Enabling Traceable eParticipation in Legislative Processes Using Akoma Ntoso

Simone Vagnoni
CIRSFID-ALMA AI
University of Bologna
Bologna, Italy
simone.vagnoni3@unibo.it

Monica Palmirani
CIRSFID-ALMA AI
University of Bologna
Bologna, Italy
monica.palmirani@unibo.it

***Abstract*—This paper presents a public consultation platform for legislative bills encoded in Akoma Ntoso (AKN) XML that keeps every comment traceable to the cited legal text. Inspired by the *crowdlaw* paradigm—the use of technology to involve citizens in lawmaking—the platform lets citizens anchor feedback to specific passages, propose modifications with a separate rationale field, and search across documents using semantic similarity. The analytics pipeline builds interpretable signals from deterministic aspect extraction, Natural Language Inference (NLI)-based stance inference assembled into Perspectivized Stance Vectors (PSV), thematic clustering via k-means, and Principal Component Analysis (PCA) on comment–anchor difference embeddings. Aggregates retain explicit back-links to comment identifiers and target segments, enabling transparent review without decoupling feedback from the legislative structure. We describe the architecture and data model, detail the NLP pipeline, and report preliminary evaluation on three Italian parliamentary bills (146 comments). Results show that SBERT anchoring substantially outperforms random assignment but is surpassed by a TF-IDF lexical baseline, reflecting high terminological overlap between comments and legal provisions. Domain-general NLI exhibits a strong neutral bias on Italian legal text (24–49% accuracy); a hybrid NLI+LLM pipeline raises accuracy to 74–94% (macro-F1 0.63–0.71).**

***Index Terms*—crowdlaw, e-democracy, e-governance, e-participation, Akoma Ntoso, stance detection, NLI, semantic search, PCA, clustering,**

I. Introduction

Public consultations on draft legislation are an essential instrument for democratic participation, allowing citizens, experts, and stakeholders to provide input before bills become law. However, large-scale consultations often collect hundreds or thousands of submissions that are hard to connect back to specific provisions in the legal text [1], [2]. Analysts tasked with synthesizing feedback face a manual process that risks losing the link between aggregate conclusions and the normative passages they concern.

The term *crowdlaw* describes the practice of using technology to draw on public intelligence and expertise to improve the quality of lawmaking [3]. Noveck identifies four stages where citizen input can add value—problem identification, solution design, drafting, and evaluation—and argues that well-designed crowdlaw processes yield better legislative outputs, not merely greater procedural legitimacy. Although a growing ecosystem of e-participation platforms supports aspects of this vision, foundational reviews [1], [2], [4] and a recent systematic analysis of 116 digital tools [5] confirm a persistent gap: few platforms integrate NLP-based analytics or maintain traceable links between citizen feedback and specific legal provisions. The OECD Guidelines [6] stress that quality consultation requires transparency about how inputs are used.

Our platform addresses these gaps with three design principles. First, it treats the legislative bill as the primary reference: every comment is anchored to an article or sub-article passage in an Akoma Ntoso (AKN) document [7], and every aggregate indicator retains the list of contributing comment identifiers and the target segment. Second, it structures citizen input around three components—a proposed modification, an optional motivation, and a removal flag—providing a consistent foundation for downstream analysis. Third, its analytics pipeline combines deterministic aspect extraction with NLI-based stance inference, thematic clustering, and PCA visualizations, each of which can be inspected, recomputed, or substituted without breaking traceability.

Our work addresses three research questions: **RQ1**: *How accurately can embedding-based similarity route citizen comments to the correct legislative article, and how does this compare to lexical baselines?* **RQ2**: *To what extent can domain-general NLI classify stance in Italian legal consultations, and does an LLM fallback improve results?* **RQ3**: *Does the proposed AKN-based workflow close the traceability and NLP analytics gaps identified in existing e-participation platforms?*

Our contributions are: (i) a traceability-first consultation workflow that preserves links from aggregates to cited passages; (ii) a structured citizen input model combining proposed modification, motivation, and removal flag; (iii) an interpretable analytics pipeline combining deterministic aspect extraction, NLI-based stance inference, and visual summaries; (iv) a feature comparison of the platform against seven existing tools showing it closes the gaps in traceability and NLP analytics (RQ3); and (v) preliminary evaluation on three Italian parliamentary bills comparing anchoring methods and assessing the hybrid NLI+LLM stance pipeline against NLI-only.

II. Related Work

A. E-participation and crowdlaw platforms

E-participation platforms can be grouped by the type of citizen input they support. *Opinion-mapping* tools like Pol.is [8] apply dimensionality reduction to vote matrices to surface opinion clusters and bridging statements; vTaiwan [9] combines Pol.is with structured deliberation phases for legislative consultations. *Participatory-process* frameworks such as Decidim [10], Consul [11], and Your Priorities [12] provide modular environments for proposals, debates, and participatory budgeting; Arana-Catania et al. [13] applied NLP to Consul's Decide Madrid deployment for proposal recommendation and comment summarization, confirming that ML can reduce information overload but stopping short of provision-level anchoring or stance analytics. *Commenting* platforms like RegulationRoom [14] add moderator facilitation to U.S. federal rulemaking, while WikiLegis [15]—developed by the Brazilian Chamber of Deputies' LabHacker—is the closest crowdlaw [3] precedent to our work, enabling citizens to propose amendments to specific articles of draft bills. However, WikiLegis does not parse bills into a machine-readable format, does not apply NLP analytics, and does not produce traceable aggregates.

