BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2026
https://excel.fit.vutbr.cz

Automated Classification of Czech Municipal Council Agenda Items

Bc. Martin Janeček*

Abstract

Czech municipalities are required by law to publish certain materials related to council meetings [1], but in practice, this public data remain difficult to use for further processing and analysis. The documents are scattered across municipal websites in a wide variety of formats and carry no thematic labels. This work presents an end-to-end pipeline for processing data from documents linked to agenda items of Czech municipal council meetings. City-specific scrapers collect materials, a schema-driven parser extracts structured text from heterogeneous PDF files, and a fine-tuned RobeCzech model performs multi-label classification into 17 thematic categories. The dataset of 2 113 manually annotated Brno City Council items was built through 18 iterative training rounds combining new annotation, error analysis, and taxonomy refinement. The final model achieves test micro F1 = 0.900.

*xjanec31@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

Municipal council meetings generate a continuous log of local governance, yet without thematic structure, this data remain an opaque heap—valuable in principle, but practically inaccessible to citizens, journalists, and researchers alike.

Czech city councils are legally required to publish certain documents related to their meetings, but there are no regulations governing the format of the data [1]. Meeting agendas, transcripts, resolutions, and other materials are scattered across individual city websites, published in inconsistent formats with no thematic labels. As a result, answering simple questions, such as how often the council has discussed affordable housing, environmental policy, or schools, requires manually reviewing countless unstructured documents.

This work addresses **multi-label text classification** of Czech council agenda items. Based on the title and annotation of an agenda item (typically 100–400 tokens after extraction), the model assigns zero or more labels from a flat taxonomy of 17 thematic categories. The task is challenging due to strong class imbalance (*property* appears in ∼40 % of items, *international cooperation* in under 3 %), overlapping labels, and frequent references to city-specific organizations.

RobeCzech [2], a monolingual RoBERTa [3] model pretrained on 4.9B Czech tokens, consistently outperforms multilingual baselines on Czech NLP benchmarks and is well suited for this domain, with a vocabulary rich in city-specific entities and abbreviations.

The main contributions are:

1. A reusable pipeline (**Figure 4**),
2. a 17-class taxonomy and 2 113-item annotated dataset (**Figure 2**),
3. a classifier reaching test micro F1 = 0.900 (**Figure 3**, **Figure 5**, **Figure 6**).

2. Data Analysis and Key Challenges

A survey of seven municipalities of varying types revealed significant diversity (**Figure 1**):

- **Format variability:** text-layer PDF, scanned PDF, DOCX, XML, and HTML.
- **Heterogeneous document structure:** each municipality uses a different document management system, resulting in incompatible PDF layouts that require specific parsing schemas.
- **Attachment noise:** agenda PDFs often embed scanned letters, maps, or budget tables that add noise for classification.
- **Unstable document–item mapping:** an item may have zero, one, or multiple documents.

This case study focuses in detail on the Brno City Council (ZMB), which has provided a stable and consistently structured corpus since 2018.

3. Pipeline and Annotation

The pipeline has four stages (**Figure 4**).

Scraping. City-specific scrapers are subclasses of `BaseScraper` and collect session lists, item metadata, and linked documents. PDF files are cached by content hash for incremental processing and deduplication.

Schema-driven PDF parsing. Each city–document-type pair is described by a JSON schema specifying the layout, regular expressions for metadata fields (session ID, date), and section delimiters (annotation, proposed resolution, reason report). This schema-based approach avoids hardcoded parsers. New municipalities can be added without modifying the base scraper.

Annotation in Label Studio [4]. Parsed records are imported with model-generated heuristic pre-labels as suggestions to speed up annotation. The taxonomy evolved through three versions: *culture* and *tourism* were merged due to high semantic overlap and consistent co-occurrence; *urban planning* was split into *urban planning* and *public space development* because the two areas are thematically distinct, and the model failed to separate them as a single class. Several categories (*family policy*, *investment*) were absorbed into other labels. All 2 113 items were annotated and reviewed by a single annotator, guided by an annotation hint list with defined boundary cases. The dataset is split into train / validation / test sets (1 683 / 216 / 214 items, approx. 80 / 10 / 10 %).

