Karlsruhe Institute of Technology

INSTITUTE OF INFORMATION SYSTEMS AND MARKETING

Identifying Municipal Climate Policy Measures in Unstructured Text: An Instrument-Based, Precision-Oriented Machine Learning Approach

Seminar Thesis

Sven M. Lutz

2663304

At the Department of Economics and Management
Institute of Information Systems and Marketing (IISM)
Information Systems I

Reviewer:	Prof. Dr. rer. pol. Christof Weinhardt
Supervisor:	Prof. Dr. Marie-Louise Arlt

21.01.2026

Abstract

Municipalities play a central role in the implementation of climate policy, yet systematic empirical analysis of implemented municipal climate policy measures remains limited. A key obstacle is that relevant evidence is dispersed across heterogeneous, unstructured municipal web texts, while reliable annotations are scarce. Existing automated approaches often rely on thematic climate relevance and therefore conflate discourse, strategic intent, and communication with concrete policy action. This study proposes a theory-driven framework for identifying municipal climate policy measures in text under severe label scarcity. Grounded in instrument-based policy design theory, policy measures are defined strictly as explicit, operational policy instruments with textual evidence of implementation, allocation, or enforcement. To operationalise this definition at scale, the approach combines a reject-option decision architecture with conservative weak supervision and confidence-controlled bootstrapping, treating abstention as a first-class outcome. Large language models are integrated as controlled discriminative components within this structured pipeline rather than as autonomous end-to-end classifiers. The framework is evaluated on a manually annotated gold standard from German municipal web texts drawn from three Bavarian municipalities (Munich, Bayreuth, and Günzburg). Results indicate that reliable identification is feasible even with a very small, highly imbalanced labelled dataset: conservative label expansion preserves a precision-first operating regime, while an unconstrained LLM-only classifier exhibits elevated false positives by over-classifying action-oriented discourse as policy measures. Overall, the study demonstrates that conservative, theory-informed machine learning can reconcile policy-theoretic rigour with scalable text analysis and provides a robust methodological basis for implementation-oriented measurement of municipal climate policy in real-world data environments.

Contents

List of Figures

List of Tables

List of Abbreviations

F1 F1-score. 11–13, 16

FN False Negative. 5, 11, 18, 22

FP False Positive. iv, 1, 5, 9, 11, 15–17, 22, 25

IE Information Extraction. 19, 20

LLM Large Language Model. iv, 2, 16–25

ML Machine Learning. 25

TN True Negative. 18

TP True Positive. 11

1. Introduction

Municipalities constitute a central implementation level in climate governance. They translate national and supranational objectives into locally binding interventions through concrete policy instruments such as building regulations, funding schemes, infrastructure investments, and administrative programmes (Hood, 1983; Peters, 2018). As a result, municipal action plays a decisive role in shaping the practical effectiveness of climate policy. Despite this importance, systematic empirical evidence on municipal climate policy implementation remains limited. A primary obstacle lies in the nature of the available data. Information on municipal climate-related activities is dispersed across heterogeneous and largely unstructured web-based texts, including press releases, service pages, strategy documents, project descriptions, and legal notices. These sources differ substantially in purpose, style, and degree of formalisation and are therefore not readily amenable to large-scale automated analysis (Gentzkow et al., 2019; Grimmer & Stewart, 2013; Wilkerson & Casas, 2017). Consequently, it remains difficult to systematically observe and compare implemented municipal policy interventions rather than communicative statements or symbolic commitments. This empirical difficulty is compounded by a conceptual challenge. Municipal climate communication frequently interweaves strategic goals, targets, narratives, and references to ongoing initiatives with descriptions of concrete interventions. While such discourse is an integral component of policy-making, it does not in itself constitute policy implementation unless it is operationalised through specific instruments. From an instrument-based policy design perspective, the policy instrument represents the operative unit of analysis: a concrete mechanism through which public authorities seek to influence behaviour, allocate resources, or shape socio-technical systems (Capano & Howlett, 2020; Hood, 1983; Lascoumes & Le Galès, 2007). Automated text analysis approaches that rely primarily on thematic relevance or climate-related keywords therefore conflate policy discourse with policy action and systematically overstate implementation activity (Gentzkow et al., 2019; Grimmer & Stewart, 2013). Under an instrument-based conception of public policy, thematic relevance is not a sufficient proxy for implementation. To address this measurement problem, this study advances a deliberately conservative task definition that departs from prevalent theme-based operationalisations in automated policy analysis. A text segment is classified as a municipal climate policy measure only if it specifies an operational policy instrument with explicit textual evidence of implementation, allocation, or enforcement, rather than merely articulating a goal, commitment, or general intention. This restriction is not a methodological limitation but a necessary condition for maintaining conceptual validity when analysing policy implementation in unstructured text. Operationalising such a definition poses substantial methodological challenges. Manually annotated training data are scarce, the positive class is rare, and misclassifications are asymmetrically costly from a policy-analytic perspective. In particular, FP identifications of policy measures would introduce spurious interventions and distort downstream analyses of policy instrument portfolios. Against this background, this study proposes a theory-driven approach for identifying municipal climate policy measures in unstructured

web-based text under conditions of severe label scarcity. Rather than treating LLMs as autonomous end-to-end classifiers, a practice that risks systematic over-attribution of policy action under instrument-based definitions, the approach embeds LLMs as controlled, discriminative components within a structured decision framework that enforces explicit uncertainty handling and abstention (Chow, 1970; Ratner et al., 2017; Settles, 2009). The methodological design prioritises precision, interpretability, and conceptual validity over exhaustive coverage and is explicitly aligned with instrument-based policy design theory. The central contribution of this study is threefold. First, it demonstrates how instrument-based policy design theory can be operationalised as a strict task definition for automated text analysis of municipal policy implementation. Second, it shows that precision-oriented identification of municipal climate policy measures is feasible even under extreme label scarcity when abstention and conservative weak supervision are treated as first-class design principles. Third, it provides empirical evidence that unconstrained end-to-end LLM classification systematically violates instrument-based validity constraints by over-classifying thematic or action-oriented discourse as implemented policy measures. The proposed approach is evaluated empirically using a manually annotated gold standard constructed from German municipal web texts. The corpus comprises documents from three Bavarian municipalities, Munich, Bayreuth, and Günzburg, selected to demonstrate methodological feasibility across heterogeneous administrative contexts rather than to enable causal or representative comparison. Under conditions of pronounced class imbalance and limited annotation resources, the evaluation focuses on precision-oriented metrics and interpretable error patterns rather than marginal performance differences, which are known to be unstable in small-sample settings (Chow, 1970; Powers, 2011). The results demonstrate that reliable identification of instrument-based municipal climate policy measures is feasible even with limited gold data, provided that weak supervision and bootstrapping are constrained by high-confidence selection and explicit abstention. The remainder of the study is structured as follows. Chapter 2 develops the instrument-based conceptual foundation and derives implications for empirical measurement. Chapter 3 presents the methodological framework and task operationalisation. Chapter 4 describes the data sources and annotation setup. Chapters 5 and 6 report the evaluation design and empirical findings. Chapter 8 discusses limitations and generalisability, and Chapter 9 concludes.