Across all categories, Shin et al. [5] confirm that among 116 digital tools, NLP capabilities remain rare and accountability features—informing citizens how their input influenced decisions—are largely absent.

B. Legal informatics and argument mining

Akoma Ntoso [7] provides a standardized XML vocabulary for legal documents, including bills, acts, and amendments. Its hierarchical structure (body, article, paragraph) enables machine-readable identification of normative provisions. Palmirani's Smart Legal Order framework [16] extends this vision by emphasizing transparency, traceability, and interoperability of normative data—principles that directly inform our platform's design. In legal NLP, argument mining identifies claims and evidence in judicial and legislative texts [17], while stance classification determines whether a text supports or opposes a given proposition [18]. A recent survey identifies 36 ML tasks—from argument mining to stance detection and summarization—that can enhance deliberative quality in online participation [19]. Lawrence and Reed [20] provide a comprehensive taxonomy of argument mining approaches, which typically focus on structure detection within existing text rather than on anchoring new citizen feedback to specific provisions, which is the focus of our platform.

C. NLP for stance and similarity

Transformer encoders and NLI benchmarks enable semantic similarity and stance inference across heterogeneous text [21], [22]. Sentence-BERT [23] provides efficient sentence embeddings for retrieval and clustering, with multilingual variants supporting cross-lingual applications. Perspectivized Stance Vectors (PSV) [24] formalize stance aggregation over concepts, enabling fine-grained (dis)agreement analysis by mapping each comment to a vector of per-aspect stance values. Laurer et al. [25] show that NLI-based transfer learning substantially reduces annotation requirements for political text classification, though they also note that domain-general NLI models can exhibit label biases on out-of-domain text—a finding consistent with our results on Italian legal consultation data. Vamvas and Sennrich [26] introduce the X-Stance dataset covering German, French, and Italian political comments, demonstrating that multilingual BERT achieves moderate cross-lingual transfer for stance detection, with a performance gap that motivates domain-specific adaptation.

TABLE I
Feature comparison of selected participation platforms. ✓ = supported; ∼ = partial; – = absent or undocumented.

Platform	*AKN parse*	*Art. anchor*	*Stance det.*	*Clustering*	*Traceability*	*Struct. input*	*MR export*
Pol.is	–	–	∼	✓	–	–	–
vTaiwan	–	–	∼	✓	–	–	–
Decidim	–	∼	–	–	✓	✓	∼
Consul	–	✓	–	–	✓	✓	∼
RegulationRoom	–	✓	–	–	–	✓	–
Your Priorities	–	–	–	∼	–	✓	∼
WikiLegis	–	✓	–	–	–	✓	–
Our platform	✓	✓	✓	✓	✓	✓	✓

∼: Pol.is/vTaiwan perform opinion clustering, not stance against legislative text. Decidim anchors at paragraph/proposal level but does not support arbitrary text-span selection within a provision. Consul supports span-level text selection (Collaborative Legislation phase) but lacks NLP analytics and AKN export. Decidim/Consul offer JSON API or CSV export but not annotated AKN export. Your Priorities provides an AI similarity engine for content clustering and an API with key-based data export, but neither feature is publicly documented as a core deliberative mechanism.

D. Positioning

Table I compares our platform with existing participation tools across seven dimensions that reflect the crowdlaw design requirements identified by Noveck [3]: structured connection to legislative text, analytical support for synthesizing input, and traceability for institutional accountability.

III. System Overview

A. Architecture

The platform is organized as a single-page application (SPA) frontend communicating via REST with a FastAPI backend that persists data in SQLite. Figure 1 shows the component structure. The frontend uses a three-panel grid layout: document management on the left, the legal text with article-level rail indicators in the center, and analytics (comments, clusters, PCA scatter plot) on the right. All state is managed client-side in a single JavaScript object, enabling rapid navigation without server round-trips for view changes.

The consultation lifecycle follows a repeatable sequence: (1) upload the AKN bill, which triggers parsing into articles and paragraph-level segments; (2) open the comment window for citizen submissions; (3) recompute analytics on each new

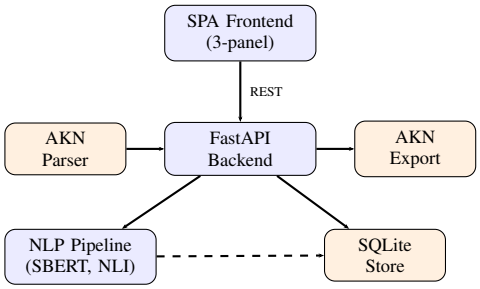

Fig. 1. System architecture. Solid arrows indicate request flow; the dashed arrow indicates NLP write-back to the database.

comment via `process_document()`; (4) optionally trigger PSV computation for concept-level stance aggregation; and (5) export an annotated AKN document embedding all comments as `<note>` elements under `<analysis>`, preserving identifiers and anchors for downstream institutional use.

B. AKN ingestion

During ingestion, the parser extracts all `<article>` elements from the AKN XML using namespace-agnostic traversal (wildcard matching). For each article, the `eId` attribute, `<num>`, and `<heading>` are recorded, and each `<p>` element becomes a separate segment for fine-grained anchoring. The document title is extracted from `<docTitle>` or `<shortTitle>` with a preamble-scanning fallback that filters institutional metadata (chamber names, legislature numbers) to isolate the bill's descriptive title.

