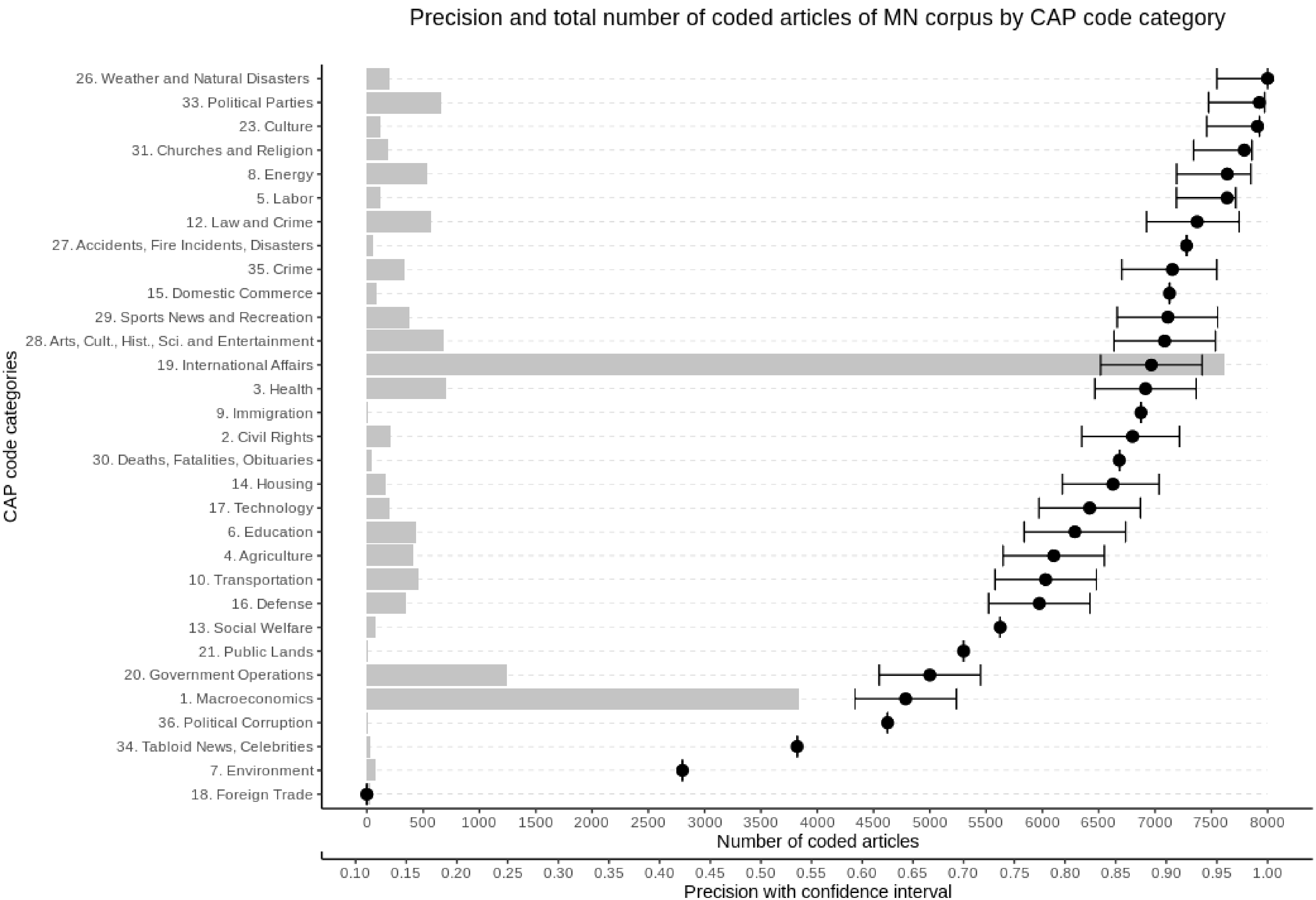


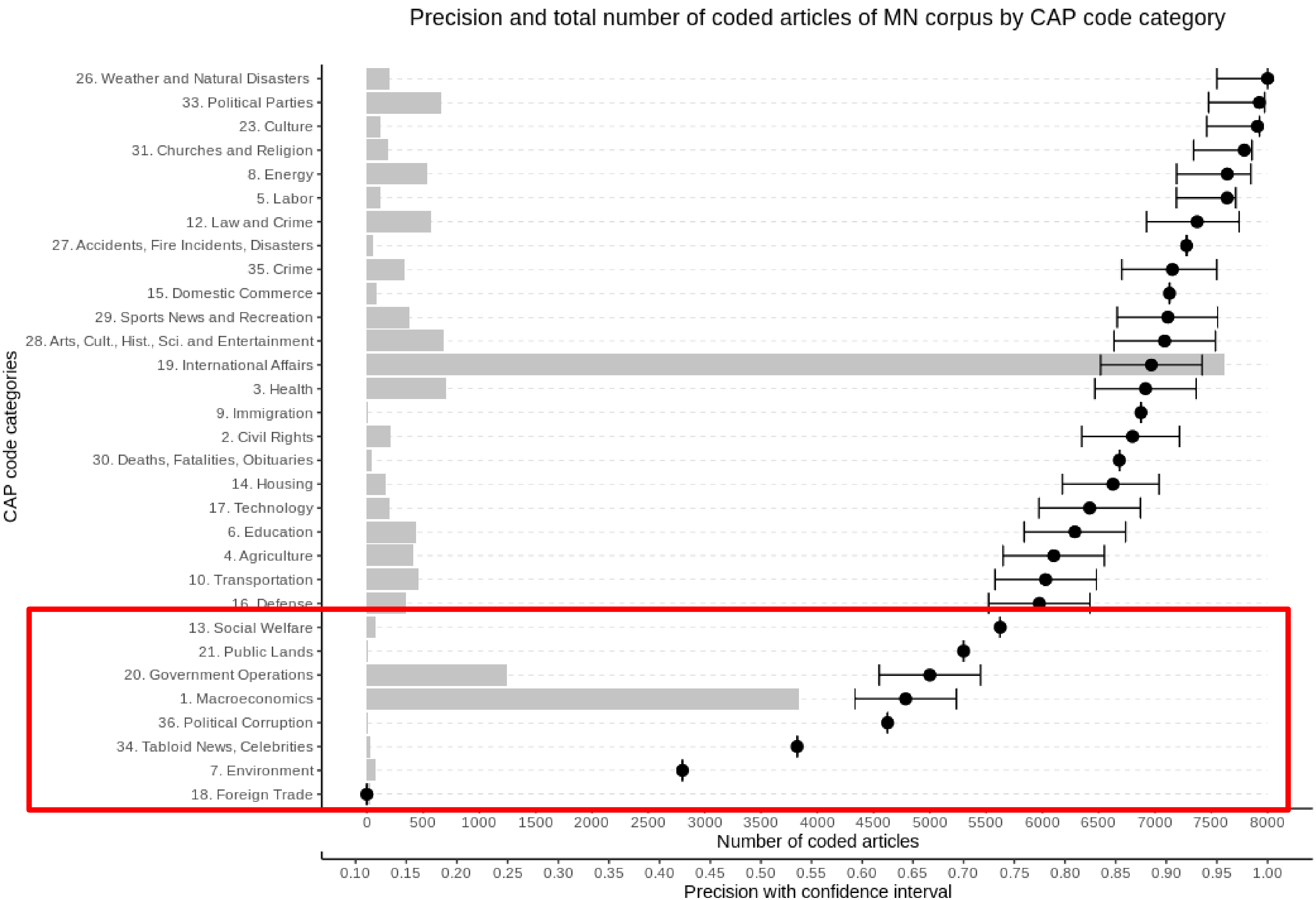
Precision of MN corpus coding by CAP code category
CAP code categories
26. Weather and Natural Disasters
33. Political Parties
23. Culture
31. Churches and Religion
8. Energy
5. Labor
12. Law and Crime
27. Accidents, Fire Incidents, Disasters
35. Crime
15. Domestic Commerce
29. Sports News and Recreation
28. Arts, Cult., Hist., Sci. and Entertainment
19. International Affairs
3. Health
9. Immigration
2. Civil Rights
30. Deaths, Fatalities, Obituaries
14. Housing
Total Result
17. Technology
6. Education
4. Agriculture
10. Transportation
16. Defense
13. Social Welfare
21. Public Lands
20. Government Operations
1. Macroeconomics
36. Political Corruption
34. Tabloid News, Celebrities
7. Environment
18. Foreign Trade
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90 0.95 1.00
Precision with confidence interval


