2026 European Conference on Computing in Construction
Corfu, Greece
July 12-15, 2026

FROM BUILDING PERMIT REVIEW COMMENTS TO COMMENT-BASED PROCESS INDICATORS: AN ESTBERT-BASED CLASSIFICATION WORKFLOW

Pille-Riin Peet[1], Ergo Pikas[2], and Aime Ruus[1]
[1] Tartu College School of Engineering, Tallinn University of Technology, Puiestee 78, 51008, Tartu, Estonia
[2] Building Lifecycle Research Group, Department of Civil Engineering and Architecture, Tallinn University of Technology, Ehitajate tee 5, 19086, Tallinn, Estonia

Abstract

Current building permit evaluation methods use broad metrics like duration and interviews, lacking detail and objectivity for targeted improvements. This paper introduces a workflow to convert unstructured reviewer comments into process indicators using EstBERT, a taxonomy, and supervised multi-label classification. Comments are categorized by issue type and domain, then combined into indicators of rework reasons, inconsistency, and transparency. On a validation set of 365 comments, the classifier scores micro-F1 of 0.470 and macro-F1 of 0.395. Results indicate reviewer comments can reliably support baseline assessment and inform targeted interventions across applications, municipalities, and work types.

Introduction

The building permit process and related procedures are a critical regulatory step in the delivery of construction projects (Hagedorn et al., 2025). However, it is often described as lacking transparency and predictability, with interpretive subjectivity and inconsistencies stemming from the volume and fragmentation of regulations, which lengthen the process and increase construction project costs (Peet et al., 2025). The process is complex, involving numerous stakeholders and disciplinary domains, and its implementation varies across jurisdictions due to differences in administrative practices, values, legal frameworks, and institutional structures (Fauth et al., 2024).

Given this complexity, commonly used metrics, such as total processing time and activity counts (e.g., review rounds and number of remarks), are insufficient because they describe outcomes without providing insight into why rework and iterations occur. On the other hand, effectiveness is frequently assessed through interviews and stakeholder surveys that capture concepts such as transparency, robustness, and citizen-friendliness. These perception-based assessments can be influenced by role-dependent and reporting biases, limiting their suitability as objective baselines. Emerging quantitative approaches, such as stakeholder management analysis metrics (Motahar et al., 2025), provide complementary insights into coordination and communication structures. However, they do not offer scalable, content-level evidence on the issues that cause review iterations in individual cases. In other words, current methods do not adequately support the diagnosis of process problems in a way that is sufficiently scalable and objective to enable benchmarking and monitoring of improvements.

Building permit process activity logs include unstructured comments that capture what reviewers found missing, unclear, inconsistent, or non-compliant, thereby providing direct evidence of why iterations occur. This paper examines whether building permit reviewer comments can be classified into a structured two-layer taxonomy with sufficient reliability to support the development of comment-based process indicators. More specifically, the study evaluates whether an Estonian-language transformer model (EstBERT) (Tanvir et al., 2021) can support multi-label classification of reviewer comments by issue type and disciplinary domain in a way that is useful for baseline assessment and intervention monitoring. The aim is to complement qualitative analyses with scalable indicators that reveal content-level reasons for rework requests across applications, municipalities, and work types. Rather than treating classification as an end in itself, this study uses it as an intermediate step in a broader workflow for transforming unstructured reviewer comments into structured evidence for process diagnosis.

Thus, the main contributions of this study are as follows:

- a workflow for deriving comment-based process indicators from unstructured building permit activity logs;
- a two-layer taxonomy for structuring permit review comments by issue type and disciplinary domain;
- an EstBERT-based multi-label classification pipeline that enables scalable application of this taxonomy to permit review comments.

This paper is structured as follows. The Methodology section describes the data source and the main workflow. In Data Preprocessing, we describe how the data were cleaned, how multi-item remarks were split, and how anonymization was implemented. Comment Classification covers the taxonomy, labeling strategy, and the EstBERT-based multi-label classification method. Deriving Metrics defines comment-based indicators and evaluation measures. The Results section describes the main findings, and the Discussion section discusses the implications, limitations, and possible directions for future research.