C. Data model and traceability

Each comment stores a document id, the user-selected article, the selected text span (if any), the system-resolved anchor (`predicted_article_id`), and provenance metadata including the similarity score, stance label, confidence, and classification method. Aggregations keep the full list of contributing comment ids per segment, enabling direct drill-down from a rail marker to the underlying evidence. Tables II and III summarize the main entities and comment-level metadata.

Traceability is enforced at three levels: (i) anchoring records both the user selection and the system-resolved segment; (ii) aggregate structures store the full list of contributing comment ids; and (iii) exports preserve these identifiers so that external reviewers can reconstruct how any indicator was computed.

D. User interaction

To submit feedback, users select a passage in the center panel and enter a *proposed modification* plus an optional *motivation*. A removal checkbox captures deletion-only feedback. The rail view (Figure 2) encodes net orientation (color) and conflict (texture) per segment with drill-down to the associated comments. The PCA map provides an exploratory overview of semantic distance among comments, with both a global document view and a per-article view.

TABLE II
CORE ENTITIES IN THE DATA MODEL.

Entity	Purpose
Document	Raw AKN XML, title, upload metadata
Article	Article number, heading, full text
Segment	Sub-article text units for anchoring
Comment	Author, timestamp, content, selection, metadata
Aspect	Segment-level concepts (TF–IDF / SBERT)
Cluster	Thematic cluster id and label
PCA	2D projection coordinates per comment

TABLE III
COMMENT METADATA USED FOR TRACEABILITY.

Field	Description
`selection_text`	Quoted passage from the bill (optional)
`selection_kind`	Selection type (`range` or `article`)
`predicted_article_id`	Auto-anchored article when not selected
`similarity`	Cosine similarity score used for routing
`stance`, `stance_conf`	Stance label and NLI/LLM confidence
`stance_method`	Provenance (`nli` or `targeted`)
`cluster_id`, `pca_x/y`	Analytic projections

The platform supports three roles: citizens submit comments and review their contributions; administrators upload documents and manage datasets; analysts use the rails, search, and inspector to synthesize evidence for institutional reporting. Semantic search retrieves the most relevant passages for a query (top-8 per document, top-12 cross-document), giving citizens a direct path to relevant legal text without requiring knowledge of formal article numbering.

IV. METHODS

A. Aspect extraction

Segments are represented with SBERT embeddings (`paraphrase-multilingual-MiniLM-L12-v2`, L2-normalized [23]). Candidate aspects are extracted using TF–IDF over uni/bi-grams [27] with Italian stopwords. Near-duplicate candidates are merged via cosine similarity: any two candidates with similarity ≥ 0.82 are grouped, and the first term in each cluster is retained as canonical. The result is truncated to K aspects per segment (K=6 for PSV computation, K=8 for display). The aspect list is deterministic: given a fixed document, the same concepts are produced on every run. An optional LLM path (temperature 0.0) refines candidates by synonym merging, producing at most 8 canonical aspects per segment.

B. Stance inference and PSV

Stance is computed with respect to the referenced legal text. A multilingual NLI cross-encoder (`mDeBERTa-v3-base-xnli-multilingual-nli-2mil7` [21], [22]) evaluates two hypothesis templates against each comment:

- *Support*: "The author supports the following statement: {target}"