2. Theoretical Background

This chapter does not merely review instrument-based policy design theory, but derives explicit conceptual constraints for the empirical identification of municipal climate policy measures in text. Building on established policy design literature, it defines policy instruments as the primary unit of analysis and distinguishes them explicitly from policy discourse, strategic goals, and symbolic commitments. These distinctions are translated into measurement-relevant criteria that delimit what can, and cannot, be treated as evidence of policy implementation in unstructured textual data. In doing so, the chapter establishes the normative and analytical boundaries within which automated identification can be conducted without conflating discourse with implementation (Capano & Howlett, 2020; Peters, 2018).

2.1 Municipal Climate Policy and Policy Instruments

Municipal climate policy is operationalised through concrete interventions that translate higher-level political objectives into locally binding or administratively actionable measures. In policy analysis, such interventions are commonly conceptualised as policy instruments, understood as the specific means through which public authorities seek to influence behaviour, allocate resources, or shape socio-technical systems (Hood, 1983; Lascoumes & Le Galès, 2007; Peters, 2018). Policy instruments are analytically distinct from policy goals, strategies, or thematic commitments. While goals articulate desired outcomes and strategies define orientations or priorities, instruments refer to identifiable interventions that can, in principle, be adopted, implemented, funded, enforced, or administered by public authorities. Typical examples include regulatory requirements, financial incentives, informational programmes, and infrastructural investments (Hood, 1983). This distinction is central to empirical policy analysis because it delineates observable policy action from communicative or symbolic expressions of intent. From a measurement perspective, policy instruments constitute the only unit of analysis that can be treated as empirically observable evidence of policy implementation in text. All other policy-related statements lack the evidential specificity required for reliable identification under automated conditions. Importantly, not all theoretically conceivable policy instruments are equally observable or evidentially recoverable in text. Instruments may exist in administrative practice without being specified in publicly accessible documents at a level of detail sufficient for automated identification. The distinction between theoretical existence and textual observability is therefore critical for policy analysis based on unstructured web data and necessitates a restrictive interpretation of what counts as an identifiable policy measure in text.

2.2 Instrument-Based Policy Design

Instrument-based policy design provides an analytical framework that foregrounds the concrete means of policy intervention rather than declared objectives or thematic orientations (Capano & Howlett, 2020; Peters, 2018). This perspective is widely shared in

contemporary policy design research and emphasises that policies can only be meaningfully identified, compared, and evaluated when the instruments through which they operate are specified explicitly (Howlett, 2019). From this viewpoint, references to climate protection, sustainability, or decarbonisation are analytically insufficient unless they are linked to concrete interventions, such as funding schemes, legal requirements, or infrastructure projects (Peters, 2018). For empirical research, instrument-based policy design therefore implies a conservative operationalisation. Only textual instances that refer to specific, actionable instruments and provide explicit textual indications of implementation, allocation, or enforcement can be treated as evidence of policy implementation (Capano & Howlett, 2020). This restriction is not a limitation of the analytical framework but a necessary condition for maintaining conceptual clarity, comparability, and construct validity across heterogeneous policy contexts. By enforcing a strict separation between instruments and other policy-related statements, instrument-based policy design establishes a principled measurement anchor for empirical identification in text-as-data applications.

2.3 Policy Discourse vs. Policy Measures

A central challenge in automated policy analysis is the distinction between policy-related discourse and concrete policy measures. Municipal texts frequently combine strategic visions, progress narratives, and descriptions of specific interventions within the same document (Peters, 2018). This empirical pattern aligns with long-standing insights from implementation research, which emphasise that formal commitments and policy rhetoric often diverge from operational action during the implementation process (Pressman & Wildavsky, 1984; Schmidt, 2008). Policy discourse refers to statements expressing intentions, orientations, or general commitments. While such discourse is politically and communicatively relevant, it does not in itself indicate the existence of a policy instrument. In contrast, policy measures denote specific interventions that are formally adopted or sufficiently specified to allow implementation, funding, or administrative execution (Peters, 2018). From a policy-theoretic perspective, analytical approaches that conflate discourse with instruments systematically overstate policy activity and undermine the validity of empirical claims about policy implementation (Capano & Howlett, 2020; Grimmer & Stewart, 2013). Crucially, the boundary between discourse and measures is not always semantic but evidential. Texts may employ instrument-related terminology, such as references to programmes, regulations, or funding schemes, without documenting an actual decision, allocation, or implementation step. From an empirical standpoint, such cases remain part of policy discourse, even if they resemble instruments in abstract policy design terms. Excluding these instances is therefore not an omission, but a deliberate measurement decision aimed at preventing systematic overestimation of policy implementation in automated text analysis.

2.4 Implications for Measurement and Evaluation

The conceptual distinction between policy discourse and policy instruments has direct implications for empirical measurement in automated policy analysis. Most importantly, it

implies an asymmetric error structure in which FP classifications pose a greater threat to conceptual validity than False Negatives (FNs). False positives introduce interventions that are not empirically substantiated as implemented policy instruments and thereby distort downstream analyses of policy instrument portfolios (Peters, 2018; Powers, 2011). From a measurement perspective, systematically biased measurement compromises inference more severely than incomplete observation of genuine measures. This asymmetry is not a technical artefact of classification but a direct consequence of instrument-based policy design theory. FPs violate the underlying policy concept by attributing implementation where no explicit evidential basis exists, whereas FNs primarily reflect limits of textual observability rather than conceptual error. Prioritising validity over coverage is therefore consistent with established principles of empirical research, which emphasise that biased measurement constitutes a more serious threat than missing data (King & Zeng, 2001). Accordingly, empirical identification strategies should prioritise precision over recall, ensuring that identified instances correspond to genuine policy instruments rather than thematic or rhetorical references. From this perspective, abstention and non-classification are methodologically preferable to speculative inclusion, even at the cost of reduced coverage. This precision-oriented stance provides the conceptual justification for the evaluation logic adopted in the following chapters, without implying claims about completeness or exhaustiveness of identified policy measures.

3. Methodology

This chapter presents a theory-driven methodology for identifying municipal climate policy measures in unstructured web-based text under conditions of severe label scarcity. Building on the instrument-based conception of public policy established in Chapter 2, the approach operationalises policy measures strictly as explicit, instrument-based interventions and distinguishes them from climate-related discourse and goal statements. To enable scalable analysis beyond a small manually annotated dataset, this conceptual foundation is combined with conservative text-as-data methods and abstention-capable weak and semi-supervised learning strategies (Gentzkow et al., 2019; Grimmer & Stewart, 2013; Peters, 2018). Throughout, the methodological design prioritises conceptual validity and precision over coverage. Importantly, a distinction is maintained between the task definition, which is normative and grounded in policy design theory, and the learning problem, which constitutes a pragmatic approximation under severe data constraints. The proposed methodology is explicitly designed to respect this distinction.