Methodology

The building permit process activity logs are from the Estonian Building Registry (EBR). EBR is a database designed to store, provide, and disclose information on planned, under-construction, and existing buildings, as well as related procedures. It is a nationwide environment used for all formal communication between the applicant and the reviewers during the building permit process. For this study, five municipalities representing a range of municipal sizes and annual application volumes (see Table 1) were selected. The building permit activity logs query was submitted to the EBR analyst for all applications in the five selected municipalities that ended in 2025 and started after 01.06.2022 (this date was selected because, in May 2022, EBR underwent a major update affecting all available services and related data).

Table 1: Selected municipalities by population size and annual application volume

	Size of the Municipality (inhabitants as of 2025)	Building permit and building notification applications in a year
A	< 5000	< 500
B	5000 – 10 000	< 500
C	10 000 – 20 000	< 500
D	20 000 – 50 000	500 – 1000
E	> 50 000	> 1000

Types of work planned in the building:

1. Renovation of a building
2. Building construction
3. Building extension
4. Demolition of a building
5. Replacement of part of a building with an equivalent

Table 2: Count of applications by municipality and work type

	A	B	C	D	E
1	0	24	47	111	1029
2	28	105	18	436	632
3	2	13	5	62	176
4	1	2	3	30	93
5	0	0	3	14	52

The dataset was exported as a .csv file containing administrative fields and process names, with column names in Estonian. To improve readability, we use English labels for these fields throughout the paper, while retaining the original column names in an accompanying data dictionary for reproducibility. This paper includes only the subset of fields used for preprocessing, model training, and indicator derivation. Additional fields present in the exports are documented but not used.

The proposed workflow uses reviewer comments as a source of structured evidence on the causes of rework in building permit review. To enable this, comments are classified using a two-layer label system and then aggregated into municipality- and work-type-level comment-based process indicators. Because reviewer comments are unstandardized and vary substantially in style, length, and specificity, the workflow combines rule-based preprocessing with supervised classification. Comment texts were tokenized using the tartuNLP/EstBERT checkpoint, accessed via the Hugging Face model hub and the Transformers library. EstBERT generated (Tanvir et al., 2021) contextual word embeddings were then used to train a multi-label logistic regression classifier using scikit-learn (Pedregosa et al., 2012). All preprocessing, classification, and indicator computation scripts are available on GitHub (https://github.com/pillepidu/EBR-comment-classification.git).

Data preprocessing

Before analysis, the raw building permit activity logs (91,056 rows) were cleaned to ensure data quality and relevance. The dataset was first filtered to include only document types classified as building permit applications ('Ehitusloa taotlus') and building notifications ('Ehitusteatis'), as these were the focus of the analysis. Rows where the comment goal ('eesmark') was marked as internal purpose ('EESMARK_SISEMINE') were removed, as these comments were deemed irrelevant for external classification. Furthermore, any rows with entirely empty values across the comment ('kommentaar'), reason ('pohjus'), and comment goal fields were excluded, as these likely represented system entries without meaningful information. To facilitate chronological analysis and ensure uniform representation, date columns such as application start ('algus'), application end ('lopp'), and date of entry ('tehtud_kp') were converted into pandas `datetime` objects (dd/mm/yyyy), and missing dates are kept as null values. The personal identification code ('isikukood') and registration code ('reg_kood') columns were converted into binary indicators, showing whether each column has a value or not.

The raw comment ('kommentaar') fields contained embedded HTML syntax, which is extraneous for NLP tasks and could introduce noise. To address this, HTML syntax was systematically removed using the `BeautifulSoup` library. Regular expressions were used to identify additional image tags and replace various image tag formats, such as `<img src="data:image/...">` (base64-encoded images) and `<img src="//:0">` (other image references), with the standardized placeholder `[pilt]`. This approach preserved the semantic information that an image was present in the original comment while removing lengthy and irrelevant character sequences.