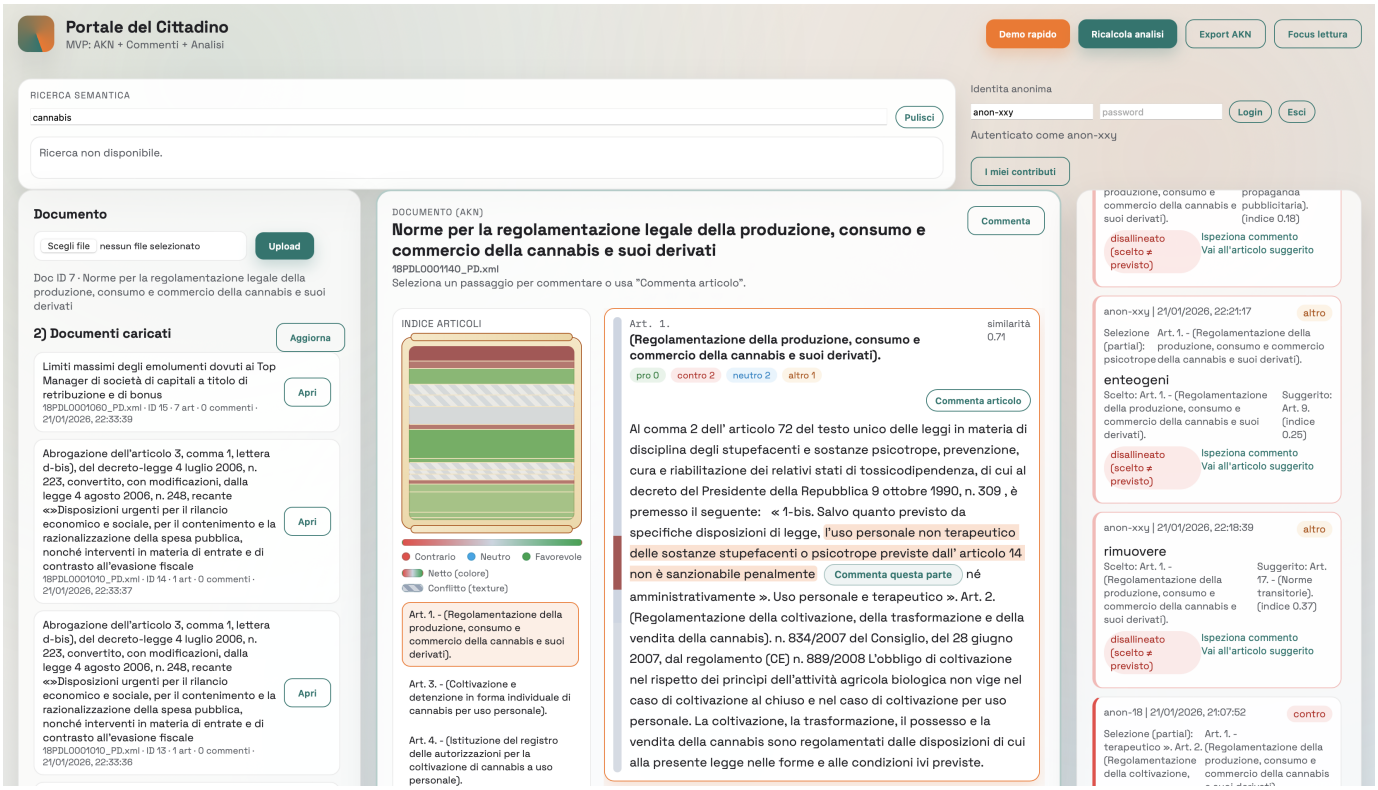

Fig. 2. Platform interface: document navigation (left), legal text with acceptance rails (center), analysis panel with comments, clusters, and PCA scatter plot (right).

- *Oppose*: "The author opposes the following statement: {target}"

where the target is the user's selected passage (truncated to 1 400 characters) or the full article text. Both pairs are scored simultaneously via softmax over entailment, neutral, and contradiction classes. Let s^+ and s^- denote the entailment probabilities for the support and opposition hypotheses, respectively, and let $p_{\mathrm{n}}^+, p_{\mathrm{n}}^-$ denote their corresponding neutral probabilities. The decision rule is:

$$\text{label} = \begin{cases} \textit{pro} & \text{if } \max(s^+, s^-) \geq 0.5 \ \wedge \ s^+ \geq s^- \\ \textit{contra} & \text{if } \max(s^+, s^-) \geq 0.5 \ \wedge \ s^- > s^+ \\ \textit{neutral} & \text{if } \max(p_{\mathrm{n}}^+, p_{\mathrm{n}}^-) \geq 0.5 \\ \textit{other} & \text{otherwise} \end{cases}$$

In the default `auto` mode, if the NLI confidence ($\max(s^+, s^-)$) falls below a configurable threshold (default 0.6), an LLM-based fallback via OpenRouter is invoked (temperature=0.0, max_tokens=120). Each label is stored with its confidence and a provenance flag (`nli` or `targeted`).

The resulting labels are assembled into Perspectivized Stance Vectors (PSV) over the segment's aspects [24]. For a comment c_i and aspect set $\{a_1, \ldots, a_k\}$, the PSV is $\vec{v}_i = [s_{i1}, \ldots, s_{ik}]$ with $s_{ij} \in \{-1, 0, +1\}$, where each entry is derived from NLI stance classification of the comment against the aspect hypothesis. All concept hypotheses for a comment are batched into a single cross-encoder call for efficiency. Segment- and article-level aggregations are then derived from the PSV matrix.

To summarize convergence across PSV vectors, we define a weighted dispersion score. Let $\bar{\vec{v}}$ be the component-wise mean PSV and $w_i = |\{j : s_{ij} \neq 0\}|/k$ the proportion of non-neutral stances in $\vec{v}_i$. The convergence score is:

$$C_w = 1 - \frac{\sum_{i=1}^{N} w_i \|\vec{v}_i - \bar{\vec{v}}\|_1}{2k \sum_{i=1}^{N} w_i} \tag{1}$$

$C_w \in [0, 1]$: values near 1 indicate high agreement, values near 0 indicate maximal dispersion.

C. Clustering and PCA

Comments are clustered in embedding space with k-means (k=round($\sqrt{N}$), bounded to $[2, 6]$; fixed random seed). Cluster labels are the top-5 TF–IDF terms over each cluster's comments. For PCA, difference vectors $\vec{d}_i = \vec{e}_{\text{comment}_i} - \vec{e}_{\text{anchor}_i}$ are computed, where the anchor embedding derives from the user's selected passage or the full article; 2D PCA on the difference matrix projects each comment by its semantic displacement from the cited passage rather than its absolute position in embedding space. Both global and per-article projections report explained variance ratios for interpretability.

TABLE IV
SELECTED CONFIGURATION PARAMETERS.

Parameter	Purpose
`SBERT_MODEL`	Embedding model for similarity and PCA
`NLI_MODEL`	Cross-encoder for stance inference
`STANCE_ENGINE`	Classification mode (`auto`/`nli`/`llm`)
`STANCE_NLI_MIN_CONF`	NLI→LLM fallback threshold (default 0.6)
`PSV_NLI_MIN_CONF`	PSV-level NLI confidence threshold

D. Consensus indicators

We compute two heuristic indicators for each segment:

$$\text{net} = \frac{n_{\text{pro}} - n_{\text{against}}}{n_{\text{pro}} + n_{\text{against}}}, \quad \text{conflict} = \left(1 - \frac{|n_{\text{pro}} - n_{\text{against}}|}{n_{\text{pro}} + n_{\text{against}}}\right) \cdot \frac{n_{\text{pro}} + n_{\text{against}}}{n_{\text{total}}}$$

The *net* score ranges from -1 (unanimous opposition) to $+1$ (unanimous support). The *conflict* score combines balance (how evenly split pro/against are) with engagement (the fraction of non-neutral comments), capturing segments where opinions are both strong and divided. These indicators guide exploration rather than provide a formal measure of collective preference; they are displayed as color (net) and texture (conflict) in the rail visualization.

