VotIE: Information Extraction from Meeting Minutes

Anonymous ACL submission

Abstract

Municipal meeting minutes record key decisions in local democratic processes. Unlike parliamentary proceedings, which typically adhere to standardized formats, they encode voting outcomes in highly heterogeneous, free-form narrative text that varies widely across municipalities, posing significant challenges for automated extraction. In this paper, we introduce VotIE (Voting Information Extraction), a new information extraction task aimed at span-level extraction of voting-event arguments in narrative deliberative records, and establish the first benchmark for this task using municipal minutes, building on the CitiLink-Minutes corpus. Our experiments yield two key findings. First, under standard in-domain evaluation, fine-tuned encoders, specifically the multilingual XLM-R-CRF, achieve the strongest performance, reaching 93.2% exact-match macro F1, outperforming generative approaches. Second, in a cross-municipality setting that evaluates transfer to unseen administrative contexts, DeBERTa-CRF generalises best among encoders. In this setting, Gemini demonstrates greater robustness, with significantly smaller declines in performance. Despite this generalization advantage, the high computational cost of generative models currently constrains their practicality. As a result, lightweight fine-tuned encoders remain a more practical option for large-scale, real-world deployment. To support reproducible research in administrative NLP, we publicly release our benchmark, trained models, and evaluation framework.

1 Introduction

Natural Language Processing (NLP) has substantially advanced information extraction in governmental and administrative texts. However, most existing work focuses on document types with clear and standardized structures, such as legal documents (Luz de Araujo et al., 2018; Albuquerque et al., 2022) or parliamentary records (Erjavec et al., 2023), where entities and decisions are explicitly marked and consistently formatted. In contrast, municipal council minutes record local decision-making in a highly variable, narrative form, which differs markedly across municipalities. Voting outcomes and policy decisions are often embedded in lengthy procedural descriptions and deliberations, without explicit boundaries or standardized templates. This heterogeneity poses significant challenges for automatic processing and large-scale analysis of local voting behavior (Arana-Catania et al., 2021; Brown and Weber, 2022).

To address these challenges, we introduce VotIE (Voting Information Extraction), a task for extracting voting-event arguments from narrative deliberative records. VotIE formulates extraction as an event-argument identification problem: given a text segment describing a voting procedure, the task is to identify spans corresponding to voting event arguments, including the subject under deliberation, participants, and their respective stances (in favor, against, abstention, or absent), and vote counts. Building on this task formulation, we focus on defining the task and characterizing its difficulty. To this end, we introduce a benchmark built on the CitiLink-Minutes corpus (Campos et al., 2026), and use it to systematically evaluate feature-based models, fine-tuned Transformer encoders, and generative Large Language Models (LLMs), allowing us to characterize task difficulty and generalization behavior. Crucially, our experimental design extends beyond standard in-domain evaluation by explicitly assessing cross-municipality generalization, measuring model robustness to variation in local administrative writing styles.

Our contributions are: (i) We formalize VotIE as a structured event-argument extraction task for deliberative voting procedures and establish the first benchmark with standardized evaluation protocols, including cross-municipality generalization tests that quantify robustness to administrative writing-style variation; (ii) we provide a comprehensive set of baselines spanning both discriminative and generative extraction paradigms; (iii) we publicly release our benchmark splits, a fine-tuned model on all six municipalities[1], training and evaluation

[1] https://huggingface.co/Anonymous3445/XLM-RoBERTa-CRF-VotIE

code [2], and an interactive web demonstration. [3]

2 Related Work

Previous research has explored information extraction from administrative texts, particularly legal documents and parliamentary proceedings, yet the structured extraction of voting events from municipal council minutes has received little attention. In legal corpora, such as UlyssesNER-Br (Albuquerque et al., 2022) and LeNER-Br (Luz de Araujo et al., 2018), methods have primarily targeted general entity types (e.g., persons, organizations, legislation) rather than complex event structures. Feature-based approaches have also been applied to Portuguese municipal minutes for specific entities, such as subsidies (Rodrigues et al., 2010). However, these methods did not target voting events, relying on a narrow subsidy schema without voter roles, voting evidence, or counting expressions, and used traditional, context-independent features rather than modern context-aware embeddings.

Research on parliamentary and meetings corpora has focused on tasks such as topic modeling (Erjavec et al., 2023), summarization (Hu et al., 2023; Nedoluzhko et al., 2022), or decision detection (Carletta et al., 2006). In contrast to both parliamentary records, where votes are structured, and conversational meeting transcripts, where decisions are implicit, municipal council minutes often present formal voting events embedded within narrative prose.

However, this setting has not been explicitly addressed as a structured information extraction task. Sequence labeling approaches for recovering event arguments often offer strong in-domain performance, but may struggle to generalize across municipalities with distinct administrative writing styles, whereas Generative Large Language Models (LLMs), which have demonstrated promise for zero-shot Information Extraction (Wadhwa et al., 2023; Xu et al., 2024), have yet to be systematically applied to administrative records.