In the 34,250 rows of activity logs, some of the comments didn't have a value, and many comments were a list of multiple comments, either bulleted (e.g., using '·') or numbered (e.g., '1. Text', '1) Text'). To enable granular analysis and accurate classification, rows with empty comments were filtered out. Comments that were lists were split into individual comment entries. The splitting logic specifically targeted patterns indicative of valid list items: a number followed by a period or parenthesis, and then an uppercase letter (marking the start of a new sentence). This allowed differentiation from legal references, which often contain numbers but lack the specific pattern indicating the start of a sentence. Each identified list item was then treated as a separate row, with all other metadata fields duplicated from the original entry. This resulted in 19,585 unique rows. The classifier was trained and evaluated on these split comment items. Indicator calculations then aggregate predicted labels over all comment items associated with the same application ID (and, where relevant, review round), treating a label as present if it occurs at least once within the group.

To protect personal data and comply with privacy regulations, a three-step anonymization process was applied to the comment fields. In the first two steps, regular expressions were used to anonymize email addresses and phone numbers. Standard email address patterns (e.g., `example@domain.ee`) were identified and replaced with the placeholder `[e-mail]`. Phone numbers that comply with Estonian regulations (*Eesti Numeratsiooniplaan–Riigi Teataja*, 2018) were detected and replaced with `[telefoni number]`. To prevent misclassification of regulation references, numbers containing a slash or preceded by 'nr.', 'nr', or 'nr.' were excluded from this anonymization. The third step was the anonymization of persons' names using the EstNLTK `NerTagger` (Laur et al., 2020), which assigned the 'PER' (person) tag to identify personal names. Once identified, these names were replaced with a sequence of 'X' characters of equal length (e.g., 'Kalle Kask' became 'Xxxxx Xxxx'). This method ensured that personal identifiers were obscured while preserving the textual structure.

Comment classification

To capture the content and context of building permit review comments beyond general procedural status labels, we developed a two-layer label system in which the primary labels refer to the type of problem raised in the comment, and the second layer, the domain tag, describes the disciplinary context. The initial list of categories was developed based on an existing legal analysis of the requirements for construction projects in Estonia (Sorainen & TalTech, 2024). Because reviewer comments often contain multiple requests or refer to closely related problems, the label system was designed as a multi-label scheme in which a single comment may receive multiple primary labels and multiple domain tags. Some category boundaries are necessarily adjacent, particularly between comments about missing information, inconsistency, incorrect content, and requests for clarification. To reduce ambiguity, short decision rules were used during manual labeling. However, some overlap remained and is treated as an inherent characteristic of the comment data rather than as a purely technical modeling issue.

To train a multi-label classifier, we first extracted a subset of 2048 unique preprocessed comments for manual labeling. The subset of comments was randomly selected; to ensure an even distribution across municipalities and application types, we ensured that there were 200 unique comments for each municipality and application type. If there weren't enough of them, then all of them were chosen for the manual training set.

During the manual labeling phase, the label set was refined to better reflect recurring patterns in the reviewers' comments. The manually labeled subset was divided into a training set of 1682 labeled comments and a validation set of 365 labeled comments. The training set was used to train the initial classifier, which was applied to automatically label the remaining comments.

Primary Labels:

P1. Missing Documentation or Information
Required documents, parameters, information, or attachments are missing and must be provided.

P2. Incorrect or Error
There is an incorrect value, detail, or error in the submitted content.

P3. Inconsistent
Two submitted documents contradict one another (e.g., an explanatory document and a site plan).

P4. Non-Compliance with Regulations
Comment states non-compliance with law/regulation/standard (often with legal reference).

P5. Non-Compliance with Design Requirements or Detail plans or General Plans
Conflict with the detailed plan, general plans, or design conditions. Often explicitly stated.

P6. Unclear or Needs Clarification
Information exists but is ambiguous. The reviewer requests clarification or justification.

P7. Format or Presentation (e.g., missing or wrong signatures)

P8. Process or status (e.g., EBR form or procedural issues, formal Statements of approval or refusal without explanation)
Remarks regarding EBR forms or workflow issues rather than technical content.