V. IMPLEMENTATION

The backend is a FastAPI service with SQLite persistence, exposing REST endpoints for document management, comment submission, analytics retrieval, semantic search, and AKN export. The SBERT and NLI models described in Section IV are lazy-loaded as singletons to avoid repeated initialization. Comment and target texts are truncated to 1 200 and 1 400 characters, respectively, to fit within cross-encoder input limits. Runtime parameters are exposed through environment variables (Table IV).

Analytics are recomputed on demand: each call to `process_document()` runs anchoring, stance classification, clustering, and PCA in a single pass. Embedding generation is linear in the number of comments and segments ($O(N+S)$); similarity-based anchoring computes a full cosine matrix ($O(N \cdot S)$) via dot product on L2-normalized vectors, which remains responsive for consultation-scale datasets (hundreds of comments, tens of articles). For larger deployments, approximate nearest-neighbor indices could reduce this cost to $O(N \log S)$.

The AKN export (Section III-A) encodes each comment as a `<note>` element with `eId` and `refersTo` attributes, so downstream tools can resolve back-links without reimporting the database.

Authentication is pseudonymous: no personal data is required, and the system stores pseudonymous credentials, role information, and server-side session tokens for protected operations. The platform enforces three roles (citizen, analyst, admin) via server-side session tokens; role assignment is validated at login and checked on protected endpoints. Because the default models are multilingual, the pipeline can be applied to non-English consultations without architectural changes.

VI. PRELIMINARY EVALUATION

We conducted a preliminary evaluation on three Italian parliamentary bills from Legislature XVIII (2018–2022) to assess anchoring accuracy and stance detection across different legislative domains.

A. Setup

Bill A (*Cannabis regulation*, 17 articles, 74 comments) addresses legalization of cannabis production, commerce, and personal use. **Bill B** (*Executive compensation limits*, 7 articles, 38 comments) regulates salary caps, exit bonuses, and sanctions for top managers. **Bill C** (*Conflict of interest*, 15 articles, 34 comments) addresses incompatibility rules, asset disclosure, and sanctions for public officials. All three sets comprise synthetic comments written by the research team with pre-assigned stance labels (*pro*, *contra*, *neutral*, *other*) designed to exercise the full range of annotation categories, and include both anchored (with text selection) and unanchored submissions. For Bill A, 5 of the 74 comments carry seed-assigned labels from a prior NLI pass rather than manual assignment; these are excluded from the stance evaluation, giving $N_{\text{stance}} = 69$ while anchoring is evaluated on all 74. Evaluation was run under two configurations: `STANCE_ENGINE=nli` (NLI-only baseline) and `STANCE_ENGINE=auto` (NLI with LLM fallback for low-confidence predictions, using DeepSeek-V3 via OpenRouter at threshold $\tau = 0.6$).

B. Anchoring accuracy (RQ1)

Table V reports article-level anchoring accuracy (top-1 match between system-predicted and ground-truth article) alongside three baselines: random assignment, majority-article assignment, and TF-IDF cosine similarity. The random baseline ranges from 4.1% to 18.4%, and the majority baseline reaches at most 18.4%. TF-IDF cosine similarity is a strong lexical baseline, achieving 91.9%, 97.4%, and 82.4% on the three bills.

SBERT embedding similarity achieves 78.4%, 57.9%, and 55.9%—substantially above the random and majority baselines but below TF-IDF. This gap reflects the high lexical overlap between articles and comments in legal text: terms are reused precisely because comments reference specific provisions, so keyword overlap is highly informative. SBERT embeddings capture semantic generalization that is less necessary when exact vocabulary is shared, motivating domain-adapted embeddings as future work.

Anchoring is best when articles cover topically distinct sub-topics (Bill A: 17 articles on cannabis production, use modalities, and regulation) and degrades when provisions share vocabulary (Bill B: 7 articles all concerning remuneration and sanctions). Unanchored comments perform worse in most settings, dropping to 38.5% for Bill B, confirming that user-provided text selections substantially improve routing accuracy.

TABLE V
ANCHORING ACCURACY (ARTICLE-LEVEL, TOP-1) WITH BASELINES.

Bill	N	Random	Majority	TF-IDF	SBERT
A (Cannabis, 17 art.)	74	4.1%	17.6%	91.9%	78.4%
B (Top Mgr., 7 art.)	38	18.4%	18.4%	97.4%	57.9%
C (Confl. Int., 15 art.)	34	8.8%	11.8%	82.4%	55.9%
Anchored subset (with user selection):					
A	72	–	–	–	80.6%
B	25	–	–	–	68.0%
C	22	–	–	–	45.5%
Unanchored subset (no selection):					
B	13	–	–	–	38.5%
C	12	–	–	–	75.0%

TABLE VI
STANCE CLASSIFICATION RESULTS: NLI-ONLY VS. NLI+LLM HYBRID (DEEPSEEK-V3).