3.1 Task Definition and Conceptual Scope

The task consists of identifying concrete municipal climate policy measures in unstructured text. The central challenge is to distinguish actionable, instrument-based interventions from climate-related discourse that lacks an operational mechanism (Grimmer & Stewart, 2013; Peters, 2018). A municipal climate policy measure is defined as a text segment that specifies an instrument-based intervention enacted or administered by a public authority with the intention of influencing climate-related outcomes (Capano & Howlett, 2020; Peters, 2018). This definition requires explicit reference to a policy instrument, understood as an operational mechanism translating political intent into implementation (Hood, 1983; Lascoumes & Le Galès, 2007). In addition, the identification of policy measures requires explicit textual evidence of implementation, allocation, or enforcement, rather than mere instrument terminology. Texts that articulate goals, visions, narratives, or general commitments without specifying such an instrument are excluded. Accordingly, the task is analytically distinct from thematic climate detection, which tends to overestimate policy activity by conflating discourse with intervention (Gentzkow et al., 2019; Grimmer & Stewart, 2013). Formally, the task is framed as a binary classification problem: given a textual unit x, the objective is to predict whether x constitutes a municipal climate policy measure ($y = 1$) or not ($y = 0$). This formulation supports robust evaluation under conditions of severe class imbalance and label scarcity and aligns with methodological guidance for rare-event settings, where validity takes precedence over exhaustive coverage (King & Zeng, 2001).

3.2 Annotation Scheme and Operational Scope

The annotation scheme follows the conceptual distinction between policy goals and policy instruments: goals articulate intent, whereas instruments specify the concrete means by

Not a policy measure	**Policy measure**
Climate-related content without operational instruments	Instrument-based municipal interventions
Strategic visions and long-term goals	Regulatory provisions
Sustainability narratives	Financial incentives
Informational service pages	Binding restrictions
Awareness and participation campaigns	Formal policy programmes
Descriptive project reports	

Table 3.1: Conceptual distinction between climate-related municipal text content and municipal climate policy measures. A text qualifies as a policy measure only if it specifies an actionable policy instrument rather than merely stating goals, narratives, or informational content.

which public authorities influence behaviour or system outcomes (Hood, 1983; Lascoumes & Le Galès, 2007; Peters, 2018). In municipal climate policy, targets such as climate neutrality qualify as policy measures only when they are linked to explicitly specified instruments that imply implementation, enforcement, or resource allocation (Capano & Howlett, 2020; Peters, 2018). Figure 3.1 illustrates this relationship and motivates the instrument-based operationalisation adopted in this study.

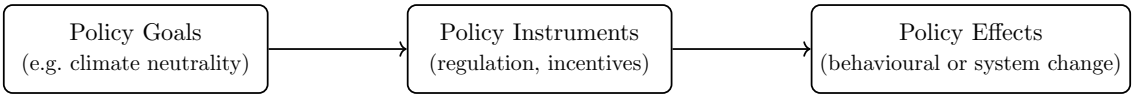

Figure 3.1: Relationship between policy goals, policy instruments, and policy effects. While goals express intent, only instruments constitute operational mechanisms that enable implementation. This distinction forms the basis for the instrument-based definition used for annotation and classification.

To ensure consistency, interpretability, and inter-annotator reliability, the annotation scheme is deliberately restricted to aspects that are directly observable in text. The primary label captures the binary distinction between policy measures and non-measures. Additional attributes are annotated only when explicitly stated, while dimensions that require external administrative data, legal interpretation, or outcome information are excluded by design (Artstein & Poesio, 2008; Gentzkow et al., 2019). Formal inter-annotator agreement statistics are not reported, as the extreme rarity of positive instances and the conservative exclusion of ambiguous cases render such measures uninformative for this task.

3.3 Segmentation and Annotation Strategy

The corpus is segmented into paragraph-level units, which typically correspond to the granularity at which policy instruments are described in municipal web texts. This unit of

Included annotation aspects	Deliberately excluded aspects
Binary identification of policy measures	Budget size or financial volume
Policy instrument type	Legal form and formal enforceability
Policy mechanism	Enforcement strength or sanctions
Target group (if specified)	Temporal duration and implementation phase
Textual specification of instruments	External policy outcomes or impacts

Table 3.2: Scope of the annotation scheme. Included aspects are selected for textual observability and reliability, while excluded aspects typically require additional administrative data or legal interpretation.

analysis provides sufficient local context while limiting contamination from unrelated page sections and aligns with established text-as-data practice (Gentzkow et al., 2019; Grimmer & Stewart, 2013). Annotation follows a two-stage strategy. First, a small, high-precision gold standard is manually annotated according to the instrument-based definition. Ambiguous cases are resolved conservatively to protect label semantics and inter-annotator reliability. Second, the labelled dataset is expanded through abstention-capable weak supervision and confidence-controlled pseudo-labelling. Gold-standard instances are strictly excluded from the unlabeled pool prior to any label expansion to prevent leakage. Abstention is treated as a first-class outcome throughout the annotation and learning process, ensuring that uncertain cases are explicitly excluded rather than forcibly assigned to a class.

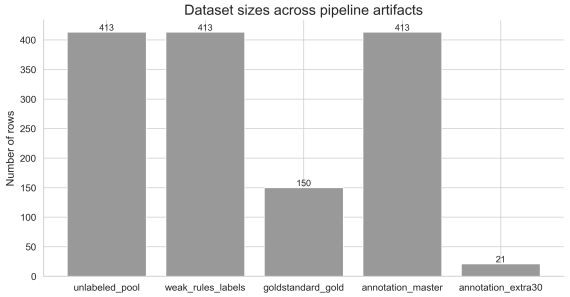

Figure 3.2: Sizes of dataset artefacts across the annotation pipeline. The manually annotated gold standard provides a high-precision anchor but represents only a small fraction of the available municipal text, motivating conservative label expansion.

3.4 Weak Supervision and Conservative Bootstrapping

Weak supervision is employed to expand labels under strict precision constraints. The objective is not maximal coverage, but the identification of a limited set of highly reliable additional training instances (Ratner et al., 2017; Settles, 2009). Two complementary

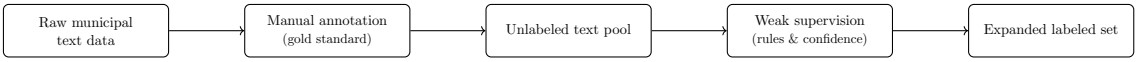

Figure 3.3: Two-stage annotation workflow. Manual annotation produces a high-precision gold standard. The remaining corpus is partially labelled through conservative weak supervision, yielding an expanded training set while maintaining strict separation between verified and inferred labels.

signal sources are used. First, domain-informed heuristic rules detect explicit instrument-related cues and are allowed to abstain when evidence is insufficient. Second, model-based confidence estimates derived from a classifier trained on the gold standard are used to select high-confidence positive and negative instances from the unlabeled pool. Label expansion is deliberately conservative. Only instances exceeding predefined confidence thresholds are added to the expanded training set, while uncertain cases remain unlabeled. Bootstrapping is restricted to a single iteration in order to prevent semantic drift and error amplification under asymmetric error costs. The gold standard remains fixed and isolated throughout and is never modified or augmented. This design reflects the asymmetric conceptual costs of FP classifications in policy identification and follows established findings that iterative self-training under label noise risks systematic error amplification, particularly in imbalanced settings (Peters, 2018; Ratner et al., 2017; Zhu & Goldberg, 2009).