P9. Requirements for stakeholder approval
Requires consent or approval from another authority or party (e.g., neighbors, utility service providers, Environmental agencies).

P10. No Issues
Explicit statement that there are no remarks.

Domain Labels:

D1. Fire Safety
D2. Heritage
D3. Site
D4. Utilities

D5. Administrative or Procedural
D6. Environmental
D7. Waste Management
D8. Energy efficiency

EstNLTK (Laur et al., 2020) was used to preprocess Estonian-language comment texts before automatic classification. Each comment was sentence-segmented and tokenized into word tokens. For tokenization, each comment text was converted into an EstNLTK `Text` object and tokenized using the `words` layer. For sentence segmentation, the same `Text` object was annotated with the `sentence` layer to provide sentence boundaries for downstream NLP processing. Each comment was then tokenized using EstBERT (Tanvir et al., 2021) based embeddings, a transformer-based language model trained on Estonian text. Contextual embeddings were generated using EstNLTK `BertTagger`, which assigns a BERT embedding to each token in the comment text after annotating the text with the `words` and `sentences` layers. Token-level embeddings were then aggregated into a single fixed-length vector per comment by calculating the arithmetic mean across all token embeddings (mean pooling). This comment-level tokenization supports downstream similarity-based analysis while allowing comments to be associated with multiple issue categories in subsequent steps, even when a single comment contains multiple distinct issues.

Deriving metrics

Following model training, the final classifier was applied to the complete preprocessed dataset to assign primary labels and domain tags to each reviewer comment. These predicted labels were then added to the activity logs and aggregated into comment-based process indicators for rework, inconsistency (e.g., review rounds and recurring issue types), and procedural transparency (e.g., the presence of explicit legal references). The indicators were calculated for the full dataset and also by municipality and intended construction work type. For multi-label outputs, each assigned label was counted once when calculating category frequencies. For application-level metrics (e.g., whether an application contains issues about missing information or documentation), a category is counted as present if it occurred in at least one comment associated with that application.

Model performance was evaluated using the validation set of 365 manually labeled comments (20% of the labeled dataset). Because the task is multi-label and the label distribution is imbalanced, the primary evaluation metrics are micro- and macro-averaged precision, recall, and F1 score, together with per-label precision, recall, and F1 for the primary labels. For completeness, we also report label-cell accuracy as a supplementary metric. For each label ℓ, we compute true positives (TPℓ), false positives (FPℓ), and false negatives (FNℓ) from binary predictions. Macro-averaged precision, recall, and F1 were calculated by averaging the per-label scores, giving equal weight to each label. Micro-averaged precision, recall, and F1 are calculated by first summing TPℓ, FPℓ, and FNℓ across all labels, and then computing precision, recall, and F1-score, with label weights proportional to their frequency. Additionally, we conducted a qualitative error analysis by reviewing misclassified samples to identify the causes of systemic failures (e.g., overlapping category boundaries, multi-problem comments not fully separated during preprocessing, and domain-specific phrasing).

In addition to the predicted labels, each comment was tagged for the presence of an explicit legal reference (has_legal_reference: True/False) using regular expression pattern matching for citation markers and keywords commonly used in Estonian, such as "§", "EhS", "määrus", "standard", "juhend", "seadus", and "nõuded". This flag was used to quantify the proportion of comments that provide explicit legal justification, both overall and by municipality.

Table 3 summarizes the comment-based process indicators used in the quantitative analysis.

Table 3: Overview of metrics derived from EBR activity logs

Theme	Question	Analysis and derived metrics
Inconsistencies and subjectivity	How many applications require multiple review rounds?	Calculates R=max('ringi_nr') per application. Reports the percentage of applications where R > 1 overall, by municipality, and work type.
	What are the most common reasons for revision requests?	Identifies comments, both primary and domain tags. Reports the most common labels overall, by municipality, and work type.
Procedural transparency and a generally worded legal framework	How many comments contain an explicit reference to any legal document, standard, or guide as justification for the comment?	Identifies comments where `has_legal_reference`=True. Reports the percentage of comments with legal references overall, by municipality, and by work type.
Inconsistency and procedural transparency	How often do requirements shift across review rounds within the same application?	Tracks both the primary and domain tags per application that first appear when the review round > 1. Reports the percentage of applications with late-appearing labels, the most common late-appearing labels overall, and how many times they occur.