3 VotIE Task Definition

VotIE adopts the argument-role vocabulary of Event Extraction (EE) (Doddington et al., 2004; Ebner et al., 2020; Wang et al., 2024), but operationalizes the problem as span-level event-argument identification rather than full event reconstruction. Each voting decision involves role-specific arguments, that jointly define a structured voting event. However, instead of explicitly reconstructing event graphs or argument links, the task focuses on identifying the textual spans corresponding to voting-event arguments. Formally, given a document $\mathbf{X} = (w_1, \ldots, w_n)$, the goal is to extract the set $\mathcal{S} = \{(s_j, e_j, t_j)\}$ of all voting-related spans, where s_j, e_j are character offsets and t_j is the argument type. In contrast to canonical EE benchmarks such as ACE and RAMS, where arguments are often short and lexically localized around discrete event mentions, VotIE arguments tend to be long, procedural, and semantically coupled. Figure 1 illustrates the coupling that motivates the schema: outcome gains procedural meaning only in relation to the accompanying Voter-Abstention and Voter-Against spans that define who falls outside the majority, requiring joint interpretation of procedural voting language rather than recognition of isolated lexical mentions.

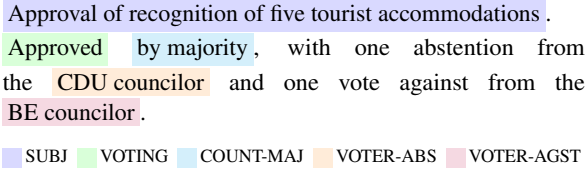

Figure 1: The VotIE task: extracting structured voting event arguments from narrative municipal minutes.

Schema Definition: Our extraction schema $\mathcal{T}$ defines five categories of voting event arguments, comprising twelve argument types. **Subject** captures the matter under deliberation (e.g., "Budget approval"). **Voter Roles** identify participants according to their voting stance, with four subtypes: favor, against, abstention, and absent. **Voting Evidence** marks the decision action (e.g., "approved", "deliberated"), while **Counting Expressions** denote vote aggregation (e.g., "unanimously", "by majority") and explicit numeric tallies typically used in secret ballots (e.g. "3 votes in favor/against/blank"). Finally, **Voting Method** identifies the procedural modality of the vote (e.g., "secret ballot").

4 Experimental Setup

Annotated Dataset: We use the CitiLink-Minutes corpus (Campos et al., 2026), which com-

[2] https://github.com/Anonymous3445/VotIE

[3] https://huggingface.co/spaces/Anonymous3445/VotIE-demo

prises 120 anonymized municipal meeting minutes (2021-2024) from six Portuguese municipalities. Annotation procedure, IAA (Cohen's κ=0.91 on the voting layer), and full corpus statistics are reported in the source paper, and voting layer statistics are reproduced in Appendix B. For the standard benchmark, we used a stratified 70/15/15 chronological split, allocating the oldest documents to the training set and the most recent to the test set, balanced for entity-type representation. For cross-municipality evaluation, we perform leave-one-municipality-out (LOMO) experiments. Splits, prompts (Appendix A), and training/evaluation code are released at our repository.

Extraction Paradigms: We investigate two distinct computational approaches to extract $\mathcal{S}$ from input X. In the **Sequence Labeling** paradigm, we employ standard BIO encoding, mapping input tokens X to a label sequence Y via encoder representations H. The model maximizes $P(Y|X)$ using either a linear classifier or a Conditional Random Field to capture transition dependencies (Lafferty et al., 2001). Spans are reconstructed by aggregating contiguous tags. In the **Generative Extraction** paradigm, the task is formulated as text-to-structure generation, where the model estimates $P(\mathbf{Z}|\mathbf{X})$, and $\mathbf{Z}$ is a structured JSON object. Each predicted span specifies the exact character substring from X and a type drawn from a closed label set. Generated strings are then aligned back to the source X to recover the start and end indices (s, e), using proximity-based disambiguation when multiple occurrences exist. Spans that cannot be exactly re-aligned to $\mathbf{X}$ are discarded rather than treated as predictions. We intentionally require deterministic character-level re-alignment to preserve traceability and avoid heuristic ambiguity introduced by approximate matching

Evaluation Metrics. We report macro-averaged entity-level precision, recall, and F1. Our primary metric uses exact match following CoNLL conventions (Tjong Kim Sang and De Meulder, 2003), which requires both identical boundaries and type agreement. We also compute partial match F1 following SemEval-2013 conventions (Segura-Bedmar et al., 2013), which requires type agreement and boundary overlap, to ensure legal traceability. For discriminative models, BIO predictions are converted to spans and evaluated using seqeval library (Nakayama, 2018), while for generative models, character offsets are recovered via exact substring matching (Wang et al., 2023). Macro averages are taken over the 11 entity types present in the test set, Count-Against appears in the schema but has zero test instances (Appendix B) and is therefore excluded from evaluation.

Training Setup: We evaluate eight architectures across two paradigms, as listed in Appendix C, Table 9. Training follows standard practices for sequence labeling, without any hyperparameter tuning (Souza et al., 2019): AdamW optimization (learning rate 5×10^{-5}), class-weighted cross-entropy loss, maximum 10 epochs with early stopping (patience 3), and 512-tokens sliding windows with 50-token overlap. Invalid BIO transitions are corrected post-decoding. For generative extraction, we test Gemini 2.5 Pro via API (Comanici and et al., 2025), GPT 5.5 via API (OpenAI, 2026), and the AMALIA model (8B) (Simplício et al., 2026), an European Portuguese open-source LLM. All the models were tested in both zero-shot and few-shot (5 examples) configurations, employing schema-constrained decoding and chunking using the langextract library (Goel, 2025), and greedy decoding to ensure deterministic generation. Our goal was not to maximize LLM performance through complex prompting strategies, but to compare paradigms under a controlled extraction setup.