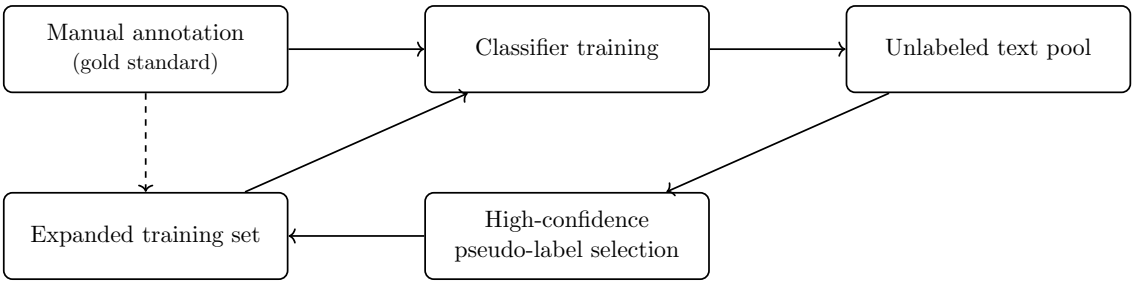

Figure 3.4: Conservative bootstrapping procedure. A classifier trained on the gold standard generates confidence estimates on an unlabeled pool. Only the most reliable instances are pseudo-labelled and added to the expanded training set, while the gold standard remains unchanged and strictly separated, limiting error propagation.

4. Data and Annotation Setup

This chapter documents the construction of the empirical dataset and explicates the data, annotation, and evaluation design choices underlying the study. Rather than merely describing the data pipeline, it motivates these choices as deliberate safeguards for conceptual validity under conditions of severe label scarcity, evidential ambiguity, and asymmetric error costs. In doing so, the chapter translates the theoretical and methodological commitments established in Chapters 2 and 3 into a concrete and auditable empirical setup.

4.1 Data Sources and Collection

The corpus consists of textual content collected from official municipal websites, which represent the primary publicly accessible channel through which municipalities document policies, programmes, regulations, and administrative activities. Focusing on official sources ensures institutional authorship, legal relevance, and a direct link to administrative practice, while excluding secondary reporting or interpretative media content that would confound policy discourse with external commentary. Data collection targets pages containing climate-related references in order to maximise topical coverage while retaining institutional provenance. Crawling and preprocessing procedures systematically remove duplicate content, boilerplate elements, and purely navigational text. No attempt is made to normalise stylistic or structural heterogeneity, as such variation reflects genuine differences in municipal communication practices and constitutes a defining characteristic of the empirical problem. The resulting corpus is therefore intentionally unstructured and heterogeneous with respect to length, format, and linguistic style, providing a realistic testbed for instrument-based policy identification in web-based text.

4.2 Segmentation and Unit of Analysis

Texts are segmented into paragraph-level units, which constitute the basic unit of analysis throughout the study. This granularity reflects how policy instruments are typically specified in municipal web texts and balances two competing requirements. On the one hand, paragraph-level units provide sufficient local context to identify explicit interventions and evidential markers of implementation. On the other hand, they limit contamination from unrelated page sections that would arise under coarser segmentation strategies. The choice of paragraph-level segmentation is therefore not merely technical, but conceptual. It operationalises the evidential criterion derived in Section 3.1 by aligning the unit of analysis with the level at which policy instruments are most plausibly documented in text.

4.3 Annotation Procedure and Gold Standard

As no labelled datasets exist for municipal climate policy measures under an instrument-based definition, a manual annotation process is employed to construct a gold standard. Annotation follows the strict task definition introduced in Section 3.1. A subset

of paragraph-level units is labelled as either a policy measure or a non-measure. Only text segments that provide explicit textual evidence of an implemented or formally specified policy instrument are assigned to the positive class. Ambiguous cases are systematically excluded from the positive class in order to preserve label semantics and conceptual validity. The resulting gold standard is intentionally small and strongly imbalanced. This is not a practical limitation but a direct consequence of the restrictive evidential criteria imposed by the instrument-based definition and the empirical rarity of explicit policy instruments in municipal web texts. Expanding the gold standard through speculative or ambiguous labels would increase apparent coverage at the cost of systematically introducing false positives, thereby undermining the validity of downstream analysis. The gold standard thus serves as a high-precision anchor rather than an exhaustive representation of policy activity. Formal inter-annotator agreement statistics are not reported. Under extreme class imbalance and conservative exclusion of ambiguous cases, such measures provide limited additional information and risk conflating conceptual uncertainty with annotation disagreement. Instead, consistency is ensured through explicit annotation guidelines and strict adherence to the instrument-based definition.

4.4 Dataset Splits and Leakage Prevention

The gold standard is partitioned into disjoint training and test subsets using a fixed hold-out split. The test set is strictly excluded from all stages of model training, weak supervision, post-processing, and bootstrapping. Weakly supervised and pseudo-labelled instances are drawn exclusively from an unlabelled pool that excludes all gold-standard test instances. Bootstrapping is restricted to a single iteration and does not modify or augment the original gold standard. This design choice reflects the asymmetric conceptual costs of classification errors and mitigates the risk of semantic drift and error amplification under label noise. All model configurations are evaluated on the same held-out test set, ensuring comparability across experimental conditions and preserving the integrity of the ground truth.

4.5 Evaluation Setup and Metrics

Evaluation follows a binary classification framework with explicit abstention. Performance is reported using precision, recall, and F1-score (F1), computed exclusively over non-rejected (classified) instances. Let TruePositive(TP), FP, and FN denote true positives, false positives, and false negatives among classified instances. Precision, recall, and F1 are defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Instances for which the model abstains are excluded from these quantities and are instead captured by coverage, defined as the proportion of instances for which a classification decision is produced.

This evaluation logic follows directly from the asymmetric error structure derived in Chapter 2. Precision is treated as the primary criterion of interest, as false positive classifications violate the instrument-based definition by attributing implementation where no explicit evidential basis exists. Recall and F1 are reported for completeness but are interpreted in light of limited positive support and explicit abstention behaviour.

4.6 Reproducibility and Implementation Details

All experiments are implemented using a modular pipeline that separates data processing, annotation, weak supervision, classification, post-processing, and evaluation. This separation enables transparent auditing of each stage and prevents unintended information leakage across components. Hyperparameters, confidence thresholds, and rule specifications are fixed prior to evaluation and applied consistently across all configurations. Preprocessing steps, annotation guidelines, and evaluation scripts are documented, and random seeds are fixed where applicable to support reproducibility. Together, these measures ensure that reported results reflect intrinsic system behaviour under predefined conceptual and methodological constraints rather than artefacts of post-hoc optimisation.

5. Evaluation Design

This chapter specifies the evaluation design used to assess the proposed classification pipeline under conditions of severe label scarcity, pronounced class imbalance, and explicit reject-option decision making. The evaluation framework is not treated as a purely technical component, but as a direct consequence of the instrument-based task definition and the asymmetric error structure derived in Chapter 2. Accordingly, the design prioritises auditability, conceptual validity, and interpretability over exhaustive performance estimation or optimistic generalisation. All evaluation choices are motivated as deliberate safeguards against spurious precision and post-hoc optimisation under small-n conditions (Chow, 1970; Ratner et al., 2017).