Results

Classification performance

Table 4 summarizes overall performance using micro- and macro-averaged precision, recall, and F1, reported separately for the primary labels (P1–P10), the domain labels (D1–D8), and all labels combined. Label-cell accuracy is also reported as a supplementary metric across the validation set.

Overall, domain labels (D1–D8) are predicted more reliably than primary labels (P1–P10), with higher label-cell accuracy (0.893 vs 0.778) and macro-F1 (0.437 vs 0.362). The gap between micro and macro scores indicates uneven performance across labels, with labels with less training data lowering the macro-averaged results. Table 5 reports per-label performance for the primary labels. Performance varies substantially across labels, with the highest F1 observed for "Process or status" (P8, F1 = 0.662) and the lowest for "No issues" (P10, F1 = 0.143).

Table 4: Overall performance of the classifier

	Primary labels (P1-P10)	Domain labels (D1-D8)	All labels (primary + domain)
Micro Precision	0.332	0.395	0.351
Micro Recall	0.689	0.761	0.712
Micro F1	0.448	0.520	0.470
Macro Precision	0.274	0.349	0.307
Macro Recall	0.620	0.638	0.628
Macro F1	0.362	0.437	0.395
Accuracy	0.778	0.893	0.829

Table 5: Performance for primary labels (P1-P10). P1 - Missing Documentation or Information, P2 - Incorrect or Error, P3 – Inconsistent, P4 - Non-Compliance with Regulations, P5 - Non-Compliance with Design Requirements or Detail plans or General Plans, P6 - Unclear or Needs Clarification, P7 - Format or Presentation, P8 - Process or status, P9 - Requirements for stakeholder approval, P10 - No Issues

	Precision	Recall	F1
P1	0.437	0.874	0.582
P2	0.126	0.621	0.209
P3	0.358	0.763	0.487
P4	0.125	0.438	0.194
P5	0.136	0.75	0.231
P6	0.363	0.589	0.449
P7	0.194	0.619	0.295
P8	0.62	0.71	0.662
P9	0.247	0.686	0.364
P10	0.133	0.154	0.143

Additionally, we conducted a qualitative error analysis by manually reviewing misclassified validation samples (false positives and false negatives) to identify recurring causes of systematic failures. Two main problems were detected. First, overlapping category boundaries led to confusion between similar labels, especially when a comment could belong to more than one primary category. Second, the splitting preprocessing function failed to capture multiple requests or issues in a single text. The model tended to either assign only a subset of the relevant labels (false negatives) or add an additional label triggered by a specific phrase (false positives). This helps explain the uneven per-label performance and highlights that both label-definition clarity and data-preparation exceptions are key drivers of systematic misclassification.

Frequency and reasons for rework

Table 6 reports the share of applications that required multiple review rounds (R > 1) by municipality (marked A-E as described in Methodology) and intended work type (marked 1-5 as described in Methodology). Across the full dataset (N = 2,804 applications), 1,498 applications (53.4%) required more than one review round, with an average of 2.10 rounds per application.

Table 6: Percentage of applications that require multiple review rounds by municipality and work type (%)

	A	B	C	D	E
1	0.0	41.7	38.3	43.2	50.4
2	25.0	31.4	61.1	46.8	72.2
3	50.0	38.5	20.0	54.8	63.6
4	0.0	0.0	33.3	30.0	62.4
5	0.0	0.0	33.3	85.7	25.0

Table 7 summarizes the most common issue types driving revision requests, broken down by municipality and work type. Across all municipalities and work types, the classifier assigned 51,051 primary labels in total (each label counted once per comment). The three most frequent issue types were "Missing Documentation or Information" (P1, 9,573), "Unclear or Needs Clarification" (P6, 8,519), and "Incorrect or Error" (P2, 7,711). This indicates that rework mainly results from incomplete submission documentation, unclear information, and specific errors in the submitted documents.

