Posts Talk Policy, Stories Don't: Policy-Issue Detection on Instagram with Fine-Tuned Transformers and Prompted LLMs

Michael Achmann-Denkler[1], **Mario Haim**[2], **Christian Wolff**[1]
[1]Universität Regensburg
Regensburg, Germany
{michael.achmann, christian.wolff}@ur.de
[2]Ludwig-Maximilians-Universität
München, Germany
haim@ifkw.lmu.de

Abstract

Policy issues are central to election campaigns, yet systematic analyses of issue communication on Instagram remain scarce—particularly for ephemeral Stories. We develop and evaluate automated methods for detecting the binary presence of policy issues in Instagram posts and Stories from the 2021 German federal election. Drawing on a gold-standard dataset of 1,357 annotated documents across three textual channels (captions, OCR-extracted image text, and speech transcripts), we compare a fine-tuned German transformer (GBERT) with multiple LLM prompting strategies (zero-shot, few-shot, retrieval-augmented). Both approaches prove effective: GBERT achieves a cross-validated macro F_1 of 0.90, closely matched by GPT-o3 under few-shot prompting (0.88). Substantively, policy visibility varies far more by content format than by party: 70% of posts contain policy references compared to only 17% of Stories, a pattern that holds consistently across all eight parties. An exploratory topic model confirms that parties reproduce familiar issue-ownership profiles within the subset of policy-relevant texts. Our results establish binary issue detection as a feasible foundation for studying policy communication in multimodal, ephemeral social media environments.

Keywords: policy issues, Instagram, political communication, text classification, LLMs, German federal election

1. Introduction

Issue ownership theory holds that campaigns help set the criteria of electoral choice by strategically foregrounding issues for which parties enjoy reputational advantages (Petrocik, 1996). Yet systematic analyses of policy-issue communication on social media remain limited: research on political actors' social media strategies has focused primarily on populist communication, disinformation, mobilization, and personalization, while the substantive policy content of parties' messages has received comparatively little attention (Bene et al., 2024). This gap is particularly pronounced for Instagram Stories—an ephemeral, relatively underexamined campaign format whose integration into political communication research remains limited (Towner and Muñoz, 2024; Bast, 2021).

Instagram poses distinct methodological challenges for issue detection. Its "digital architecture" (Bossetta, 2018) conditions how campaign communication is organized across multiple content spaces: the persistent feed and the ephemeral Story format, which disappears after 24 hours and follows distinct production conventions (Towner and Muñoz, 2024, 2022). In the empirical context we study—German federal elections—manually coded analyses suggest that 40–65% of Instagram posts contain at least one policy issue (Haßler et al., 2023, 2021), yet these analyses do not extend to Stories, which distribute meaning across multiple text-bearing components (on-screen text, spoken content) rather than a single caption field (Towner and Muñoz, 2022). While Stromer-Galley and Rossini (2024) demonstrate that binary issue detection is feasible for campaign messages on Twitter and Facebook, analogous approaches remain untested on Instagram, where heterogeneous text traces and scarce training resources for German-language content compound the challenge. Extending binary issue detection to this platform—including its ephemeral formats and dispersed text traces—thus constitutes both a practical and theoretically meaningful step.

The 2021 German federal election provides a suitable empirical setting. Issue ownership theory suggests that parties strategically emphasize issues they are perceived to handle best while avoiding issues associated with competitors (Walgrave et al., 2015), making issue differentiation a central mechanism of campaign competition. In the 2021 campaign, Instagram had already become an established campaign channel (Haßler et al., 2023), and the open succession following Angela Merkel's departure produced an unusually competitive race among multiple chancellor candidates. This paper addresses these gaps by developing

Proceedings of The 3rd Workshop on Natural Language Processing for Political Sciences (PoliticalNLP 2026) @ LREC 2026, pages 120–130
16 May 2026, Palma, Mallorca (Spain).

and evaluating automated methods for detecting the presence of policy issues in Instagram posts and Stories from this election. We compare supervised fine-tuning (GBERT) with multiple LLM prompting strategies across Instagram's heterogeneous text types (RQ1), examine how issue visibility varies by text type, format, and party (RQ2), and provide exploratory evidence on which policy domains parties emphasise (RQ3). Specifically:

RQ1 How well do prompted LLMs and a fine-tuned transformer (GBERT) perform in binary classification of policy-issue presence, and how do prompting strategies and model choice affect performance?