5.1 Hold-Out Protocol under Label Scarcity

All quantitative evaluation is conducted exclusively on the manually annotated gold standard. The gold standard is partitioned into disjoint training and test subsets using a fixed hold-out protocol. The test set is strictly excluded from all stages of model training, weak supervision, post-processing, and bootstrapping. This strict separation ensures that reported results reflect generalisation to unseen, manually verified instances rather than artefacts of label leakage or iterative self-training. Under severe label scarcity and extreme class imbalance, a single transparent hold-out split provides a more auditable evaluation basis than repeated resampling procedures. Cross-validation is deliberately avoided, as repeated reuse of the same rare positive instances would create an illusion of robustness without increasing evidential support. Moreover, cross-validation is incompatible with the reject-option design adopted in this study, as thresholded abstention behaviour cannot be meaningfully averaged across folds without reintroducing post-hoc optimisation. Quantitative results are therefore complemented by qualitative inspection of misclassifications to assess error structure and conceptual alignment.

5.2 Metrics and Coverage

System performance is reported using precision, recall, F1, and accuracy computed exclusively on non-rejected (classified) instances. In addition, coverage is reported as the proportion of instances for which the system produces a classification decision. This explicit separation between classification quality and decision availability reflects the reject-option logic of the pipeline and prevents conflation of abstention behaviour with misclassification. Precision is treated as the primary evaluation criterion, as false positive classifications directly violate the instrument-based definition by attributing policy implementation where no explicit evidential basis exists. Recall and F1 are reported for completeness but are interpreted in light of limited positive support and explicit abstention. Accuracy is included only as a secondary reference measure, as it is known to be dominated by the majority class under severe imbalance and provides limited insight into instrument identification performance (Powers, 2011). Aggregate metrics such as ROC or precision–recall curves

are deliberately not employed. Under small-sample conditions, such curves are highly sensitive to single observations and encourage post-hoc operating-point selection. Fixed operating points combined with explicit coverage reporting provide a more interpretable and conceptually aligned evaluation under the constraints of this study.

5.3 Reject-Option Classification and Abstention

The evaluation framework incorporates reject-option behaviour following reject-option decision theory (Chow, 1970). Model outputs are interpreted as confidence-weighted predictions, and instances falling below predefined decision thresholds are rejected rather than forced into binary classifications. Abstention is treated as a valid and informative outcome, reflecting epistemic uncertainty rather than model failure. Decision thresholds are fixed prior to evaluation and applied consistently across all configurations. No post-hoc optimisation of operating points is performed. This restriction ensures that reported performance reflects intrinsic system behaviour under predefined reliability constraints rather than opportunistic threshold tuning. In the context of instrument-based policy identification, abstention serves as a conceptual safeguard: rejecting uncertain cases is preferable to speculative inclusion that would compromise construct validity. The evaluation design therefore operationalises abstention as a first-class component of model performance rather than a residual error category.

6. Results

This chapter reports the empirical results of the proposed pipeline under the evaluation design specified in Chapter 5. All results are computed exclusively on the held-out gold test set. Weakly labelled instances generated during conservative bootstrapping are used solely for training and are excluded from evaluation to preserve the integrity of the ground truth (Ratner et al., 2017). Results are organised along three dimensions: (i) precision stability under conservative label expansion, (ii) robustness under extreme label scarcity and class imbalance, and (iii) interpretability of residual errors relative to the instrument-based definition of municipal climate policy measures.

6.1 Supervised Baseline Performance

A supervised baseline classifier is trained exclusively on the gold training split and evaluated on the held-out gold test set. This configuration provides a lower-bound reference for system behaviour under strict manual supervision, without weak labelling, abstention-aware expansion, or bootstrapping. Given the pronounced class imbalance and limited positive support, aggregate metrics are interpreted cautiously, with confusion-matrix inspection providing the most informative view of system behaviour. Consistent with the asymmetric error structure of the task, precision constitutes the primary evaluative quantity of interest (Capano & Howlett, 2020; Peters, 2018; Powers, 2011).

6.2 Effect of Conservative Bootstrapping

The classifier is retrained using an expanded training set that combines gold labels with a limited number of high-confidence pseudo-labels obtained through conservative bootstrapping. Label expansion is restricted to instances exceeding predefined confidence thresholds, while an explicit abstention region excludes ambiguous cases from forced classification.

6.2.1 Confusion Matrix Comparison

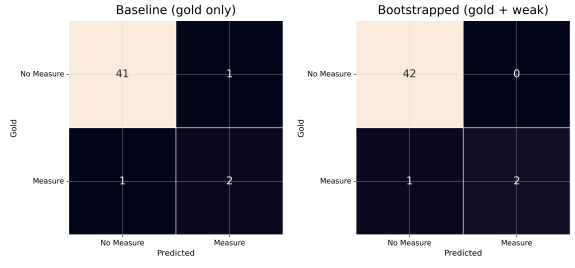

Figure 6.1: Confusion matrices on the gold test set for the supervised baseline (gold only) and the bootstrapped classifier (gold + weak). Conservative bootstrapping preserves the precision-oriented error profile and does not introduce additional FPs.

Figure 6.1 compares confusion matrices for the supervised baseline and the bootstrapped classifier evaluated on the identical gold test set. The central evaluative criterion is precision stability, that is, whether conservative label expansion introduces additional FPs.

The observed pattern indicates that bootstrapping stabilises decision boundaries without eroding precision. Abstention thereby acts as a safeguard against boundary drift rather than a mechanism for maximising coverage (Chow, 1970; Ratner et al., 2017).

6.2.2 Aggregate Metrics under Small Support

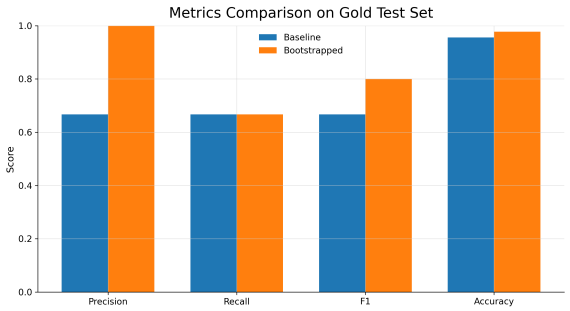

Figure 6.2: Performance metrics for baseline and bootstrapped classifiers on the gold test set. Reported values are computed on classified instances only and should be interpreted in light of limited positive support.

Figure 6.2 reports precision, recall, F1, and accuracy for both configurations. While these values represent point estimates under limited positive support, they indicate that conservative bootstrapping preserves overall performance and does not compromise the intended precision-oriented operating regime.

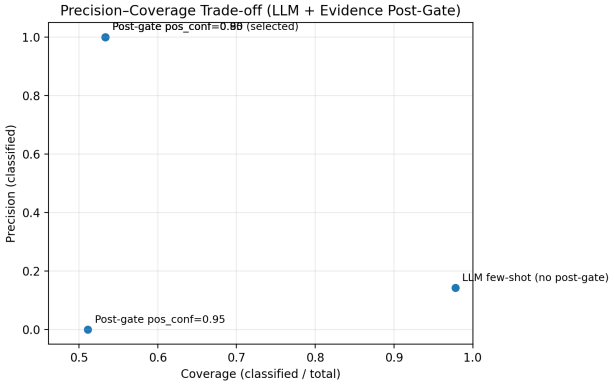

Figure 6.3: Trade-off between precision and coverage under varying decision thresholds. Higher precision is achieved at the cost of reduced coverage, reflecting the explicit reject-option design of the pipeline.