	Bill A		Bill B		Bill C	
	Acc	F1	Acc	F1	Acc	F1
NLI-only	49.3%	0.365	26.3%	0.233	23.5%	0.148
NLI+LLM	**73.9%**	**0.633**	**89.5%**	**0.661**	**94.1%**	**0.706**
N	69		38		34	

C. Stance accuracy (RQ2)

Table VI reports stance classification results for both pipeline configurations. **NLI-only:** Bill A achieves 49.3% accuracy (macro-F1: 0.365); Bills B and C score 26.3% and 23.5%—close to the four-class random baseline of 25%. The confusion matrix for Bill A (Table VII) shows that NLI correctly identifies most opposition comments (18/23 *contra*) but classifies no *neutral* comments correctly (0/16) and confuses *pro* with *other* (6 misclassifications). Bills B and C exhibit a stronger neutral bias: the model classifies 50.0% and 73.5% of all comments as *neutral*, versus 18.4% and 20.6% in the ground truth.

NLI+LLM hybrid: Replacing low-confidence NLI predictions ($\tau < 0.6$) with DeepSeek-V3 classifications yields 73.9% accuracy (macro-F1: 0.633) on Bill A, 89.5% (F1: 0.661) on Bill B, and 94.1% (F1: 0.706) on Bill C. On Bill A, 24 of 69 comments triggered the LLM fallback; on Bills B and C, 30 of 38 and 32 of 34 did so, reflecting the lower average NLI confidence on these more argumentative texts. The hybrid pipeline nearly eliminates the neutral-bias failure mode.

The contrast between the two modes is consistent with comment register: Bill A contains direct stance markers (*"sono favorevole"*, *"sono contrario"*), while Bills B and C feature longer argumentative text expressing stance indirectly through legal reasoning, which the domain-general NLI model cannot reliably decode but DeepSeek-V3 handles well.

D. Illustrative example

To illustrate the end-to-end pipeline, consider a comment on Bill A, Article 3 (personal cultivation): *"L'autoproduzione per uso personale deve essere limitata a due piante per evitare abusi."* The system anchors this to Article 3 via embedding similarity ($\cos = 0.71$), classifies it as *pro* (NLI entailment $s^+ = 0.62$), and maps it to the PSV aspect "coltivazione personale" with value $+1$. The comment appears in the PCA scatter near other cultivation-related submissions and contributes to the green (supportive) rail segment for Article 3. An analyst clicking the rail segment can drill down to this specific comment and verify the anchoring, the stance label, and the cited passage.

TABLE VII
CONFUSION MATRIX FOR BILL A, NLI-ONLY MODE (ROWS = GROUND TRUTH, COLS = PREDICTED).

	pro	contra	neutral	other
pro	15	4	2	6
contra	0	18	5	0
neutral	6	4	0	6
other	2	0	0	1

E. Summary

RQ1: SBERT anchoring substantially outperforms random and majority baselines but falls below TF-IDF cosine similarity on all three bills. The gap reflects the high lexical specificity of legal text, where exact terminology strongly predicts article membership. User-provided text selections improve accuracy in most settings.

RQ2: Domain-general NLI alone is insufficient for Italian legal consultations, performing near chance on Bills B and C. The NLI+LLM hybrid (DeepSeek-V3 fallback at $\tau < 0.6$) raises accuracy to 74–94% and macro-F1 to 0.63–0.71 across all bills, eliminating most of the neutral-bias error. These results confirm that a hybrid pipeline is necessary for practical deployment on argumentative legal text.

RQ3: The feature comparison in Table I shows that the proposed platform is the only surveyed tool to combine AKN parsing, article-level anchoring, automated stance detection, thematic clustering, and machine-readable export with full end-to-end traceability, addressing all seven feature dimensions absent in at least one comparable system.

VII. DISCUSSION AND LIMITATIONS

A. Transparency and contestability

Our platform embeds three mechanisms to support transparency in consultation processes, aligning with the OECD Guidelines for Citizen Participation [6]. First, *traceability* ensures that every aggregate indicator retains the full list of contributing comment identifiers and cited passages, enabling reviewers to reconstruct how conclusions were derived. Second, *provenance metadata* records the classification method (`nli` or `targeted`), confidence scores, and similarity metrics for each comment, providing auditors with the information to assess the reliability of automated inferences. Third, *machine-readable export* encodes all anchoring and stance metadata in the exported AKN document—including cited passage, predicted article, stance label and confidence, classification method, cosine similarity, author role, and removal flag—so

downstream tools and independent analysts can verify system outputs without relying solely on the platform's interface.

These design choices align with the transparency and accountability principles in the OECD Guidelines for Citizen Participation [6], which stress that consultation outputs must be traceable to individual contributions and explainable to participants. By making the classification pipeline inspectable and substitutable, our platform enables institutional reviewers and citizen groups to challenge specific stance inferences or anchoring decisions.

B. Risks of algorithmic bias and the neutral default

The preliminary evaluation reveals a critical limitation: domain-general NLI models trained on news and web corpora exhibit a strong neutral bias when applied to Italian legal consultation text, classifying 50–74% of comments as neutral in Bills B and C despite ground-truth neutral rates of 18–21%. If stance detection systematically undercounts expressed support or opposition, aggregate indicators may misrepresent the distribution of citizen views.

This finding aligns with Laurer et al. [25], who show that while NLI-based classifiers are effective for zero-shot political text classification, domain shift degrades performance—a result consistent with the neutral bias we observe on Italian legal text. Replacing low-confidence NLI predictions with DeepSeek-V3 classifications (threshold $\tau = 0.6$) raises accuracy to 74–94% and macro-F1 to 0.633–0.706, demonstrating that the hybrid pipeline substantially corrects the neutral bias. However, the LLM fallback introduces dependencies on proprietary APIs and raises questions about classification consistency across model versions.