5 Results and Discussion

Standard Benchmark Performance. Table 1 reports performance on the test set, averaged over three seeds for the three encoder-CRF models. The multilingual **XLM-R-CRF** achieves the highest Exact Match Macro F1 with 93.2% $\pm$1.5, ahead of DeBERTa-CRF (91.5$\pm$1.3) and BERTimbau-CRF (90.5$\pm$1.5). A paired segment-level bootstrap (1,000 iterations, pooled over seeds) confirms XLM-R-CRF significantly outperforms BERTimbau-CRF (p=0.018) but not DeBERTa-CRF (p=0.19). The three encoder-CRFs form a single statistical tier on Relaxed Match criteria. The effect of the CRF layer is architecture-dependent: it provides a 8.0-point F1 gain for XLM-R and 13.0 for BERTimbau, whereas for DeBERTa it primarily balances precision and recall. A substantial 32 point gap separates the best encoder from the top-performing generative model, which decreases to 8 points under Relaxed Match criteria. This disparity suggests that while LLMs are effective at identifying relevant semantic regions, they struggle to match the strict token-level boundaries required

for legal traceability. Details regarding the pipeline computational costs can be found in Appendix C.

Table 1: Macro averaged results of all baselines on the standard benchmark. For the three encoder-CRF models, P/R/F1 report the mean over three seeds (42, 13, 123), in %. All other rows report the results from seed 42.

	Exact Match			Relaxed Match		
Model	**P**	**R**	**F1**	**P**	**R**	**F1**
DeBERTa-CRF	89.0	94.8	91.5±1.3	96.0	96.8	96.1±1.2
DeBERTa-Lin.	86.0	94.5	89.8	91.4	98.3	94.6
XLM-R-CRF	**91.8**	**95.0**	**93.2**±1.5	**96.7**	97.5	**97.0**±0.8
XLM-R-Lin.	79.4	94.6	85.2	86.0	**98.9**	91.0
BERTimbau-CRF	88.5	93.1	90.5±1.5	96.6	96.2	96.2±0.8
BERTimbau-Lin.	70.3	91.0	77.5	77.6	97.9	84.9
BiLSTM-CRF	71.7	68.4	69.7	79.3	70.8	74.6
CRF	84.9	79.0	81.0	88.2	81.9	84.1
GPT (5s)	59.7	61.2	60.2	86.5	89.0	87.0
GPT (0s)	53.5	52.8	52.6	86.2	83.9	84.0
AMALIA (5s)	37.7	41.1	34.3	47.3	58.2	46.0
AMALIA (0s)	32.5	23.9	26.5	50.8	41.4	43.3
Gemini-2.5 (5s)	**61.3**	**62.1**	**61.3**	**89.0**	**90.8**	**89.3**
Gemini-2.5 (0s)	58.3	60.1	58.7	86.8	89.3	87.0

Cross-Municipality Generalization.

Municipal meeting minutes overlap in topic coverage but diverge in surface form, which motivates our LOMO protocol. Cross-municipality variation is primarily lexical, with pairwise TF-IDF similarity being about three times higher within than across municipalities (Cohen's d=5.1), while multilingual sentence-BERT embeddings (Reimers and Gurevych, 2019) show a much smaller gap (d=1.0). Table 2 reports macro F1 scores on held-out municipalities. DeBERTa-CRF generalises best among encoders, with a drop of 35% relative to the standard benchmark, and significantly outperforms XLM-R-CRF on M02, M04, and M06 (paired bootstrap, $p < 0.05$). These results highlight a clear trade-off between in-domain performance and generalization. The fine-tuned encoders struggle on unseen writing styles, particularly in M06, suggesting overfitting to specific phrasing patterns present in the training data. In contrast, Gemini exhibits superior cross-municipality robustness, consistently outperforming the encoders on held-out municipalities. This suggests that the LLM relies less heavily on municipality-specific lexical patterns, making it a more reliable choice for deployment in regions with no available training data.

6 Error Analysis

Table 3 categorizes errors into Missing (MIS), Spurious (SPU), Boundary (BND), and Type (TYP).

Table 2: Cross-municipality (LOMO) macro F1 (%). XLM-R/DeBERTa are 3-seed means.

Model	**M01**	**M02**	**M03**	**M04**	**M05**	**M06**	**Mean**
Exact Match							
BERTimbau	56.2	44.4	54.9	46.1	51.8	33.4	47.8
XLM-R	57.8	59.1	61.4	64.5	57.0	28.2	54.7
DeBERTa	57.4	**73.7**	62.2	**70.3**	59.7	31.1	59.1
Gemini	**66.8**	69.0	**69.7**	62.0	**76.0**	**45.4**	**64.8**
Relaxed Match							
BERTimbau	72.8	60.7	64.3	50.3	63.1	42.1	58.9
XLM-R	74.7	71.6	67.0	73.6	69.4	37.5	65.6
DeBERTa	74.6	**79.4**	69.8	77.1	69.1	42.2	68.7
Gemini	**80.5**	77.9	**80.7**	**94.8**	**85.8**	**63.6**	**80.6**

Among encoders, BERTimbau-CRF yields the lowest absolute error count but the lowest macro F1, as it underperforms on rare types; DeBERTa-CRF and XLM-R-CRF make more boundary errors yet achieve higher macro F1. Gemini exhibits the opposite pattern, high segment-level recall but substantial over-generation, resulting in 32% of its errors being spurious. Boundary errors remain a consistent challenge across all architectures, primarily due to the length and syntactic complexity of Subject spans (Appendix D).