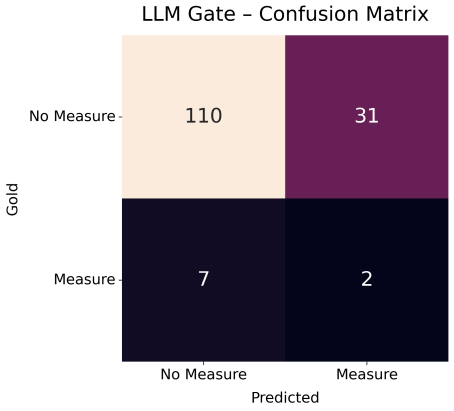

Figure 6.4: Confusion matrix of an unconstrained LLM-based gate on the gold test set. The elevated FP rate reflects systematic over-classification of thematic or rhetorical content as policy measures.

6.2.3 Precision–Coverage Trade-off

Figures 6.3 and 6.5 illustrate the relationship between decision reliability and decision availability under varying thresholds. Higher precision is obtained by accepting lower coverage, consistent with the reject-option design of the pipeline.

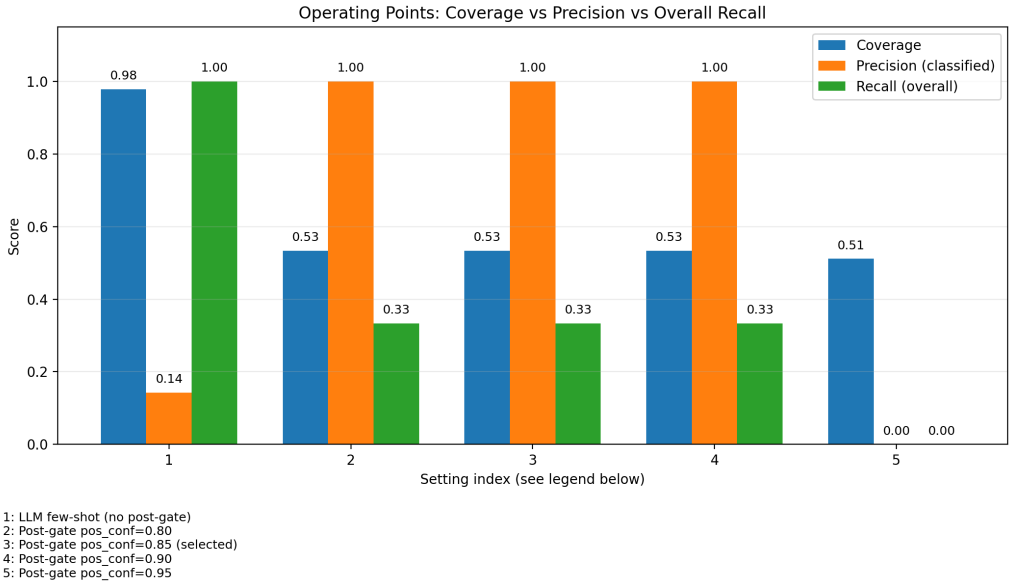

Figure 6.5: Operating points across configurations: coverage, precision (classified instances), and overall recall on the gold test set. The selected post-gate threshold (pos_conf=0.85) reflects the intended high-precision regime under reject-option decision logic.

6.3 Behaviour of an Unconstrained LLM Gate

To contextualise the system design, an unconstrained LLM-only gate is evaluated as a contrastive configuration. This setup directly classifies text segments without rule-based filtering, explicit abstention, or conservative thresholding. As shown in Figure 6.4, the unconstrained gate exhibits a substantially elevated FP rate on the gold test set. These errors predominantly correspond to thematically climate-related or action-oriented segments that lack explicit textual evidence of operational policy instruments, illustrating a characteristic failure mode of single-stage LLM deployment for instrument-based policy identification.

6.4 Qualitative Error Analysis

Quantitative results are complemented by qualitative inspection of misclassified cases to assess alignment with the instrument-based conceptual boundary.

6.4.1 False Positives

FPs primarily arise from action-oriented language in the absence of explicit policy instruments. Recurring patterns include strategic visions without operational mechanisms, descriptive project narratives, and informational communication that signals intent without documenting implementation, allocation, or enforcement. These cases correspond to violations of the instrument criterion and reflect conceptual ambiguity rather than random model error.

6.4.2 False Negatives

FNs most frequently occur when genuine policy instruments are expressed only implicitly, for instance through legalistic, condensed, or cross-referential administrative language. In these cases, the underlying intervention exists but is not recoverable at the paragraph level without additional contextual or inferential information. By contrast, True Negatives (TNs) correspond to text segments that do not describe an implemented policy instrument in the first place and are therefore correctly rejected. The observed FNs are thus an expected consequence of conservative evidential criteria, paragraph-level segmentation, and severe supervision scarcity rather than systematic model failure.

6.4.3 Implications

Residual errors produced by the conservative pipeline predominantly reflect genuine conceptual boundary cases rather than systematic model drift or instability. By contrast, the unconstrained LLM gate highlights the limitations of single-stage prompting for conceptually strict policy identification tasks. Overall, the results demonstrate that the proposed pipeline enables reliable and interpretable identification of municipal climate policy measures under realistic annotation constraints, prioritising conceptual validity and precision over maximal coverage.

7. Related Work

This chapter situates the study within prior research on computational text analysis in public policy, Information Extraction (IE) from administrative texts, weak supervision under label scarcity, and the use of LLMs for IE. The review focuses on methodological approaches relevant to identifying concrete policy interventions and highlights their limitations with respect to conceptual validity, precision, and interpretability in implementation-oriented policy analysis.

7.1 Policy Text Analysis and Text-as-Data

Computational text analysis has become a central methodology for large-scale research on public policy and governance. Foundational work in political science conceptualises legislative, administrative, and policy-related documents as text-as-data, enabling systematic analysis of themes, attention, framing, and discursive dynamics (Gentzkow et al., 2019; Grimmer & Stewart, 2013; Wilkerson & Casas, 2017). These approaches have been widely applied to the study of policy agendas, issue salience, and political communication across institutional contexts. In many applications, policy activity is operationalised indirectly through thematic prevalence, keyword frequencies, or latent topic structures. While such measures are well suited for analysing attention and discourse, they are not designed to distinguish between strategic communication and concrete policy action. As a result, thematic relevance is frequently used as a proxy for implementation. From an instrument-based policy design perspective, this constitutes a substantive limitation, as policy measures are defined in terms of concrete instruments rather than discursive content (Capano & Howlett, 2020; Peters, 2018). Accordingly, text-as-data approaches that rely primarily on topical signals risk systematically overstating policy activity when applied to implementation-oriented research questions.

