

Analysis and Visualization of Documents from Council Meetings

CZ: Analýza a vizualizace dokumentů ze zasedání zastupitelstev



Author **Bc. Martin Janeček** · Supervisor **Ing. Jiří Hynek, Ph.D.**
Master's thesis, defended at the **Faculty of Information Technology**,
Brno University of Technology · Department of Information Systems, 2026

Czech municipal councils decide on land sales, budget amendments and major infrastructure, publishing each decision as an unstructured PDF scattered across various portals without thematic labels. **I built an end-to-end system that transforms those PDFs into searchable, thematically tagged records**, from automated data collection to a fine-tuned Czech language model that classifies each agenda item into 17 predefined topics.

THE PROBLEM

Thousands of public decisions a year, with no way to search by topic, context, or agenda links.

While Czech municipalities are legally required to publish their council agendas, **every city uses different format**: scanned PDFs, HTML tables, XML exports, every city on its own portal, none of them labelled by thematic topic.

A journalist reviewing housing decisions of the past four years, or a resident following transport projects in their district, has to open hundreds of documents manually. **The data is public, yet effectively unsearchable.**

EXAMPLE OUTPUT

Tramvajová trať Plotní — majetkoprávní vypořádání pozemků

Zastupitelstvo souhlasí s nabytím pozemků dotčených stavbou tramvajové trati Plotní do vlastnictví statutárního města Brna.

doprava

majetek

územní plánování

THE SYSTEM

From a city portal to thematic labels, without a human in the loop.

- 1 Automated document ingestion**
Automated collection of public council documents from dynamic city portals, with content-hash deduplication.
- 2 Schema-driven PDF parser**
One JSON schema per municipality, adding a new city only requires writing a config, not new code.
- 3 RobeCzech multi-label classifier**
Czech RoBERTa fine-tuned on 2,113 hand-annotated items using an independent sigmoid per class.
- 4 Per-class thresholds**
Tuned on the validation set (0.30–0.85) to control precision and recall on rare topics.

Annotation quality beat data volume. Fixing incorrect labels on an unchanged set of 1,807 items gained 2.0 pp of test F1, more than any comparable increase in dataset size across 18 experiments.

RESULTS & IMPACT

0.900

TEST
MICRO F1

94.6 %

TRANSFER TO
A NEW CITY

2,113

ANNOTATED
AGENDA ITEMS

0.900 test micro F1, with strong zero-shot generalization to unseen cities.

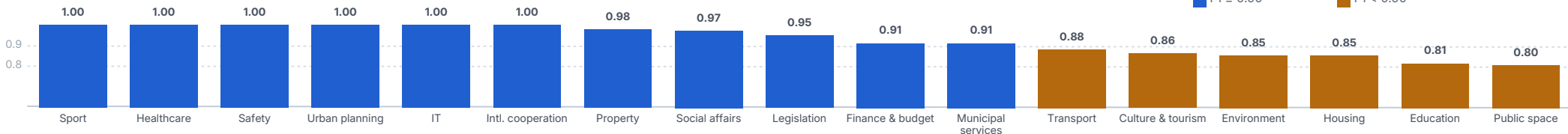
Model (identical data split)	MICRO F1	MACRO F1
TF-IDF + logistic regression	0.812	0.679
TF-IDF + LinearSVC	0.861	0.765
XLM-R base (multilingual)	0.899	0.871
CZERT-B base (Czech BERT)	0.917	0.892
RobeCzech, fine-tuned	0.900	0.884

Fine-tuning gains **+8.8 pp** micro and **+11.9 pp** macro F1 over the strongest non-neural baseline, LinearSVC. The 1.7 pp gap to CZERT-B equals roughly six labels in a 213-item test set, which is not significant without a statistical test.

Zero-shot transfer to Hradec Králové: 94.6% precision

712 agenda items from a city completely absent from training set were classified with no fine-tuning. Manual review of 93 sampled items confirmed that every item assigned a label was at least partially correct.

PER-CLASS PERFORMANCE ACROSS THE 17-TOPIC TAXONOMY



Per-class F1 on the validation set with per-class thresholds. Every one of the 17 topics exceeds 0.80 F1. The weakest classes align directly with the highest human annotator disagreement (mean pairwise Jaccard index of 0.663 across 50 re-labeled items). These category boundaries are inherently fuzzy, placing a natural ceiling on classification performance.



TRY IT

huggingface.co/spaces/martin0925/council-voting
Explore the live interactive demo. The pipeline feeds an **interactive visualization** layer built over Czech municipal council voting data.

RECOGNITION

Awarded by an industry partner
at **Excel@FIT 2026**.

CONTINUING AFTER THE DEFENCE

The pipeline is expanding to Prague, Most, Čelákovice, and the small village of Trpín. Successfully processing datasets from the capital down to a village of 400 residents proves that the system **generalizes seamlessly** across municipalities of any scale.