Table 3: Error classification analysis in the standard benchmark under exact match criteria.

Model	**Total**	**MIS**	**SPU**	**BND**	**TYP**
BERTimbau-CRF	**186**$_{\pm 8}$	46$_{\pm 8}$	**40**$_{\pm 11}$	**96**$_{\pm 4}$	4$_{\pm 1}$
DeBERTa-CRF	219$_{\pm 11}$	**35**$_{\pm 3}$	47$_{\pm 10}$	133$_{\pm 6}$	3$_{\pm 2}$
XLM-R-CRF	247$_{\pm 16}$	45$_{\pm 4}$	48$_{\pm 16}$	152$_{\pm 3}$	**2**$_{\pm 2}$
Gemini (5s)	1010	222	319	437	32

7 Conclusion

We introduced VotIE, a novel task and benchmark for extracting voting-event arguments from deliberative records, with baselines spanning discriminative and generative paradigms. Two findings stand out. First, fine-tuned encoders remain superior for precise span extraction, with XLM-R-CRF achieving the strongest results (93.2±1.5% F1), significantly outperforming few-shot LLMs. Second, while these encoders perform well in-domain, they degrade on unseen municipalities, whereas Gemini demonstrates superior stability, making it better suited for low-resource settings. Although generative models offer superior adaptability, their high inference costs makes discriminative encoders the more practical choice for large-scale applications where annotated data is available.

8 Limitations

Task Formulation Limitations. Our formulation intentionally prioritises robust event-argument identification as a first benchmarking step towards full deliberative event reconstruction. VotIE is formulated with an event-argument schema, but the benchmark evaluates flat span identification only: we do not score argument-to-trigger binding, role attachment, or event grouping. The research challenge VotIE isolates is the identification of these arguments themselves (long, procedurally embedded, semantically coupled, and stylistically heterogeneous across municipalities), rather than the downstream assembly of correctly identified arguments into events, which becomes a well-defined attachment problem. This separation is benign for the 95.3% of test segments that contain at most one voting trigger, where the attachment of arguments to the single trigger is implicit. For the remaining 4.7% of multi-vote segments (motions paired with amendments, or composite procedural blocks), the models may identify every argument correctly yet cannot disambiguate which Voter span attaches to which Subject span. Therefore, full event assembly lies outside our current evaluation and is the natural next step. We expect Relation Extraction or full Event Extraction paradigms with explicit argument-role binding to be the appropriate formalism, and we leave this, together with the construction of multi-event gold attachments to future work.

Coreference Resolution: Our evaluation focuses on segment-level extraction, treating each segment in isolation, which ignores cross-segment context and, importantly, coreference. In administrative texts, voters are frequently referred to by role (e.g., "the Councilor") or pronominal references rather than proper names, often relying on earlier context or external records. Consequently, while models can detect mentions of voters, they do not resolve these mentions to specific individuals in a knowledge base. Integrating coreference resolution is therefore a necessary next step to support detailed analysis of individual voting behavior.

Dataset limitations: A key observation of our experiments was that the Leave-One-Municipality-Out (LOMO) experiments (Table 2) reveal performance degradation when models encounter new municipalities. Although encoder models remain robust, variation in writing style and lexical choices introduce domain shifts that these models cannot fully overcome. Addressing this may require larger, geographically diverse datasets or unsupervised domain adaptation techniques to improve generalization. Moreover, the secret-ballot minority classes contain a very low number of annotated examples (Count-*, Voting-Method) carry single-digit test support, with Count-Against absent entirely from the test split. Per-type F1 on these categories is therefore noisy at best and unmeasurable at worst. Reliable assessment of secret-ballot extraction requires targeted annotation of additional minutes containing such procedures

Language: Finally, this study focuses exclusively on Portuguese municipal records. While the VotIE schema (Subject, Voter, Stance, Count) is theoretically language-agnostic, the syntactic and stylistic patterns of administrative Portuguese differ significantly from other languages. Further research is needed to assess whether the architectural trends observed here, specifically the superior performance of fine-tuned encoders over LLMs, generalize to other linguistic contexts.

9 Ethical Considerations

Intended Use and Misuse Potential: The primary goal of VotIE is to enhance transparency in local governance by making unstructured data more accessible. However, applying automated information extraction to public administrative records carries a risk of unintended misinformation. As noted in our Error Analysis section 6, neither fine-tuned encoders nor generative LLMs achieve perfect accuracy, with the latter showing a pronounced tendency towards spurious generation. In a real-world civic technology context, extraction errors that misattribute a vote or stance to a political actor could misrepresent their public position, potentially damaging their reputation and distorting public perception. Consequently, we emphasize that systems built on VotIE should be deployed as human-in-the-loop support tools for journalists and civil society organizations, rather than fully autonomous authorities on legislative truth. Their purpose is to assist in navigating voluminous records, not to replace official, legally binding minutes.

Privacy and Data Usage: The dataset used in this work comprises public municipal records, which are, by law, open to the public. However, to mitigate potential harm to private citizens occasionally mentioned in these proceedings (e.g.,

petitioners), we relied on the anonymized version of the CitiLink-Minutes corpus, in which all personal identifiers were manually removed prior to model training.

Usage of AI: In accordance with conference policy, we state that AI assistants were used to enhance clarity and readability, improving language flow and presentation. Moreover, code assistance systems were used to write some subroutines. All scientific claims, experimental designs, and data analyses are the original work of the authors.