7.2 Information Extraction from Policy and Administrative Texts

IE aims to transform unstructured text into structured representations such as entities, relations, or events. Early work established rule-based and supervised pipelines for extracting structured information from news and administrative documents (Grishman & Sundheim, 1996). Contemporary IE systems typically rely on supervised learning and are often optimised for recall-oriented extraction under the assumption that relevant instances can be exhaustively annotated. In policy-related applications, IE techniques have been used to identify actors, institutions, and thematic relations in legal and administrative texts. However, the identification of policy measures poses distinct challenges. Policy instruments are abstract, context-dependent, and linguistically heterogeneous, and their presence cannot be inferred reliably from topical relevance alone. Many existing approaches implicitly treat policy-related extraction as a form of topical or semantic classification, thereby assuming that policy action follows from relevance. This assumption

conflicts with instrument-based theories of policy design, which conceptualise implementation in terms of tangible instruments of action rather than semantic proximity (Peters, 2018). Consequently, standard IE approaches risk extracting discursive references to climate policy without reliably identifying instrument-based interventions.

7.3 Weak Supervision under Label Scarcity

Weak supervision addresses limited labelled data by replacing manual annotation with heuristic rules, distant supervision, or noisy labelling functions (Ratner et al., 2017). Data programming frameworks explicitly model noise and dependencies among labelling sources in order to generate probabilistic training labels and enable learning at scale under label scarcity. While weak supervision has been successfully applied to a wide range of text classification and extraction tasks, it presupposes that the target concept is clearly defined and consistently observable in text. In the context of policy measures, the primary challenge lies not only in label scarcity but also in the conceptual ambiguity between policy discourse and policy action. Without an explicit theory-driven task definition, weak supervision risks propagating ambiguous or conceptually invalid labels by conflating discursive references with policy instruments (Capano & Howlett, 2020; Peters, 2018). This limitation motivates conservative labelling strategies that prioritise conceptual validity and evidential certainty over coverage when learning under severe supervision constraints.

7.4 Large Language Models for Information Extraction

LLMs have demonstrated strong performance in zero-shot and few-shot IE tasks (Brown et al., 2020). By reformulating extraction as conditional generation or discriminative scoring, LLMs enable flexible task specification and reduce reliance on task-specific architectures. At the same time, generative extraction introduces challenges related to hallucination, schema violations, and output inconsistency. In policy analysis contexts, such errors directly compromise the validity of claims about policy implementation. Recent work therefore emphasises the controlled integration of foundation models into structured decision pipelines, where LLMs provide probabilistic or semantic signals rather than acting as autonomous end-to-end classifiers (Bommasani et al., 2021). This system-oriented perspective aligns with the requirements of precision, auditability, and conceptual control in implementation-focused policy research.

7.5 Research Gap

Taken together, existing research provides powerful tools for analysing large collections of policy-related text. However, these approaches are not designed to identify concrete, instrument-based policy measures with high conceptual precision. Text-as-data methods primarily capture discourse, IE systems prioritise recall, and weak supervision techniques depend on well-defined target concepts that are often not explicitly operationalised in policy analysis. This study addresses this gap by combining instrument-based policy design theory with a conservative, precision-oriented extraction framework. By explicitly

distinguishing policy discourse from policy measures and embedding LLMs as controlled components within a structured decision pipeline, the proposed approach contributes a methodologically complementary perspective to the automated analysis of municipal climate policy implementation.

8. Discussion and Limitations

This chapter interprets the empirical findings in light of the study's objective to identify instrument-based municipal climate policy measures in heterogeneous municipal web text under conditions of severe label scarcity. Rather than maximising coverage, the proposed pipeline is deliberately designed to prioritise conceptual validity, precision, and auditable decision logic. It integrates instrument-based policy design theory with reject-option classification and conservative weak supervision (Capano & Howlett, 2020; Chow, 1970; Peters, 2018; Ratner et al., 2017).

8.1 Precision as a Conceptual Safeguard

Across all evaluated configurations, the pipeline exhibits the intended precision-oriented error profile. FPs are rare, while recall is deliberately constrained through conservative thresholding and explicit abstention. This behaviour directly reflects the substantive requirements of instrument-based policy analysis, where FP classifications introduce non-instrumental content, such as strategic visions, rhetorical commitments, or thematic narratives, into downstream analyses of policy instruments and portfolios, thereby biasing empirical inference (Capano & Howlett, 2020; Peters, 2018). Qualitative inspection confirms that residual errors concentrate on theoretically expected boundary cases. FNs primarily arise from instruments expressed implicitly through condensed or legalistic administrative language, whereas FPs are associated with action-oriented rhetoric that lacks a clearly specified operational mechanism. The systematic alignment of these error patterns with the goal–instrument distinction indicates that the pipeline operationalises the intended conceptual boundary rather than relying on topical or semantic cues (Lascoumes & Le Galès, 2007; Peters, 2018).

8.2 Weak Supervision as Boundary Stabilisation

In this study, conservative bootstrapping does not function as a recall-oriented augmentation strategy. Instead, it serves to stabilise decision boundaries under extreme supervision scarcity by selectively incorporating high-confidence weak labels while deliberately excluding ambiguous cases from the learning process. This interpretation is consistent with established work on weak supervision and active learning (Ratner et al., 2017; Settles, 2009). Methodologically, the results demonstrate that weak supervision can be aligned with a precision-first regime when combined with explicit abstention and strict separation between trusted evaluation labels and weak training labels. Under these conditions, conservative label expansion is preferable to aggressive coverage gains, as the conceptual cost of FPs outweighs the cost of unresolved cases in instrument-based policy identification (Capano & Howlett, 2020; Peters, 2018).

8.3 Large Language Models: Capability versus Decision Logic

Results from the unconstrained LLM-only gate demonstrate that conceptual conservatism does not emerge automatically from instruction-driven inference. In the absence of ex-

plicit thresholds and abstention mechanisms, the model tends to over-classify semantically salient or action-oriented segments, even when no operational policy instrument is specified. This finding motivates a strict separation between model capability and decision logic. Rather than acting as autonomous classifiers, LLMs are treated as providers of probabilistic signals embedded within a reject-option decision framework and complemented by interpretable weak supervision (Chow, 1970; Ratner et al., 2017). This system-level separation is essential for deploying LLM-based components in policy analysis contexts that require transparency, error control, and construct validity (Bommasani et al., 2021).

8.4 Interpretation under Severe Label Scarcity

Given the small and strongly imbalanced gold standard, absolute metric values must be interpreted with caution. Single misclassifications can substantially affect precision estimates, while limited positive support constrains the stability of aggregate performance measures. Accordingly, the evaluation emphasises error structure and qualitative inspection over marginal differences in aggregate metrics, in line with best practices for imbalanced classification and small-sample evaluation (Powers, 2011). Within these constraints, the most robust empirical findings are that conservative label expansion does not erode precision and that observed failures correspond to theoretically plausible ambiguity classes rather than arbitrary model instability (Capano & Howlett, 2020; Peters, 2018).