RQ2 How does the distribution of issue presence vary across text types (caption, OCR, transcript), formats (posts vs. Stories), and parties?

RQ3 Which types of policy issues appear in the campaign texts, and how do parties emphasise these issue areas?

2. Related Work

Conceptual Background

Policy issues are central to political communication, and theoretical traditions converge on why detecting their presence matters. The Comparative Agendas Project (CAP) provides a standardised taxonomy for analysing issue attention across institutions and contexts (Baumgartner et al., 2019). Issue ownership theory holds that parties strategically emphasise issues with which they are perceived as competent (for a review, see Walgrave et al. 2015), while framing and agenda-setting research shows that selective topic emphasis can itself function as a framing mechanism (Klamm et al., 2022). Detecting whether a message contains policy content is thus a necessary first step for assessing ownership patterns, framing choices, or issue competition—a step that becomes especially important on social media, where platform conditions can weaken or reconfigure traditional patterns of issue emphasis (Bene et al., 2024).

Classification of Policy Issues

The computational classification of policy issues is well established for formal political texts. Early supervised approaches using SVMs on bag-of-words features demonstrated that CAP topic assignments can be reproduced at scale for legislative documents (Purpura and Hillard, 2006; Hillard et al., 2008), while dictionary-based methods offered transparent alternatives for lexically distinctive domains (Albaugh et al., 2013; Sevenans et al., 2014). More recently, fine-tuned transformers have advanced the state of the art: Klamm et al. (2022) report micro-F_1 scores in the mid-0.70s for German parliamentary debates, and Sebők et al. (2025) introduce a multilingual pipeline achieving weighted macro-F_1 above 0.75 across several language–domain combinations. LLMs have also shown promise as CAP coders, with few-shot GPT models reaching human-level agreement on congressional texts (Rytting et al., 2023) and hybrid LLM workflows approaching custom-trained baselines (Gunes and Florczak, 2025), though the latter still perform best overall when sufficient training data are available.

Transferring these methods to social media poses distinct challenges. Dictionary-based CAP tools achieve low recall on tweets (Praet et al., 2021), and supervised classifiers lose accuracy when applied across text types or time periods (Burscher et al., 2015). Moreover, a substantial share of social-media campaign content contains no policy content at all (Praet et al., 2021; Hemphill and Schöpke-Gonzalez, 2020), and campaign discourse encompasses communicative functions—mobilisation, relationship-building, ceremonial messaging—that CAP-based schemes leave unaddressed (Hemphill and Schöpke-Gonzalez, 2020; Stromer-Galley and Rossini, 2024). Notably, Stromer-Galley and Rossini (2024) show that a binary *issue* classifier achieves F_1 scores between 0.75 and 0.93 on Facebook posts and tweets, making it one of the most reliably detected categories in their campaign-message scheme. Binary issue detection thus represents a logically prior and more tractable step than multi-class topic assignment, yet it remains largely untested with LLM prompting strategies, fine-tuned German-language transformers, and on platforms beyond Twitter and Facebook.

Issue Communication in Social-Media Campaigning

A separate strand of research examines what parties actually communicate on social media, though this literature remains comparatively limited (Bene et al., 2024) and unevenly distributed across platforms. On Twitter, parties exhibit distinctive issue profiles that reflect ideological positioning (Praet et al., 2021) and participate in mutually influential agenda-setting dynamics with legacy news media and politicians' social media agendas (Gilardi et al., 2022). For the 2021 German federal election specifically, Hellwig et al. (2024) identify COVID-19, climate policy, digitisation, and financial policy as prominent topics in party tweets. On Instagram, evidence is considerably sparser. In a study spanning three nationwide elections Haßler et al. (2023) show that 40–65% of posts across German

elections addressed at least one policy issue, and Haßler et al. (2021) document pronounced cross-party differences in 2017: the Greens addressed policy issues in 52% of posts (predominantly environmental policy), while the CDU did so in only 14%. Exploratory work on the 2021 election using unsupervised topic modeling suggests a format asymmetry, with policy-related topics appearing more frequently in permanent posts than in ephemeral Stories (Achmann and Wolff, 2023), but these patterns remain unvalidated by supervised methods. For Stories, there is no systematic evidence of policy issues, motivating the present study's focus on developing supervised detection methods for Instagram campaign content across both formats.

We address this gap by systematically comparing supervised fine-tuning and LLM prompting strategies for binary issue detection across Instagram's heterogeneous text types.