A Prompts

For our generative baselines, the models were provided with detailed system instructions defining the taxonomy and annotation rules, followed by five few-shot examples. The prompt and examples used were in Portugues, while the examples below are translated for readability.

A.1 System Instructions and Taxonomy

Below is the system prompt used to define the task boundaries, the 12 defined entity types, and the extraction constraints.

> **You are an expert information extraction system for Portuguese municipal meeting minutes (atas de câmara municipal).**
>
> **Task:** Extract all entity spans from the input text and return a JSON object with an "entities" array.
>
> Each entity must have exactly two fields:
>
> - `text`: the EXACT character sequence from the input (verbatim, do not paraphrase)
> - `type`: one of the 12 allowed entity types listed below
>
> **Entity Type Taxonomy:**
>
> - *Voter roles (named individuals or groups):*
> - `VOTER-FAVOR` — participant who voted in favour (e.g., "o Executivo Municipal", "os eleitos pelo PS")
> - `VOTER-AGAINST` — participant who voted against (e.g., "Vereador do BE")
> - `VOTER-ABSTENTION` — participant who abstained (e.g., "a eleita pelo Nós, Cidadãos")
> - `VOTER-ABSENT` — participant who was absent and did not vote (e.g., "a eleita pelo Nós, Cidadãos")
> - *Voting event:*
> - `VOTING` — verb or verbal phrase that expresses the voting act (e.g., "deliberou", "aprovou", "Aprovada")
> - `SUBJECT` — noun phrase describing what is being voted on; exclude section headings and action verbs
> - `COUNTING-UNANIMITY` — textual expression of a unanimous outcome (e.g., "por unanimidade", "unanimidade")
> - `COUNTING-MAJORITY` — textual expression of a majority outcome, including its breakdown when stated as a single span (e.g., "por maioria", "com 1 voto contra e 3 abstenções")
> - *Secret-ballot aggregate counts (numeric):*
> - `COUNT-FAVOR` — aggregate count of votes in favour (e.g., "6 votos a favor")
> - `COUNT-BLANK` — aggregate count of blank votes (e.g., "7 votos em branco")
> - *Procedure:*
> - `VOTING-METHOD` — expression naming the voting modality (e.g., "escrutínio secreto", "voto secreto")
>
> **Annotation rules:**
>
> 1. **EXACT SPAN:** Extract the verbatim character sequence. Do not add, remove, or alter any characters.
> 2. **SUBJECT extraction:** Extract the noun phrase for the matter being voted on; omit section headings and numeric labels. Strip leading action verbs (e.g., "aprovar", "ratificar") from the extracted span. When the same matter appears in multiple forms, prefer the more specific description.
> 3. **COUNTING vs COUNT-*:** Use `COUNTING-UNANIMITY` / `COUNTING-MAJORITY` for natural-language outcome expressions; use `COUNT-FAVOR` / `COUNT-BLANK` only for explicit aggregate numeric counts (secret ballot).
> 4. **VOTER tagging inside COUNTING-MAJORITY:** Named voters that appear within a majority-breakdown expression are tagged both as `VOTER-*` and as part of the `COUNTING-MAJORITY` span.

A.2 Few-Shot Demonstrations

To enable robust in-context learning, the models were provided with five extraction examples, covering diverse deliberation scenarios. The actual prompt used Portuguese inputs and Portuguese gold spans, while the examples below are translated for readability.

Example 1: Majority with named against and abstention voters

> **Input:** "13. Recognition of IMI and IMT exemption for the buildings whose description appears in the table in Annex I. Approved, with 1 vote against from the Councilor for the BE and 3 abstentions from the Councilors for the PS and the CDU."
>
> **Extracted Entities:**
>
> - `SUBJECT`: "Recognition of IMI and IMT exemption for the buildings"
> - `VOTING`: "Approved"
> - `COUNTING-MAJORITY`: "1 vote against from the Councilor for the BE and 3 abstentions from the Councilors for the PS and the CDU"
> - `VOTER-AGAINST`: "Councilor for the BE"

- VOTER-ABSTENTION: "Councilors for the PS" *and* "the CDU"

Example 2: Unanimity with an absent voter

Input: "4. APPROVAL OF MINUTES. The Chair presented minutes no. 02, dated 18.01.2023, to the meeting. Having considered and analysed the matter, the Municipal Executive resolved unanimously, without the participation of the councilor elected for the Nós, Cidadãos as she was not present, to approve minutes no. 02, dated 18.01.2023."

Extracted Entities:

- VOTER-FAVOR: "the Municipal Executive"
- VOTING: "resolved"
- COUNTING-UNANIMITY: "unanimously"
- VOTER-ABSENT: "the councilor elected for the Nós, Cidadãos"
- SUBJECT: "minutes no. 02, dated 18.01.2023"

Example 3: Secret ballot with aggregate counts

Input: "14. Designation of the Chair of the Board of Directors of the Porto Digital Association. By secret ballot, approved, with 6 votes in favor and 7 blank votes."

Extracted Entities:

- SUBJECT: "Designation of the Chair of the Board of Directors of the Porto Digital Association"
- VOTING-METHOD: "secret ballot"
- VOTING: "approved"
- COUNT-FAVOR: "6 votes in favor"
- COUNT-BLANK: "7 blank votes"

Example 4: Multi-vote segment with distinct voting events

Input: "Acceptance of the Donation of 74 'Debuxo' drawings. The Council resolved to accept the donation of 74 'Debuxo' drawings, made by Micael Matias. It further resolved to approve and execute the corresponding deed of donation."