8.5 Generalizability and Domain Dependence

The study is situated within the institutional and linguistic context of German municipal web communication. While policy instruments constitute a general analytical concept, their linguistic realisation and administrative framing vary across jurisdictions and governance systems (Capano & Howlett, 2020; Hood, 1983; Lascoumes & Le Galès, 2007). Differences in communicative style, ranging from legalistic to narrative, may induce domain effects even within the same national setting. Accordingly, the results should be interpreted as demonstrating methodological feasibility rather than universal calibration. Transfer to other contexts is likely to require adaptation of prompts, heuristic rules, and decision thresholds, while the underlying design principles, instrument-based operationalisation, explicit abstention, and strict separation between trusted and weak labels, are expected to remain applicable (Chow, 1970; Peters, 2018; Ratner et al., 2017).

8.6 Limitations and Future Research

The binary task formulation deliberately abstracts from richer policy dimensions such as instrument subtypes, target groups, calibration, and discretion. While this restriction is methodologically appropriate under severe label scarcity, future work could extend the annotation scheme once larger and more diverse gold standards become available (Capano & Howlett, 2020; Peters, 2018). Paragraph-level segmentation may fail to capture policy measures distributed across multiple text segments. Document-level aggregation

or structured extraction could improve coverage but would introduce additional complexity, reflecting broader trade-offs in text-as-data research (Gentzkow et al., 2019; Grimmer & Stewart, 2013). Finally, fixed decision thresholds enhance interpretability and reproducibility but do not guarantee calibrated probability estimates. Additional labelled data could enable systematic calibration and uncertainty-aware operating-point selection while preserving reject-option behaviour (Chow, 1970).

8.7 Summary

Overall, the findings support the central claim that conceptually valid identification of municipal climate policy measures is feasible under severe label scarcity when policy design theory is operationalised explicitly and LLM–based signals are embedded within conservative decision logic and weak supervision mechanisms (Capano & Howlett, 2020; Chow, 1970; Peters, 2018; Ratner et al., 2017). The resulting pipeline prioritises precision, interpretability, and methodological transparency, which are essential for empirical policy analysis based on noisy and heterogeneous municipal text sources (Gentzkow et al., 2019; Grimmer & Stewart, 2013).

9. Conclusion

This study addressed the challenge of identifying municipal climate policy measures in unstructured web-based text under conditions of severe label scarcity, strong class imbalance, and strict conceptual constraints. It demonstrated how a theoretically precise notion of public policy, grounded in instrument-based policy design, can be operationalised within a conservative, weakly supervised Machine Learning (ML) pipeline (Capano & Howlett, 2020; Peters, 2018). The central contribution of the study lies in the integration of three mutually reinforcing design principles. First, municipal climate policy measures are defined strictly in terms of operational policy instruments, enforcing a clear analytical separation between actionable interventions and goals, narratives, or thematic discourse (Hood, 1983; Lascoumes & Le Galès, 2007). Second, annotation and learning are organised around a precision-oriented strategy that explicitly incorporates abstention and uncertainty rather than enforcing exhaustive labelling under ambiguity (Chow, 1970; Ratner et al., 2017). Third, LLMs are employed as controlled discriminative components embedded within a structured decision architecture, rather than as autonomous end-to-end classifiers. Empirically, the results show that reliable identification of municipal climate policy measures is feasible even with a very small and highly imbalanced gold standard. Across all evaluated configurations, the pipeline consistently prioritises precision over recall, thereby limiting FP classifications that would otherwise distort downstream analyses of policy instruments and policy portfolios (Capano & Howlett, 2020; Peters, 2018; Powers, 2011). Observed error patterns align closely with theoretically anticipated boundary cases, indicating that remaining errors predominantly arise from genuine conceptual ambiguity rather than arbitrary model instability or noise. From a methodological perspective, the study yields a clear system-level insight. LLMs are most effective for conceptually strict policy analysis tasks when embedded within explicitly constrained decision pipelines. Raw LLM outputs alone are insufficient to enforce instrument-based definitions; instead, conservative thresholding, explicit abstention, and strict separation between manually verified and automatically inferred labels are essential for preserving reliability, interpretability, and auditability (Chow, 1970; Peters, 2018; Ratner et al., 2017). This finding underscores the importance of deliberate system design in high-stakes analytical domains such as public policy research (Bommasani et al., 2021). By deliberately restricting scope with respect to policy dimensionality and document-level aggregation, the study demonstrates methodological feasibility without overfitting conceptual complexity to limited annotation resources. At the same time, the proposed framework provides a robust foundation for future extensions, including multi-dimensional policy classification, uncertainty-aware threshold calibration, document-level reasoning, and cross-municipal or cross-national transfer (Capano & Howlett, 2020). Overall, this work demonstrates that conservative, theory-informed ML approaches can successfully reconcile policy-theoretic rigour with scalable text analysis. The proposed framework offers a methodologically sound pathway for large-scale, empirically grounded analysis of municipal climate policy measures in real-world data environments (Gentzkow et al., 2019; Grimmer & Stewart, 2013).

Bibliography

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, *34*(4), 555–596. https://doi.org/10.1162/coli.2008.34.4.555

Bommasani, R., Hudson, D. A., Adeli, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint*.

Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*.

Capano, G., & Howlett, M. (2020). The knowns and unknowns of policy instrument analysis. *Policy and Society*, *39*(1), 1–21. https://doi.org/10.1080/14494035.2019.1708282

Chow, C. K. (1970). On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, *16*(1), 41–46. https://doi.org/10.1109/TIT.1970.1054417

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, *57*(3), 535–574. https://doi.org/10.1257/jel.20181020

Grimmer, J., & Stewart, B. M. (2013). Text as data. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Grishman, R., & Sundheim, B. (1996). Message understanding conference–6: A brief history. *Proceedings of COLING*.

Hood, C. C. (1983). *The tools of government*. Macmillan.

Howlett, M. (2019). *Designing public policies: Principles and instruments*. Routledge. https://doi.org/10.4324/9781315150951

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, *9*(2), 137–163. https://doi.org/10.1093/oxfordjournals.pan.a004868

Lascoumes, P., & Le Galès, P. (2007). Introduction: Understanding public policy through its instruments. *Governance*, *20*(1), 1–21. https://doi.org/10.1111/j.1468-0491.2007.00342.x

Peters, B. G. (2018). *Policy problems and policy design*. Edward Elgar Publishing.

Powers, D. M. W. (2011). Evaluation: From precision, recall and f-measure to roc. *Journal of Machine Learning Technologies*, *2*(1), 37–63.

Pressman, J. L., & Wildavsky, A. (1984). *Implementation: How great expectations in washington are dashed in oakland* (3rd ed.). University of California Press.

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., & Ré, C. (2017). Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, *11*(3). https://doi.org/10.14778/3157794.3157797

Schmidt, V. A. (2008). Discursive institutionalism: The explanatory power of ideas and discourse. *Annual Review of Political Science*, *11*, 303–326. https://doi.org/10.1146/annurev.polisci.11.060606.135342

Settles, B. (2009). *Active learning literature survey* (tech. rep. No. 1648). University of Wisconsin–Madison.

Wilkerson, J., & Casas, A. (2017). Large-scale computerized text analysis in political science. *Annual Review of Political Science*, *20*, 529–544. https://doi.org/10.1146/annurev-polisci-052615-025542

Zhu, X., & Goldberg, A. B. (2009). *Introduction to semi-supervised learning*. Morgan; Claypool Publishers. https://doi.org/10.2200/S00196ED1V01Y200906AIM006