3. Methods

We focus on the binary classification of *policy issue presence*—whether a given Instagram text contains identifiable references to policy issues, governance measures, or legislative details. The procedure comprises developing a binary-labeled dataset through multi-coder annotation, assessing inter-annotator reliability, and benchmarking a fine-tuned transformer model alongside multiple LLMs under different prompting regimes. This design enables us to compare model performance (RQ1) and to examine the distribution of policy-issue presence across text types, formats, and party accounts (RQ2). Additionally, we apply exploratory topic modeling to the subset of policy-bearing texts to provide descriptive context on the issue domains parties emphasise (RQ3).

Corpus and Annotation

We analyse Instagram content from the 2021 German federal election campaign, collected during the final two weeks before election day (September 12–25, 2021). Posts were retrieved retrospectively via CrowdTangle, while Stories were captured daily using a `selenium`-based script simulating a human user. The dataset comprises 707 posts (1,153 images, 151 videos) and 2,208 Stories (1,246 videos) from eight parties (AfD, CDU, CSU, Grüne, Linke, FDP, Freie Wähler, SPD) and 14 leading candidates. Preprocessing included OCR on images and the first video frame, as well as transcription of video content using a fine-tuned German Whisper model[1], yielding a corpus of 4,614 text documents across three channels: captions, OCR-extracted image text, and speech transcripts.

Drawing on Haßler et al. (2023, 2021), we operationalize *policy issue presence* as the occurrence of direct, indirect, or implicit references to substantive policy topics, governance measures, or legislative details. Explicit stance-taking was not required; the threshold was defined by identifiable policy-related substance, including domain-specific references (e.g. education, pensions, agriculture) and indications of programmatic direction such as proposals or reforms. Generic political communication lacking substantive policy reference (e.g. campaign mobilization, personality-focused messaging, or event reporting) was coded as `False`, as were ambiguous cases. This scheme was applied independently to each text type.

We drew a stratified sample across text type and content format (post vs. Story). Annotation was carried out in two batches by six coders (three assigned independently per document), all native German speakers. Gold-standard labels were derived by majority vote; after excluding ties and one coder due to insufficient quality, the final dataset comprised 1,357 documents (Krippendorff's $\alpha = 0.67$).

Classification Approaches

Based on this gold-labeled dataset, we benchmarked two main classification strategies:

1. **Supervised fine-tuning** of a German-language transformer (GBERT).

2. **Prompt-based classification** using state-of-the-art large language models (LLMs).

This allowed us to compare a reproducible, supervised approach to zero- and few-shot LLM classification. We did not consider traditional ML models, as deep learning-based approaches have surpassed classical machine-learning methods on a range of text classification tasks, and transformer-based pre-trained language models have set new state-of-the-art performance across many NLP tasks (Minaee et al., 2021).

For the LLM evaluation, we adopted a sequential design: we first compared prompting strategies (zero-shot, few-shot, RAG few-shot) on a single model (GPT-4.1) to identify the most effective prompting approach, and then held that approach constant while comparing across models (GPT-4.1, GPT-o3, GPT-5). This avoids a full factorial comparison of three models by three prompting strategies, which would have added complexity disproportionate to the chapter's primary aim of developing a reliable binary detector rather than benchmarking LLM prompting in its own right. The trade-off is that we cannot confirm whether the best-performing

[1] https://huggingface.co/bofenghuang/whisper-large-v2-cv11-german

prompting strategy generalises across all models; we return to this point when discussing the results.

Supervised Model: GBERT. We fine-tuned the `deepset/gbert-large` model (Chan et al., 2020) using the Hugging Face Transformers library. Each document was tokenized and labeled as `True` or `False`. Class imbalance was addressed via class weighting in the loss function. Hyperparameters were optimized using Weights & Biases sweeps, with macro F_1 as the main target metric.[2] Training was performed on an Nvidia L4 GPU (RunPod), with stratified five-fold cross-validation for robust evaluation.

LLMs: Prompting strategies. We further evaluated several GPT-family models (`gpt-4.1-2025-04-14`, `o3-2025-04-16`, `gpt-5-2025-08-07`) and three prompting setups:

Zero-shot: Direct classification without any in-context examples.

Few-shot: Prompts containing a fixed set of `True` and `False` examples.

RAG few-shot: Retrieval-augmented prompting, dynamically retrieving examples from the annotated pool as context, inspired by recent research in LLM-based classification (Leitner et al., 2025).