Extracted Entities:

- **Event 1:**
  - SUBJECT: "Acceptance of the Donation of 74 'Debuxo' drawings"
  - VOTER-FAVOR: "The Council"
  - VOTING: "resolved"
- **Event 2:**
  - VOTING: "resolved"
  - SUBJECT: "deed of donation"

Example 5: Non-voting agenda item (empty extraction)

Input: "5.2. DEPARTMENT OF FINANCE AND ADMINISTRATIVE MODERNIZATION. No items scheduled under this point."

Extracted Entities: *None (empty list returned)*

B Dataset Statistics

The CitiLink-Minutes dataset (Campos et al., 2026) provides the annotation layer used in this work. It comprises council meeting minutes from six Portuguese municipalities (Alandroal, Campo Maior, Covilhã, Fundão, Guimarães, and Porto) spanning multiple administrative years. From this corpus, 2,885 segments were annotated, totalling 10,694 annotated spans across 12 entity types. A subset of 2,516 segments (87.2%) contains at least one voting event, and the remaining 369 are kept as negative examples. All sensitive information was manually anonymised by the dataset authors. The full dataset is publicly available online [4].

B.1 Corpus Overview

	Train	Dev	Test
Segments	1,803	553	529
w/ voting event	1,574	481	461
secret ballot	17	5	5
Total spans	6,753	2,059	1,882

Table 4: Corpus overview per split (before oversampling).

B.2 Entity Type Distribution

Table 5 reports entity counts before oversampling. The distribution is heavily skewed: Voting and Subject together account for approximately 51% of all spans, while secret-ballot types Count-* and Voting-Method collectively represent under 1%.

Entity type	Train	Dev	Test	Total
VOTING	1,749	524	489	2,762
SUBJECT	1,730	523	486	2,739
VOTER-FAVOR	1,085	336	319	1,740
COUNTING-UNANIMITY	1,008	313	326	1,647
VOTER-ABSTENTION	747	221	135	1,103
COUNTING-MAJORITY	162	57	59	278
VOTER-AGAINST	99	60	36	195
VOTER-ABSENT	61	23	18	102
VOTING-METHOD	39	1	5	45
COUNT-BLANK	39	1	5	45
COUNT-FAVOR	32	0	4	36
COUNT-AGAINST	2	0	0	2
Total	**6,753**	**2,059**	**1,882**	**10,694**

Table 5: Entity type counts per split (before oversampling). COUNT-AGAINST is part of the annotation schema but has no instances in the test set and is therefore excluded from all macro-averaged metrics reported in the paper.

[4] https://doi.org/10.25747/7KG6-1K22

B.3 Per-Municipality Distribution

The corpus spans six Portuguese municipalities. Table 6 reports segment counts per split. M03 (Covilhã) is the largest contributor (24.9% of all segments), while M04 (Fundão) is the smallest (8.1%).

ID	Municipality	Train	Dev	Test	Total
M03	Covilhã	473	144	103	720
M05	Guimarães	351	103	102	556
M01	Alandroal	315	96	93	504
M06	Porto	272	93	107	472
M02	Campo Maior	245	79	74	398
M04	Fundão	147	38	50	235
Total		1,803	553	529	2,885

Table 6: Segment counts per municipality and split. Splits were stratified to maintain approximately equal municipality representation in each fold.

B.4 Segment and Span Length Statistics

Segments vary substantially in length. Short procedural openings (5–20 words) coexist with extended deliberative passages of several thousand words within he same document. Table 7 summarises segment-level token statistics. The heavy right tail motivates the 512-token sliding-window strategy with 50-token overlap used by all transformer baselines.

	N	Mean	Median	Std	Range
Segments	2,885	276	169	454	5–8,489

Table 7: Segment-length statistics, in words, across the full corpus.

Span lengths are similarly heterogeneous across entity types (Table 8). Subject spans are the lengthier among the extracted classes, whereas types such as VOTING are typically a single token.

Type	N	Mean	Med	Std	Range
VOTING	2,762	1.0	1	0.0	1–1
SUBJECT	2,739	15.8	15	8.6	1–80
V-FAV	1,740	2.6	2	0.9	1–11
C-UNA	1,647	2.0	2	0.3	1–5
V-ABST	1,103	4.7	4	1.9	2–8
C-MAJ	278	2.1	2	1.2	1–20
V-AGN	195	3.3	3	1.5	1–8
V-ABS	102	3.1	3	1.3	1–8
V-MTH	45	2.0	2	0.0	2–2
C-BNK	45	3.9	4	0.4	3–5
C-FAV	36	4.0	4	0.0	4–4
C-AGN	2	3.0	3	0.0	3–3
All	10,694	5.7	2	7.5	1–80

Table 8: Span-length statistics (in words) per entity type.

B.5 Training Oversampling

To address class imbalance, the training split applies segment-level oversampling. Segments containing at least one span of type Count-Blank, Count-Favor, or Voting-Method are duplicated four times. This was applied identically across all encoder configurations. Development and test splits are never oversampled and bootstrap is computed on the un-oversampled test set

C Model Architectures and Computational Budget

Table 9 summarises the full baseline lineup across two paradigms. The discriminative side spans a feature-based CRF, a BiLSTM-CRF, and three Transformer encoders (BERTimbau-Large, mDeBERTa-v3, XLM-RoBERTa), each with a linear and a CRF head. The generative paradigm comprises Gemini 2.5 Pro, GPT 5.5, and the open-weights AMALIA 8B, all run via langextract in zero- and few-shot settings.