Precision of MN corpus coding by CAP code category
26. Weather and Natural Disasters
33. Political Parties
23. Culture
31. Churches and Religion
8. Energy
5. Labor
12. Law and Crime
27. Accidents, Fire Incidents, Disasters
35. Crime
15. Domestic Commerce
29. Sports News and Recreation
28. Arts, Cult., Hist., Sci. and Entertainment
19. International Affairs
3. Health
9. Immigration
2. Civil Rights
30. Deaths, Fatalities, Obituaries
14. Housing
Total Result
17. Technology
6. Education
4. Agriculture
10. Transportation
16. Defense
13. Social Welfare
21. Public Lands
20. Government Operations
1. Macroeconomics
36. Political Corruption
34. Tabloid News, Celebrities
7. Environment
18. Foreign Trade
CAP code categories
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90 0.95 1.00
Precision with confidence interval

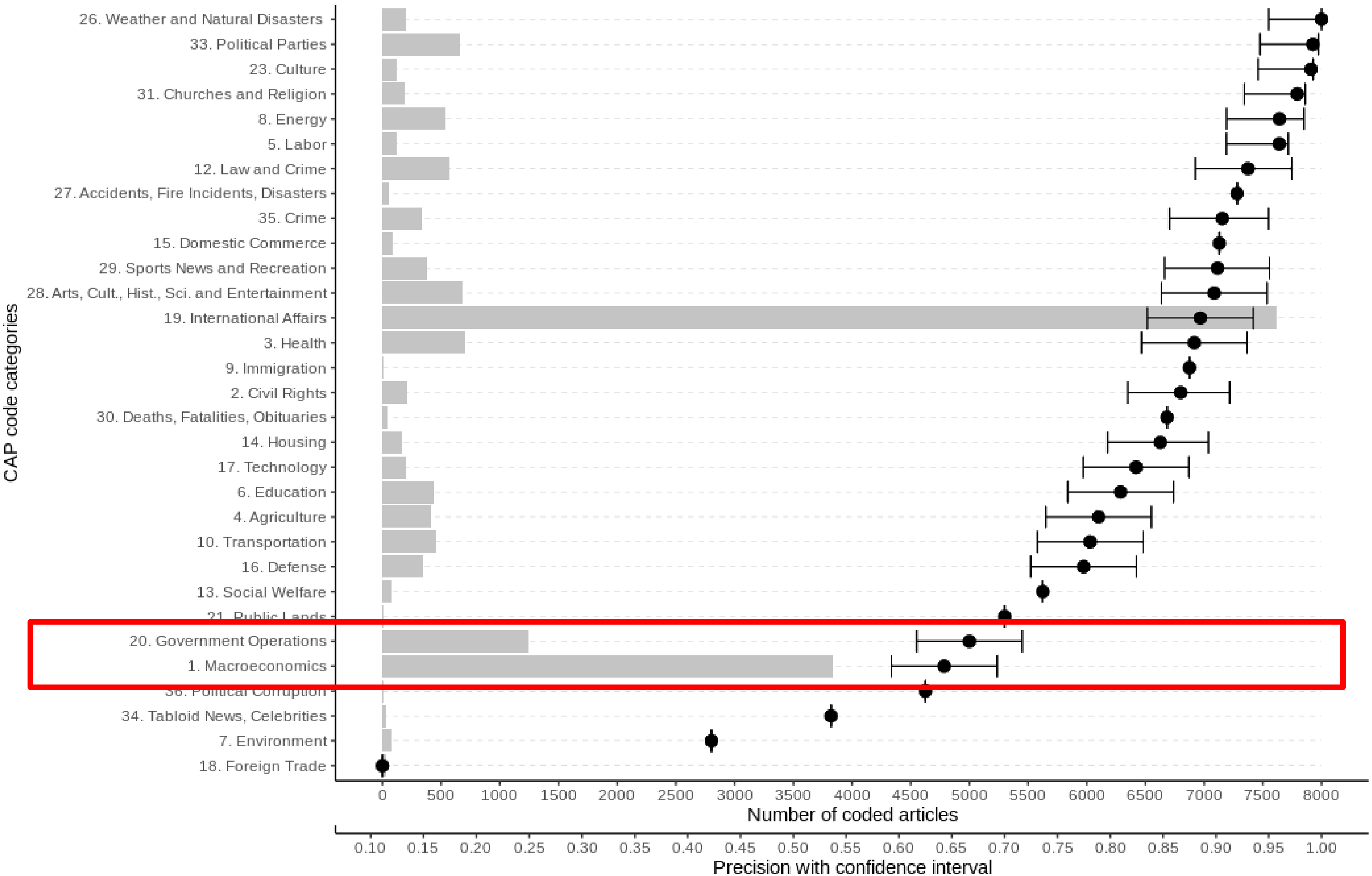

Precision of MN corpus coding by CAP code category
CAP code categories
26. Weather and Natural Disasters
33. Political Parties
23. Culture
31. Churches and Religion
8. Energy
5. Labor
12. Law and Crime
27. Accidents, Fire Incidents, Disasters
35. Crime
15. Domestic Commerce
29. Sports News and Recreation
28. Arts, Cult., Hist., Sci. and Entertainment
19. International Affairs
3. Health
9. Immigration
2. Civil Rights
30. Deaths, Fatalities, Obituaries
14. Housing
Total Result
17. Technology
6. Education
4. Agriculture
10. Transportation
16. Defense
13. Social Welfare
21. Public Lands
20. Government Operations
1. Macroeconomics
36. Political Corruption
34. Tabloid News, Celebrities
7. Environment
18. Foreign Trade
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90 0.95 1.00
Precision with confidence interval


Precision of MN corpus coding by CAP code category
26. Weather and Natural Disasters
33. Political Parties
23. Culture
31. Churches and Religion
8. Energy
5. Labor
12. Law and Crime
27. Accidents, Fire Incidents, Disasters
35. Crime
15. Domestic Commerce
29. Sports News and Recreation
28. Arts, Cult., Hist., Sci. and Entertainment
19. International Affairs
3. Health
9. Immigration
2. Civil Rights
30. Deaths, Fatalities, Obituaries
14. Housing
Total Result
17. Technology
6. Education
4. Agriculture
10. Transportation
16. Defense
13. Social Welfare
20. Government Operations
1. Macroeconomics
34. Tabloid News, Celebrities
7. Environment
18. Foreign Trade
CAP code categories
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90 0.95 1.00
Precision with confidence interval