Table 7: Most common reasons for revision requests by municipality and work type (Top 3, P - Primary label). P1 - Missing Documentation or Information, P2 - Incorrect or Error, P3 – Inconsistent, P4 - Non-Compliance with Regulations, P5 - Non-Compliance with Design Requirements or Detail plans or General Plans, P6 - Unclear or Needs Clarification, P7 - Format or Presentation, P8 - Process or status, P9 - Requirements for stakeholder approval, P10 - No Issues

	Top3	A	B	C	D	E
1	1	-	P1	P1	P1	P1
	2	-	P2	P6	P2	P2
	3	-	P6	P2	P8	P6
2	1	P6	P1	P1	P8	P1
	2	P8	P8	P2	P1	P6
	3	P1	P9	P6	P6	P2
3	1	P1	P2	P1	P1	P1
	2	P9	P1	P2	P6	P6
	3	P8	P4	P6	P2	P2
4	1	P2	P1	P8	P8	P6
	2	P6	P6	P1	P6	P1
	3	-	P8	P2	P1	P2
5	1	-	-	P1	P1	P8
	2	-	-	P3	P8	P1
	3	-	-	P2	P3	P3

Table 8 summarizes the most common issue domains driving revision requests by municipality and work type. Across all municipalities and work types, the classifier assigned 20,481 domain labels. The three most frequent were “Administrative or Procedural” (D5, 5,790), “Site” (D3, 5,094), and “Waste Management” (D7, 2,622). This shows that revision requests are influenced not only by technical compliance issues but also by procedural requirements and site-specific factors.

Table 8: Most common reasons for revision requests by municipality and work type (Top 3, D - Domain label). D1 - Fire Safety, D2 - Heritage, D3 - Site, D4 - Utilities, D5 - Administrative or Procedural, D6 - Environmental, D7 - Waste Management, D8 - Energy efficiency

	Top3	A	B	C	D	E
1	1	-	D1	D5	D5	D5
	2	-	D5	D2	D1	D3
	3	-	D7	D3	D3	D7
2	1	D5	D5	D5	D5	D3
	2	D3	D3	D3	D3	D5
	3	D1	D1	D7	D7	D7
3	1	D3	D1	D5	D5	D5
	2	D5	D5	D1	D1	D3
	3	-	D3	D7	D2	D7
4	1	-	D7	D5	D5	D5
	2	-	D5	D3	D3	D3
	3	-	-	D2	D7	D7
5	1	-	-	D5	D5	D5
	2	-	-	-	D3	D3
	3	-	-	-	D7	D1

Taken together, Tables 7 and 8 demonstrate that the primary drivers for revision requests are not limited to a single technical discipline. Instead, they combine issues related to submission completeness, information clarity and correctness, administrative procedures, and site-specific requirements. This supports the value of comment-based process indicators, as they offer a more nuanced understanding of the causes of rework than high-level process metrics alone.

Transparency of the rework requests

Table 9 reports the share of comments that explicitly reference a legal document, standard, or guide. Overall, 26.9% of comments contained explicit legal references. Variations in referencing practices across municipalities and work types indicate differences in how reviewers justify remarks and the transparency of the written feedback.

Table 9: Percentage of comments that contain an explicit reference to any legal document, standard, or guide as justification for the comment by municipality and work type (%)

	A	B	C	D	E
1	0.0	28.8	38.0	32.8	32.2
2	38.7	36.2	48.6	34.9	20.4
3	0.0	27.9	16.7	24.0	31.5
4	0.0	50.0	0.0	29.8	28.4
5	0.0	0.0	20.0	39.4	44.5

Table 10 reports the share of applications in which issue types or domain tags first appear after the first review round (“late-appearing labels”). Overall, 4.7% of applications had late-appearing labels, suggesting that most issues are identified early, but that a subset of cases experience shifting requirements across rounds. The five most frequent late-appearing labels were “Process or status” (P8, 436 times), “Inconsistent” (P3, 416), “Administrative or Procedural” (D5, 414), “Requirements for stakeholder approval” (P9, 356), and “Unclear or Needs Clarification” (P6, 350).