Future work should prioritize domain adaptation: fine-tuning stance classifiers on annotated corpora of Italian legal consultations, incorporating multi-annotator agreement measures, and evaluating performance stratified by comment length, register, and argumentation structure.

C. Participation bias and representativeness

Consultation inputs may over-represent particular stakeholder groups—organized interest associations, professional lobbyists, or activists—relative to the broader population. As Fishkin [28] argues, self-selected participation is structurally different from representative deliberation. Aggregates produced by our platform should be read as evidence of *expressed views*, not as a population measure or a binding referendum outcome. The OECD Guidelines emphasize that quality participation processes require proactive outreach, accessibility measures, and transparency about who participated [6]. Our platform supports the latter (traceability and export) but does not address recruitment or outreach, which remain the responsibility of the convening institution.

D. Technical limitations

Anchoring. The argmax rule assigns every comment to exactly one article, which may not capture comments spanning multiple provisions. Top-k anchoring with a confidence threshold would allow multi-article resolution at the cost of increased complexity for the analyst.

Stance detection. The hybrid NLI+LLM strategy mitigates the neutral bias documented above but depends on external API availability and adds latency. Domain-adapted models fine-tuned on Italian legal consultation data could improve accuracy without this dependency.

Scalability. The current implementation computes a full $N \times S$ similarity matrix for anchoring, which scales quadratically. For large-scale consultations, approximate nearest-neighbor indices (FAISS, Annoy) could reduce this to $O(N \log S)$ with negligible accuracy loss.

E. Threats to validity

The evaluation uses synthetic comments with pre-assigned ground-truth labels authored by the research team (146 comments across three bills). Using synthetic rather than real citizen submissions limits generalizability to production settings; stance labels reflect the designers' intent rather than a multi-annotator agreement process. The *other* label captures comments where neither support, opposition, nor neutrality applies, e.g., procedural questions or off-topic remarks.

F. Future work

Near-term priorities include: (i) domain-adapted NLI for legal Italian, fine-tuned on annotated consultation corpora with multi-annotator agreement; (ii) LLM-augmented evaluation with larger datasets and inter-annotator reliability; (iii) qualitative studies with policy analysts on time-to-evidence, perceived transparency, and trust in automated classifications; (iv) top-k anchoring with confidence thresholds; (v) integration of the values/frames layer into the evaluation, mapping segments to EuroVoc controlled vocabulary to enable cross-bill thematic comparison and EU legislative database interoperability; and (vi) deployment on real citizen consultations to assess the traceability model under operational conditions.

VIII. Conclusion

Our platform integrates AKN parsing, passage-level anchoring, and an interpretable NLP pipeline to support transparent public consultation on legislative bills. Every aggregate indicator retains back-links to the contributing comments and cited passages, so institutional reviewers can verify conclusions against the source text. Preliminary evaluation on three Italian parliamentary bills shows (RQ1) that SBERT anchoring substantially outperforms random and majority baselines but is surpassed by TF-IDF on lexically specific legal text, and (RQ2) that domain-general NLI performs near chance on argumentative bills (26% and 24% accuracy) while the NLI+LLM hybrid raises accuracy to 74–94% (macro-F1 0.63–0.71) across all three bills, confirming that a hybrid pipeline is necessary for practical deployment. A feature comparison against seven existing platforms (RQ3) confirms that no surveyed tool combines AKN parsing, article-level anchoring, stance detection, clustering, and traceable export.

By embedding transparency, provenance metadata, and machine-readable export into the consultation workflow, our platform addresses a persistent gap in e-participation platforms: the integration of structured citizen input with legislative text in a way that preserves traceability and supports public contestability [1], [6]. As participatory processes increasingly rely on computational tools for aggregation and synthesis, ensuring that these tools are inspectable, substitutable, and aligned with democratic values becomes essential.

References

[1] Ø. Sæbø, J. Rose, and L. S. Flak, "The shape of eParticipation: Characterizing an emerging research area," *Government Information Quarterly*, vol. 25, no. 3, pp. 400–428, 2008. [Online]. Available: https://doi.org/10.1016/j.giq.2007.04.007

[2] A. Macintosh, "Characterizing E-participation in policy-making," in *Proceedings of the 37th Hawaii International Conference on System Sciences (HICSS)*, 2004. [Online]. Available: https://doi.org/10.1109/HICSS.2004.1265300

[3] B. S. Noveck, "Crowdlaw: Collective intelligence and lawmaking," *Analyse & Kritik*, vol. 40, no. 2, pp. 359–380, 2018. [Online]. Available: https://www.degruyter.com/document/doi/10.1515/auk-2018-0020

[4] R. Medaglia, "eParticipation research: Moving characterization forward (2006–2011)," *Government Information Quarterly*, vol. 29, no. 3, pp. 346–360, 2012. [Online]. Available: https://doi.org/10.1016/j.giq.2012.02.010

[5] B. Shin, J. Floch, M. Rask, P. Baeck, C. Edgar, A. Berditchevskaia, P. Mesure, and M. Branlat, "A systematic analysis of digital tools for citizen participation," *Government Information Quarterly*, vol. 41, no. 3, p. 101954, 2024. [Online]. Available: https://doi.org/10.1016/j.giq.2024.101954

[6] OECD, "OECD guidelines for citizen participation processes," OECD Public Governance Reviews, Tech. Rep., 2022. [Online]. Available: https://doi.org/10.1787/f765caf6-en

[7] M. Palmirani and F. Vitali, "Akoma-ntoso for legal documents," in *Legislative XML for the Semantic Web*. Springer, 2011, pp. 75–100. [Online]. Available: https://link.springer.com/chapter/10.1007/978-94-007-1887-6_6