Precision and total number of coded articles of MN corpus by CAP code category
CAP code categories
26. Weather and Natural Disasters
33. Political Parties
23. Culture
31. Churches and Religion
8. Energy
5. Labor
12. Law and Crime
27. Accidents, Fire Incidents, Disasters
35. Crime
15. Domestic Commerce
29. Sports News and Recreation
28. Arts, Cult., Hist., Sci. and Entertainment
19. International Affairs
3. Health
9. Immigration
2. Civil Rights
30. Deaths, Fatalities, Obituaries
14. Housing
17. Technology
6. Education
4. Agriculture
10. Transportation
16. Defense
13. Social Welfare
21. Public Lands
20. Government Operations
1. Macroeconomics
36. Political Corruption
34. Tabloid News, Celebrities
7. Environment
18. Foreign Trade
0 500 1000 1500 2000 2500 3000 3500 4000 4500 5000 5500 6000 6500 7000 7500 8000
Number of coded articles
0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90 0.95 1.00
Precision with confidence interval


Precision and total number of coded articles of MN corpus by CAP code category
CAP code categories
26. Weather and Natural Disasters
33. Political Parties
23. Culture
31. Churches and Religion
8. Energy
5. Labor
12. Law and Crime
27. Accidents, Fire Incidents, Disasters
35. Crime
15. Domestic Commerce
29. Sports News and Recreation
28. Arts, Cult., Hist., Sci. and Entertainment
19. International Affairs
3. Health
9. Immigration
2. Civil Rights
30. Deaths, Fatalities, Obituaries
14. Housing
17. Technology
6. Education
4. Agriculture
10. Transportation
16. Defense
13. Social Welfare
21. Public Lands
20. Government Operations
1. Macroeconomics
36. Political Corruption
34. Tabloid News, Celebrities
7. Environment
18. Foreign Trade
0 500 1000 1500 2000 2500 3000 3500 4000 4500 5000 5500 6000 6500 7000 7500 8000
Number of coded articles
0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90 0.95 1.00
Precision with confidence interval

Precision and total number of coded articles of MN corpus by CAP code category



Major Topic	Coded Articles	Sample Size	Precision
1. Macroeconomics	3833	350	0.64
2. Civil Rights	207	135	0.87
3. Health	700	249	0.88
4. Agriculture	408	199	0.79
5. Labor	127	100	0.96
6. Education	436	205	0.81
7. Environment	71	71	0.42
8. Energy	542	226	0.96
9. Immigration	8	8	0.88
10. Transportation	459	210	0.78
12. Law and Crime	570	230	0.93
13. Social Welfare	72	72	0.74
14. Housing	168	118	0.85
15. Domestic Commerce	93	93	0.90
16. Defense	342	182	0.77
17. Technology	196	131	0.82
18. Foreign Trade	27	27	0.11
19. International Affairs	7617	366	0.89
20. Government Operations	1247	294	0.67
21. Public Lands	10	10	0.70
23. Culture	124	100	0.99
26. Weather and Natural Disasters	201	133	1.00
27. Accidents, Fire Incidents, Disasters	50	50	0.92
28. Arts, Cult., Hist., Sci. and Entertainment	677	246	0.90
29. Sports News and Recreation	385	193	0.90
30. Deaths, Fatalities, Obituaries	41	41	0.85
31. Churches and Religion	194	130	0.98
33. Political Parties	661	244	0.99
34. Tabloid News, Celebrities	28	28	0.54
35. Crime	339	181	0.91
36. Political Corruption	8	8	0.63
Total Result	19841	4630	0.84

ELKH Cloud

Main contributions of HBS and the present study

- Enhance ML precision and recall by both **human input** (validation) and **workflow design** (one-vs-all classification, ensemble voting)
- Start working from a **limited training set**
- Able to **maximize ROI on human coding**
- **Move between (intra-domian) corpora**
- Take advantage of **cloud infrastructure** and **parallel processing** with Apache Spark

Further work

- Implement a finishing step using **regular expressions to correct systematic errors**
  - “design” in *Environment*
  - “icerink” in *Public Lands*
- Testing the HBS approach on **further languages**
- Generalizing the method to **other domains** beyond media

Thank you for your attention!

sebok.miklos@tk.mta.hu
kacsuk.zoltan@tk.mta.hu

poltext.tk.mta.hu
cap.tk.mta.hu