All transformer encoder models were trained and evaluated on NVIDIA A100-SXM4-40GB GPUs on the Deucalion EuroHPC supercomputer. The feature-based CRF runs on CPU. Table 10 lists mean full-pipeline (train + predict + evaluate) times per run, measured on A100-40GB for the encoder models and on NVIDIA L40-48GB for the AMALIA inference pipeline. Encoder training across the standard benchmark (three seeds for the encoder-CRF models, single seed for the remaining configurations) and the six-municipality LOMO evaluation totals 47.3 GPU-hours. Local inference with the open-weights AMALIA 8B model adds a further 20.9 GPU-hours (1,058 documents, mean 71 s/document), comparable to the entire encoder-

training budget and a concrete example that generative extraction remains computationally costly even with open models. Gemini 2.5 Pro and GPT 5.5 were accessed via API and no GPU costs were measured.

Table 9: Model architectures.

Model	Architecture details
CRF	Linear-chain CRF + token/domain features (Lafferty et al., 2001)
BiLSTM-CRF	FastText + CharCNN + CRF layer (Huang et al., 2015)
BERTimbau-Large	Encoder + Linear/CRF head (Souza et al., 2020)
mDeBERTa-v3	Encoder + Linear/CRF head (He et al., 2021)
XLM-RoBERTa	Encoder + Linear/CRF head (Conneau et al., 2019)
Gemini 2.5 Pro	API + Langextract (Comanici and et al., 2025)
GPT 5.5 (OpenAI, 2026)	API + Langextract
AMALIA 8B	Ran locally + Langextract (Simplício et al., 2026)

Table 10: Mean running times on A100-40GB. Encoder rows are full-pipeline time per run; AMALIA rows are per-document inference over the 529-document test set. Aggregate time is measured over all discriminative runs (56 runs).

Configuration	Time
CRF (feat.-based, CPU)	17 min
BiLSTM-CRF (∼2M)	52 min
BERTimbau-Linear (334M)	41 min
BERTimbau-CRF (334M)	99 min
mDeBERTa-Linear (278M)	38 min
mDeBERTa-CRF (278M)	50 min
XLM-R-Linear (278M)	25 min
XLM-R-CRF (278M)	41 min
AMALIA 8B, zero-shot (per doc)	52.9 s
AMALIA 8B, few-shot (per doc)	89.2 s
Encoder training (standard + LOMO)	47.3 GPU-h
AMALIA 8B inference (total)	20.9 GPU-h

D Detailed Per-Entity-Type Results

Table 11 reports entity-level F1 broken down by type for all models on the standard benchmark test set. SUBJECT is consistently the hardest entity. Minority voter-role types (VOTER-AGAINST, VOTER-ABSENT) and the three secret-ballot types (VOTING-METHOD, COUNT-FAVOR, COUNT-BLANK) show higher variance across models due to their low support.

Table 11: Per-entity-type macro F1 (%) on the standard benchmark test set under exact match criteria. For the three encoder-CRF rows, values are the mean over three seeds (42, 13, 123). Support n is the number of gold spans per type in the test set. Bold marks the best score per column within each paradigm. V-Fav/Agn/Abst/Abs = Voter-Favor/Against/Abstention/Absent; C-Una = Counting-Unanimity; C-Maj = Counting-Majority; V-Mth = Voting-Method; C-Fav/Bnk = Count-Favor/Blank.

	Entity-level F1 (%)										
Model	VOTING n=489	SUBJ n=486	V-FAV n=318	V-AGN n=36	V-ABST n=135	V-ABS n=18	C-UNA n=326	C-MAJ n=59	V-MTH n=5	C-FAV n=4	C-BNK n=5
DeBERTa-CRF	**99.3**	74.5	**98.2**	**99.5**	97.8	84.9	99.1	98.3	93.3	83.5	**77.6**
DeBERTa-Lin.	99.1	74.2	97.4	91.1	98.5	73.2	99.4	93.5	100.0	88.9	72.7
XLM-R-CRF	99.0	**77.4**	96.6	97.7	**98.4**	**90.8**	**99.5**	**98.9**	**100.0**	**96.3**	71.1
XLM-R-Lin.	98.3	72.6	96.0	87.8	98.5	47.6	98.8	95.9	100.0	80.0	61.5
BERTimbau-CRF	**99.3**	74.8	98.1	94.7	98.1	77.3	99.0	97.0	90.9	90.9	75.2
BERTimbau-Lin.	98.5	54.6	83.3	54.5	91.4	28.1	98.8	97.5	100.0	72.7	72.7
BiLSTM-CRF	98.8	40.7	92.7	0.0	69.1	0.0	96.0	92.0	88.9	100.0	88.9
CRF	99.5	63.0	95.7	26.9	85.2	42.9	98.5	99.1	100.0	100.0	80.0
GPT (5s)	86.0	**31.3**	85.6	25.0	22.5	12.5	97.1	13.4	**100.0**	**88.9**	**100.0**
GPT (0s)	76.4	17.0	57.0	9.0	9.7	12.9	93.5	14.0	**100.0**	**88.9**	**100.0**
Amalia (5s)	27.0	17.0	30.5	21.8	23.1	5.2	80.3	**25.6**	57.1	40.0	50.0
Amalia (0s)	13.9	7.7	3.3	**30.3**	23.0	17.4	70.3	24.5	0.0	57.1	44.4
Gemini (5s)	**87.0**	18.7	**90.2**	27.3	45.1	**25.0**	**97.7**	14.9	80.0	**88.9**	**100.0**
Gemini (0s)	75.6	20.7	80.4	**30.3**	**52.2**	6.3	96.1	14.8	80.0	**88.9**	**100.0**