All prompts were in English and instructed the models to return `JSON` formatted output. Across models, we used default parameters (*top_p*, *max_tokens*, *verbosity*, *reasoning*), except for *temperature*, which was set to 0 for GPT-4.1. The annotation guidelines served as the template for LLM prompt design: Through several informal iterations, we optimized the prompt in ChatGPT.[3]

Evaluation design. For comparability, all models were evaluated on the same held-out 20% split ($n = 272$), using the remaining 80% for training and (where applicable) few-shot example selection. For the best-performing supervised model and LLM configuration, we then report stratified five-fold cross-validation on the full gold-standard dataset ($n = 1{,}357$), computing metrics on held-out folds only. For retrieval-augmented prompting, retrieved examples were restricted to the respective training fold. All metrics (precision, recall, F_1) were computed with `scikit-learn`.

[2] See our code repository for parameter ranges and details: https://github.com/michaelachmann/political-nlp-issue-detection

[3] See repository for prompt and more details.

Full-corpus inference. For final inference, we fine-tuned GBERT on the entire annotated dataset—removing the train/test split used during evaluation—to maximise the training signal available for the final classifier. The retrained model was then applied to all documents in the corpus, including those with existing human annotations, to ensure uniform labelling across the dataset.

Topic Modeling

To provide descriptive context on the policy domains parties emphasise (RQ3), we applied BERTopic (Grootendorst, 2022) to documents classified as containing a policy issue. Texts were split into overlapping two-sentence windows and filtered for minimum length (≥ 30 characters, ≥ 5 tokens), yielding 4,556 windows corresponding to 872 unique posts/Stories. We used multilingual sentence embeddings, UMAP for dimensionality reduction, and HDBSCAN for clustering, tuning parameters to prioritise interpretable higher-level clusters. Intrinsic metrics indicate moderate topic quality ($c_v = 0.47$, topic diversity $= 0.85$, intra-topic cosine similarity 0.51), with 19.5% of documents assigned to the outlier cluster. After removing outliers, one author mapped the remaining 43 clusters to coarse policy-issue categories based on keywords and sample documents. Clusters capturing meta-campaign content rather than identifiable policy domains were excluded, leaving 801 documents corresponding to 703 unique posts/Stories for interpretation. The topic model serves as a descriptive illustration and is not used for inferential analysis.

4. Results

Having outlined the annotation scheme, dataset construction, and modeling setup, we now turn to the empirical evaluation. We first address model performance (RQ1): (a) prompting strategies on GPT-4.1, (b) cross-model differences among GPT variants under the best prompt, and (c) a supervised fine-tuned transformer (GBERT). We then examine how policy-relevant content is distributed across text types, formats, and parties (RQ2).

RQ1 — Model Performance

(a) Prompting strategies. We first compared three prompting configurations on GPT-4.1: zero-shot, static few-shot, and retrieval-augmented few-shot (RAG-Few). As Table 1 shows, the binary F_1 scores are between 0.76 and 0.81, indicating that GPT-4.1 can infer the notion of policy-issue messaging without in-context examples. Few-shot prompting yielded the highest binary F_1, driven

Table 1: Evaluation results for prompting strategies and GPT variants on the 20% test split ($n = 272$). Binary F_1 corresponds to the TRUE class.

Model	Prompting	F_1 Macro	F_1 FALSE	F_1 TRUE
GPT-4.1	Zero-Shot	0.87	0.94	0.79
GPT-4.1	RAG-Few	0.84	0.93	0.76
GPT-4.1	Few-Shot	0.87	0.93	0.81
GPT-o3	Few-Shot	**0.89**	**0.95**	**0.83**
GPT-5	Few-Shot	0.82	0.92	0.73
Support		272	214	58

primarily by a large recall gain on the TRUE[4] class—though at the cost of reduced precision, indicating a tendency to over-predict policy-issue presence. Macro F_1 remained unchanged from the zero-shot condition, suggesting that few-shot prompting mainly redistributes errors rather than improving overall balance.

(b) Model comparison. Holding the few-shot setup constant, we evaluated two additional GPT-family models—GPT-o3 and GPT-5—to isolate model effects from prompt-engineering choices. GPT-o3 achieved the strongest LLM performance overall (Table 1), with a more balanced precision–recall trade-off on the minority class than GPT-4.1. By contrast, GPT-5 underperformed both alternatives, with noticeably lower recall on the TRUE class. This is somewhat surprising given expectations about newer model generations and may reflect differences in how the models handle the evaluative, borderline language that characterises much of the corpus.