Classification. The primary model input is `title + "\n\n" + annotation` ($\leq$512 tokens). If the annotation is uninformative, the opening paragraph of the reason report serves as a fallback. The output is a vector of independent sigmoid activations over the 17 classes. At inference time, each class applies an independently tuned decision threshold. Items where no class exceeds its threshold are marked *unassigned*.

4. Training and Results

Following the fine-tuning paradigm of [5], RobeCzech is extended with a linear sigmoid head and fine-tuned with `BCEWithLogitsLoss` and positive-class weighting to compensate for class imbalance. The implementation uses Hugging Face Transformers [6]. The best configuration uses 18 epochs, learning rate 2×10^{-5}, and batch size 16 (**Figure 5**). After training, decision thresholds are tuned independently for each class on the validation set, consistently outperforming a single global threshold (e.g. Exp 05: global micro F1 = 0.782 vs. per-class 0.841). As a reference point, a TF-IDF + logistic regression baseline [7] (OneVsRest, trained on the same split) achieves micro F1 = 0.653 on the test set, confirming that the task benefits substantially from contextual representations.

The final dataset of 2 113 items was assembled across 18 experimental iterations (**Figure 6**). A central finding was that **annotation quality improvements consistently outweighed raw dataset expansion**: Exp 13 added 200 new items to a then-1 607-item dataset and improved test F1 by $+0.002$, Exp 14 corrected noisy labels in the same dataset without adding new items and improved test F1 by $+0.021$. This pattern repeated across the 18 iterations and shaped the core methodology. Each round prioritized error analysis and label correction over raw data collection.

The final model achieves val micro F1 = 0.911 and **test micro F1 = 0.900**. Per-class F1 on the validation set ranges from 0.80 to 1.00 (**Figure 3**). High-support categories (*sport*, *healthcare*, *property*: F1 = 1.00) score best, while *information technology* (F1 $\approx$ 0.75, only $\sim$50 examples) is the weakest. An interactive demo with learning curves, per-class results, and a live classifier is available at Hugging Face Space[1].

5. Conclusions

This work shows that high-quality multi-label classification of Czech municipal council agenda items is achievable with a moderate annotation effort, provided that the right taxonomy design and data quality are treated as primary concerns. Correcting inconsistent labels consistently outweighed adding new data:
Exp 13 (+200 items) improved test F1 by $+0.002$, while Exp 14 (label corrections only) improved it by $+0.021$.

Although the classifier was trained exclusively on Brno data, this need not limit cross-city transferability. The thematic taxonomy is independent of a specific municipality, and a model trained on a large, single-city corpus may generalize better than one trained on sparse multi-city data. The planned next step is to run inference on parsed items from Most and Hradec Králové and use the resulting pre-labels to seed an iterative annotation process for those municipalities—the same workflow that proved effective for Brno. The ultimate goal is to integrate the classifier into the Zastupko.cz [8] platform, making thematically labelled council data openly accessible at scale.

[1] https://huggingface.co/spaces/martin0925/council-voting

Acknowledgements

I would like to thank my supervisor, Ing. Jiří Hynek, Ph.D., for his guidance and support throughout this work.

References

[1] Česká republika. Zákon č. 128/2000 sb., o obcích (obecní zřízení). Sbírka zákonů České republiky, 2000. §93 (veřejnost zasedání), §95 (zápis a zveřejňování usnesení).

[2] Milan Straka, Jakub Náplava, Jana Straková, and David Samuel. RobeCzech: Czech RoBERTa-based language model. In *Proceedings of TLT 2021*, 2021.

[3] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[4] Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. Label studio: Data labeling software. https://github.com/heartexlabs/label-studio, 2020.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, 2019.

[6] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of EMNLP 2020: System Demonstrations*, pages 38–45, 2020.

[7] Fabian Pedregosa et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[8] Zastupko.cz — vizualizace hlasování zastupitelstev. https://zastupko.cz, 2025.