References

Hidelberg O. Albuquerque, Rosimeire Costa, Gabriel Silvestre, Ellen Souza, Nádia F. F. da Silva, Douglas Vitório, Gyovana Moriyama, Lucas Martins, Luiza Soezima, Augusto Nunes, Felipe Siqueira, João P. Tarrega, Joao V. Beinotti, Marcio Dias, Matheus Silva, Miguel Gardini, Vinicius Silva, André C. P. L. F. de Carvalho, and Adriano L. I. Oliveira. 2022. UlyssesNER-Br: A Corpus of Brazilian Legislative Documents for Named Entity Recognition. In *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings*, page 3–14, Berlin, Heidelberg. Springer-Verlag.

Miguel Arana-Catania, Felix-Anselm Van Lier, Rob Procter, Nataliya Tkachenko, Yulan He, Arkaitz Zubiaga, and Maria Liakata. 2021. Citizen participation and machine learning for a better democracy. *Digit. Gov.: Res. Pract.*, 2(3).

Eva Maxfield Brown and Nicholas Weber. 2022. Councils in action: Automating the curation of municipal governance data for research. *Proceedings of the Association for Information Science and Technology*, 59:23–31.

Ricardo Campos, Ana Filipa Pacheco, Ana Luísa Fernandes, Inês Cantante, Rute Rebouças, Luís Filipe Cunha, José Isidro, José Evans, Miguel Marques, Rodrigo Batista, Evelin Amorim, Alípio Jorge, Nuno Guimarães, Sérgio Nunes, António Leal, and Purificação Silvano. 2026. Citilink-minutes: A multilayer annotated dataset of municipal meeting minutes. In *Advances in Information Retrieval*, pages 511–527, Cham. Springer Nature Switzerland.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2006. The AMI Meeting Corpus: A Pre-announcement. In *Machine Learning for Multimodal Interaction*, pages 28–39, Berlin, Heidelberg. Springer Berlin Heidelberg.

Gheorghe Comanici and et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2020. Multi-sentence argument linking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8057–8077, Online. Association for Computational Linguistics.

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór

Steingrímsson, Çağrı Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, and 9 others. 2023. The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*, 57(1):415–448.

Akshay Goel. 2025. LangExtract.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *Preprint*, arXiv:2111.09543.

Yebowen Hu, Timothy Ganter, Hanieh Deilamsalehy, Franck Dernoncourt, Hassan Foroosh, and Fei Liu. 2023. MeetingBank: A benchmark dataset for meeting summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16409–16423, Toronto, Canada. Association for Computational Linguistics.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. *Preprint*, arXiv:1508.01991.

John D. Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning*.

Pedro Henrique Luz de Araujo, Teófilo E. de Campos, Renato R. R. de Oliveira, Matheus Stauffer, Samuel Couto, and Paulo Bermejo. 2018. LeNER-Br: A Dataset for Named Entity Recognition in Brazilian Legal Text. In *Computational Processing of the Portuguese Language*, pages 313–323, Cham. Springer International Publishing.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Anna Nedoluzhko, Muskaan Singh, Marie Hledíková, Tirthankar Ghosal, and Ondřej Bojar. 2022. ELITR minuting corpus: A novel dataset for automatic minuting from multi-party meetings in English and Czech. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3174–3182, Marseille, France. European Language Resources Association.

OpenAI. 2026. GPT-5.5. Large language model, accessed via API. Model accessed via IAEDU API, on 2026-05-19.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Mário Rodrigues, Gonçalo Dias, and António Teixeira. 2010. Knowledge extraction from minutes of portuguese municipalities meetings. pages 51–54.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2013. SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, Atlanta, Georgia, USA. Association for Computational Linguistics.

Afonso Simplício, Gonçalo Vinagre, Miguel Moura Ramos, Diogo Tavares, Rafael Ferreira, Giuseppe Attanasio, Duarte M. Alves, Inês Calvo, Inês Vieira, Rui Guerra, James Furtado, Beatriz Canaverde, Iago Paulo, Vasco Ramos, Diogo Glória-Silva, Miguel Faria, Marcos Treviso, Daniel Gomes, Pedro Gomes, and 3 others. 2026. AMALIA: A fully open large language model for European Portuguese. In *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 380–391, Salvador, Brazil. Association for Computational Linguistics.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2019. Portuguese named entity recognition using bert-crf. *arXiv (Cornell University)*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Somin Wadhwa, Silvio Amir, and Byron Wallace. 2023. Revisiting relation extraction in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589, Toronto, Canada. Association for Computational Linguistics.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. Gpt-ner: Named entity recognition via large language models. *Preprint*, arXiv:2304.10428.

Xiaozhi Wang, Hao Peng, Yong Guan, Kaisheng Zeng, Jianhui Chen, Lei Hou, Xu Han, Yankai Lin, Zhiyuan Liu, Ruobing Xie, Jie Zhou, and Juanzi Li. 2024. MAVEN-ARG: Completing the puzzle of all-in-one event understanding dataset with event argument annotation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4072–4091, Bangkok, Thailand. Association for Computational Linguistics.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. 2024. Large language models for generative information extraction: a survey. 18(6):186357.