(c) Supervised model. The fine-tuned GBERT model achieved competitive performance on the held-out split (binary F_1 = 0.81), closely matching the best LLM results. Performance was balanced across classes, with strong recall on the minority class supported by moderately lower precision. These results indicate that supervised fine-tuning of a German BERT model provides a reliable approach to detecting policy issues, matching the best prompting configurations without requiring proprietary models.

Cross-validation. To assess generalisation, we subjected both GBERT and GPT-o3 (few-shot) to stratified five-fold cross-validation. GBERT achieved a mean macro F_1 of 0.90 ($\pm$0.03) with consistently strong and stable results across folds (Table 2). The emulated cross-validation for GPT-o3 yielded a slightly lower macro F_1 of 0.88 ($\pm$0.02), with wider variance on TRUE-class recall (Table 3). Both approaches generalise well, with a modest gap in the minority class.

Error analysis. Qualitative inspection of misclassified texts ($n = 21$) for GPT-o3 revealed two systematic patterns. False positives typically involve issue-domain language—e.g. attacks of opponents with demands for fiscal transparency ("@OlafScholz muss [...] Dokumente sofort freigeben") or alarmist framing of an opponent's position ("FDP will JEDES JAHR 500.000 Zuwanderer ZU UNS HOLEN!") or short slogans—that border on implicit policy advocacy. Given moderate annotator agreement ($\alpha = 0.67$), several of these cases reflect genuine conceptual ambiguity rather than clear model error; GPT-o3 appears somewhat more inclusive than annotators in treating issue-linked language as substantive policy issues. False negatives presented policy content in indirect or motivational form—linking climate protection to electoral choice, or embedding reform language in mobilisation rhetoric—rather than as explicit proposals.

GBERT exhibited similar false-positive patterns but additionally missed highly compressed or sloganized expressions (e.g., "Grenzen schützen," "TAX THE RICH") that convey clear directional meaning. This suggests that the fine-tuned model relies more strongly on explicit propositional structure and is comparatively less sensitive to condensed articulations of policy stance. These complementary error profiles—GPT-o3 slightly more inclusive, GBERT slightly more conservative—inform model selection.

Model selection. GPT-o3 is the best-performing LLM under few-shot prompting, but its advantage over GBERT is small. GBERT offers good cross-validation robustness alongside practical advantages in transparency, reproducibility, and inference cost. We therefore employ GBERT to generate labels for the substantive analyses in RQ2 and RQ3.

RQ2 — Issue Presence

Having selected GBERT for inference, we now examine what the classifications reveal about campaign communication. In line with RQ2, we describe how policy presence is distributed across Instagram content. Because posts and Stories may contain multiple textual channels (captions, OCR-extracted text, and, where available, speech transcripts), we report results at both the document level (text type) and the content level (post type and party differences). This descriptive analysis provides a basis for interpreting how political actors

[4]Throughout the results, TRUE denotes the presence of policy issues in a document, FALSE their absence.

Table 2: Five-fold cross-validation results for GBERT. Reported are mean metrics ± SD.

Class	Precision	Recall	F_1
FALSE	0.97 ± 0.02	0.94 ± 0.02	0.96 ± 0.02
TRUE	0.81 ± 0.06	0.88 ± 0.06	0.84 ± 0.05
Macro avg	0.89 ± 0.03	0.91 ± 0.03	**0.90 ± 0.03**

Table 3: Five-fold cross-validation results for GPT-o3 (FEW-SHOT). Reported are mean metrics ± SD. Macro averages derived from class means.

Class	Precision	Recall	F_1
FALSE	0.95 ± 0.02	0.95 ± 0.01	0.95 ± 0.01
TRUE	0.81 ± 0.04	0.80 ± 0.08	0.81 ± 0.03
Macro avg	0.88 ± 0.02	0.88 ± 0.038	**0.88 ± 0.02**

foreground substantive policy issues across different communicative environments on Instagram.

Distribution by text type. Captions featured policy issues most frequently, with 443 of 714 items (62.0%), followed by transcripts (241 of 652; 37.0%), and OCR text (494 of 3,248; 15.2%). A chi-square test confirms that issues are distributed unevenly across text types ($\chi^2 = 727.51$, $p < .001$), with a medium effect (Cramér's $V = .40$).