[8] C. T. Small, M. Bjorkegren, T. Erkkilä, L. Shaw, and C. Megill, "Polis: Scaling deliberation by mapping high dimensional opinion spaces," *Recerca: Revista de Pensament i Anàlisi*, vol. 26, no. 2, 2021. [Online]. Available: https://doi.org/10.6035/recerca.5516

[9] Y.-T. Hsiao, S.-Y. Lin, A. Tang, D. Narayanan, and C. Sarahe, "vTaiwan: An empirical study of open consultation process in Taiwan," 2018, socArXiv preprint. [Online]. Available: https://osf.io/preprints/socarxiv/xyhft/

[10] Decidim Team, "Decidim: A participatory democracy framework," 2019, open-source platform. https://github.com/decidim/decidim. [Online]. Available: https://decidim.org/

[11] Consul Project, "Consul: Open government tool," 2017, open-source platform. https://github.com/consul/consul. [Online]. Available: https://consulproject.org/

[12] R. Bjärnason, "Your priorities: A citizen engagement platform," Citizens Foundation, 2014, https://github.com/CitizensFoundation/your-priorities-app. [Online]. Available: https://www.yrpri.org/

[13] M. Arana-Catania, F.-A. Van Lier, R. Procter, N. Tkachenko, Y. He, A. Zubiaga, and M. Liakata, "Citizen participation and machine learning for a better democracy," *Digital Government: Research and Practice*, vol. 2, no. 3, pp. 27:1–27:22, 2021. [Online]. Available: https://doi.org/10.1145/3452118

[14] C. R. Farina, M. J. Newhart, and J. Heidt, "Rulemaking vs. democracy: Judging and nudging public participation that counts," *Michigan Journal of Environmental & Administrative Law*, vol. 2, no. 1, pp. 123–217, 2012. [Online]. Available: https://repository.law.umich.edu/mjeal/vol2/iss1/3/

[15] C. F. S. d. Faria, *O Parlamento Aberto na Era da Internet: Pode o Povo Colaborar com o Legislativo na Elaboração das Leis?* Brasília: Câmara dos Deputados, Edições Câmara, 2012, iSBN 978-85-402-0003-2. [Online]. Available: https://bd.camara.leg.br/bd/items/9ef328c9-90d6-4f95-aeb7-bf05f9b6d9de

[16] M. Palmirani, "Hybrid AI to support the implementation of the European directive," in *Electronic Government and the Information Systems Perspective (EGOVIS 2022)*, ser. LNCS, vol. 13429. Springer, 2022, pp. 110–122. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-12673-4_8

[17] M. Lippi and P. Torroni, "Argumentation mining: State of the art and emerging trends," *ACM Transactions on Internet Technology*, vol. 16, no. 2, pp. 10:1–10:25, 2016. [Online]. Available: https://dl.acm.org/doi/10.1145/2850417

[18] D. Küçük and F. Can, "Stance detection: A survey," *ACM Computing Surveys*, vol. 53, no. 1, pp. 1–37, 2020. [Online]. Available: https://dl.acm.org/doi/10.1145/3369026

[19] M. Behrendt, S. S. Wagner, C. Weinmann, M. Bormann, M. Warne, and S. Harmeling, "Natural language processing to enhance deliberation in political online discussions: A survey," *arXiv preprint arXiv:2506.02533*, 2025, arXiv:2506.02533. [Online]. Available: https://arxiv.org/abs/2506.02533

[20] J. Lawrence and C. Reed, "Argument mining: A survey," *Computational Linguistics*, vol. 46, no. 4, pp. 765–818, 2020. [Online]. Available: https://direct.mit.edu/coli/article/46/4/765/93768

[21] A. Conneau, R. Rinott, G. Lample, A. Williams, S. R. Bowman, H. Schwenk, and V. Stoyanov, "XNLI: Evaluating cross-lingual sentence representations," in *Proceedings of EMNLP*, 2018, pp. 2475–2485. [Online]. Available: https://aclanthology.org/D18-1269/

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423/

[23] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proceedings of EMNLP-IJCNLP*, 2019, pp. 3982–3992, arXiv:1908.10084. [Online]. Available: https://arxiv.org/abs/1908.10084

[24] M. Plenz, P. Heinisch, J. Gehring, P. Cimiano, and A. Frank, "From argumentation to deliberation: Perspectivized stance vectors for fine-grained (dis)agreement analysis," in *Findings of NAACL*, 2025, arXiv:2502.09644. [Online]. Available: https://arxiv.org/abs/2502.09644

[25] M. Laurer, W. van Atteveldt, A. Casas, and K. Welbers, "Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and BERT-NLI," *Political Analysis*, vol. 32, no. 1, pp. 84–100, 2024. [Online]. Available: https://doi.org/10.1017/pan.2023.20

[26] J. Vamvas and R. Sennrich, "X-stance: A multilingual multi-target dataset for stance detection," in *Proceedings of SwissText & KONVENS*, Zurich, Switzerland, 2020, arXiv:2003.08385. [Online]. Available: https://arxiv.org/abs/2003.08385

[27] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988. [Online]. Available: https://doi.org/10.1016/0306-4573(88)90021-0

[28] J. S. Fishkin, *When the People Speak: Deliberative Democracy and Public Consultation*. Oxford University Press, 2009. [Online]. Available: https://global.oup.com/academic/product/when-the-people-speak-9780199572106