Table 10: Percentage of applications with late-appearing labels by municipality and work type (%)

	A	B	C	D	E
1	0.0	12.5	19.6	2.8	4.2
2	25.0	10.6	27.8	5.8	2.5
3	0.0	23.1	0.0	0.0	0.6
4	0.0	0.0	33.3	6.7	2.3
5	0.0	0.0	33.3	14.3	0.0

Discussion

This study shows that unstructured building permit review comments can be structured into categories and aggregated into comment-based process indicators that complement both qualitative assessments (e.g., perceived transparency) and high-level quantitative measures (e.g., total application duration). The results indicate substantial variation in review rounds (Table 6), the types of issues raised in revision requests (Tables 7–8), and the use of explicit legal references in comments (Table 9). Taken together, these indicators provide a more diagnostic view of the permitting processes by showing not only how much rework occurs, but also what kinds of issues drive it and how review practices differ across jurisdictions and work types.

The proposed indicators should be interpreted as signs of process characteristics rather than direct measures of performance quality or administrative performance. For example, a higher number of comments or legal references may reflect a thorough and transparent review rather than inefficiency. Conversely, fewer comments could indicate either efficient early-stage collaboration between the applicant and the reviewer or limited feedback from the reviewer. Additional review rounds may result from project complexity, applicant experience, or local administrative practices. For this reason, comparison across municipalities or work types should be interpreted in context and, where possible, alongside qualitative analysis.

The evaluation conducted in this study is based on activity logs from the Estonian Building Registry (EBR) and on an Estonian-language transformer model, EstBERT. The taxonomy was developed through analysis of Estonian law, and the label definitions reflect local regulations and administrative practices. Applying the workflow in other contexts would therefore require adaptations, including revising the label set, collecting new annotated data, and retraining or adjusting the language model for the target language and document style. Nevertheless, the core workflow - preprocessing comments, multi-label categorization, and aggregation into comment-based process indicators - remains reusable across contexts.

Manual labeling may introduce bias, particularly when performed by a single labeler. Certain categories, such as "Missing Documentation or Information," "Unclear or Needs clarification," "Incorrect or Error," and "Inconsistent," are semantically adjacent and may therefore be interpreted differently across labelers. Short definitions and decision rules were used to reduce ambiguity, but future research should involve multiple labelers and report inter-labeler agreement. This would strengthen the reliability of label assignments, clarify category boundaries, and improve the interpretability of both the classification results and the indicators derived from them.

The activity logs show variation in municipal reviewing practices. Some municipalities provide brief comments, whereas others use long lists or aggregate multiple review requests into a single entry. The splitting method used in this paper extracts list items but fails to separate multiple issues expressed within a single sentence, or may incorrectly split text that resembles a list. These preprocessing problems can affect both classification performance and resulting indicators. Future improvements should therefore focus on more robust sentence-level splitting, improved list detection, and targeted handling of complex comment structures identified through error analysis.

The classification approach used in this study relies on EstBERT-based embeddings with mean pooling and logistic regression, providing a transparent and locally executable baseline for privacy-sensitive administrative text. However, this approach may overlook fine-grained contextual cues within longer or more complex comments and may perform weakly on rare or semantically adjacent labels. This is reflected in the gap between the micro- and macro-averaged results, reinforcing the need for cautious interpretation of category-level outputs. A promising direction for future research is to extend the workflow toward a more explainable AI setting in which the classification-based approach presented in this paper is complemented by an LLM-based method applied in parallel. In such a setup, the structured classifier would support auditable category assignments, while the LLM-based component could help interpret more complex or ambiguous comments, for example, by directing low-confidence predictions to LLM-based classification, decomposing multi-issue comments before labeling, or generating post hoc justifications for human review. Potential improvements, therefore, include threshold tuning, alternative aggregation strategies, stronger handling of long comments, and comparative or parallel evaluation against more recent language-model-based approaches, where privacy and deployment constraints permit.

Conclusions

This study demonstrates that reviewer comments in building permit activity logs can be structured into comment-based process indicators that make recurring causes of rework and review practices visible at scale. In contrast to high-level process measures such as total duration or number of review rounds, these indicators provide more specific evidence about the content of revision requests and the issues that drive iterative review.

Their main value lies in supporting baseline assessment, cross-case comparison, and the identification of dominant issue patterns across municipalities and work types. The three-stage workflow, consisting of preprocessing, classification, and indicator aggregation, is transferable to other jurisdictions by replacing the target-language transformer and the locally derived taxonomy and training set, while the indicator definitions remain unchanged.

The proposed workflow can support future building permit process improvement efforts by helping identify dominant issue categories, such as missing documentation, unclear submissions, or procedural bottlenecks within each jurisdiction. This helps inform improvement efforts, such as clearer reviewer templates, better applicant guidance, or the addition of a pre-consultation phase. Future research should evaluate how such interventions affect specific comment-based indicators over time and should further strengthen the workflow through improved preprocessing, stronger annotation procedures, and comparative or parallel model evaluation.

Acknowledgments

This work has been supported by the Estonian Research Council grant (PSG963).

References

Eesti numeratsiooniplaan–Riigi Teataja. (2018). https://www.riigiteataja.ee/akt/110042024005

Fauth, J., Nørkjær Gade, P., Kaiser, S., Raj, K., Goul Pedersen, J., Olsson, P.-O., Nisbet, N., Mastrolembo Ventura, S., Hirvensalo, A., Granja, J., Urban, H., Rutešić, S., Verstraeten, R., Raitviir, C.-R., Kallinen, A.-R., Schranz, C., Stojanov, T., & Tekavec, J. (2024). Investigating building permit processes across Europe: Characteristics and patterns. Building Research & Information, 0(0), 1–18. https://doi.org/10.1080/09613218.2024.2400467

Hagedorn, P., Fauth, J., Zentgraf, S., Seiß, S., König, M., & Brilakis, I. (2025). OntoBPR: An ontology-based framework for performing building permit reviews using standardized information containers. Advanced Engineering Informatics, 66, 103369. https://doi.org/10.1016/j.aei.2025.103369

Laur, S., Orasmaa, S., Särg, D., & Tammo, P. (2020). EstNLTK 1.6: Remastered Estonian NLP Pipeline. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds), Proceedings of the Twelfth Language Resources and Evaluation Conference (pp. 7152–7160). European Language Resources Association. https://aclanthology.org/2020.lrec-1.884/

Motahar, H., Fauth, J., & Melzner, J. (2025). Evaluating Stakeholder Management Practices in the Building Permit Process. Journal of Legal Affairs and Dispute Resolution in Engineering and Construction, 17(3), 04525035. https://doi.org/10.1061/JLADAH.LADR-1310

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., & Louppe, G. (2012). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12.

Peet, P.-R., Pikas, E., & Ruus, A. (2025). Lost in Regulations: Digitalization as the Key to Transparent and Efficient Building Permit Evaluation in Estonia. Computing in Construction, 6, 0–0. https://doi.org/10.35490/EC3.2025.432

Sorainen & TalTech. (2024). Ehitusprojektile esitatavate nõuete õigusanalüüs. https://kliimaministeerium.ee/sites/default/files/documents/2024-12/Ehitusprojektile%20esitatavate%20n%C3%B5uete%20%C3%B5igusanal%C3%BC%C3%BCs.pdf

Tanvir, H., Kittask, C., Eiche, S., & Sirts, K. (2021). EstBERT: A Pretrained Language-Specific BERT for Estonian. In S. Dobnik & L. Øvrelid (Eds), Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa) (pp. 11–19). Linköping University Electronic Press, Sweden. https://aclanthology.org/2021.nodalida-main.2/