Distribution by post type. Nearly two-thirds of posts (496 of 707; 70.2%) contained at least one policy-issue reference, whereas only 370 of 2,208 Stories (16.8%) did so. An additional 28 Stories (1.3%) were not associated with any textual content (i.e., neither OCR nor transcript could be extracted) and were therefore coded as NA for issues. This difference is substantial ($\chi^2 = 733.09$, $p < .001$), with a medium effect size (Cramér's $V = .35$).

Distribution by party. References to policy issues varied across parties but differences were limited in magnitude (see Figure 1), ranging from $\approx 18\%$ (AfD, Free Voters) to 38% (Die Linke, FDP), with CDU/CSU and SPD in between. The effect size was small ($\chi^2 = 46.83$, $p < .001$, Cramér's $V = .13$), indicating only modest cross-party variation.

Party–format interactions. The format gap held consistently across all parties (Figure 2). Post-level issue references ranged from 55–83%, with the highest rates for the Greens, FDP, and SPD. In contrast, Story-level references were uniformly low (9–21%), with most parties clustering between 14–21% and only the AfD and Free Voters falling noticeably below this range. A chi-square test confirmed a substantial interaction ($\chi^2 = 759.17$, $p < .001$, Cramér's $V = .51$) though the effect is driven primarily by format rather than between-party differences within formats. Format choice—post vs. Story—was thus the dominant factor in determining whether policy issues were foregrounded.

Table 4: Manually assigned issue labels from BERTopic clusters (post/Story-level).

Policy Issue Label	**Count**
Environment & Climate	259
Economy & Taxes	101
Social & Welfare Policy	76
Housing	54
Education	44
Digitalization & Infrastructure	42
Economy & Growth	41
Labour & Wages	40
Gender & Equality	33
Foreign & Security	28
Economy & Innovation	24
Pensions	21
Health & Care	16
Culture & Media	12
Agriculture & Animal Welfare	7
Scandals & Accountability	3

RQ3 — Issue Types and Their Emphasis Across Parties

To explore the substantive content of policy-relevant texts, we draw on the exploratory topic modeling described in the Methods section, treating results as descriptive findings rather than inferential evidence. BERTopic yielded 43 clusters, of which a subset mapped to coherent policy themes (Table 4).[5]

Environment & Climate emerged as the dominant domain, followed by multiple economy-related categories and social policy areas. At the party

[5]See repository for full topic list and mapping.

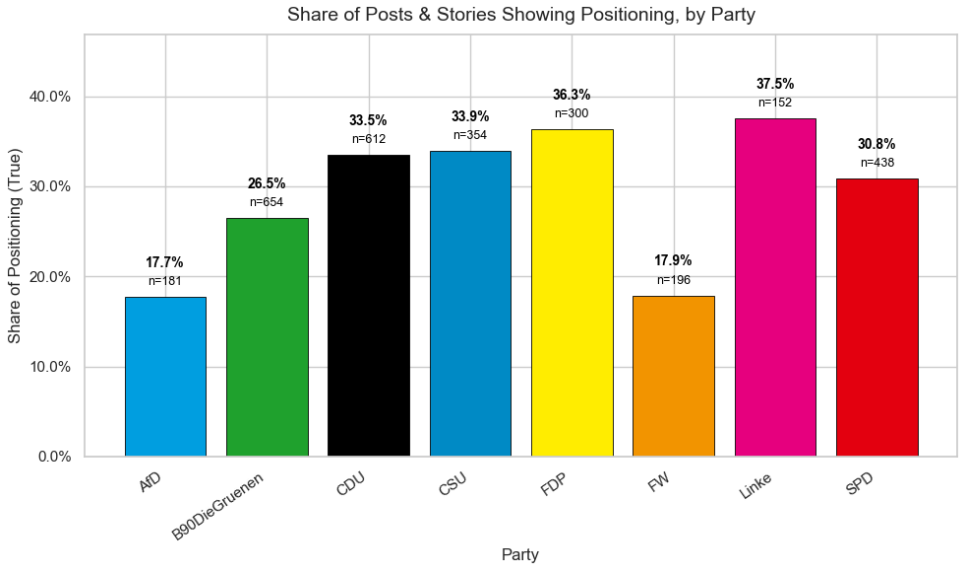

Figure 1: Share of posts and Stories containing policy-issue references by party, collapsed across format. Reference rates vary moderately by party (≈ 18˘38%), with The Left, FDP, CDU/CSU, and SPD at the upper end and AfD and Free Voters at the lower end.

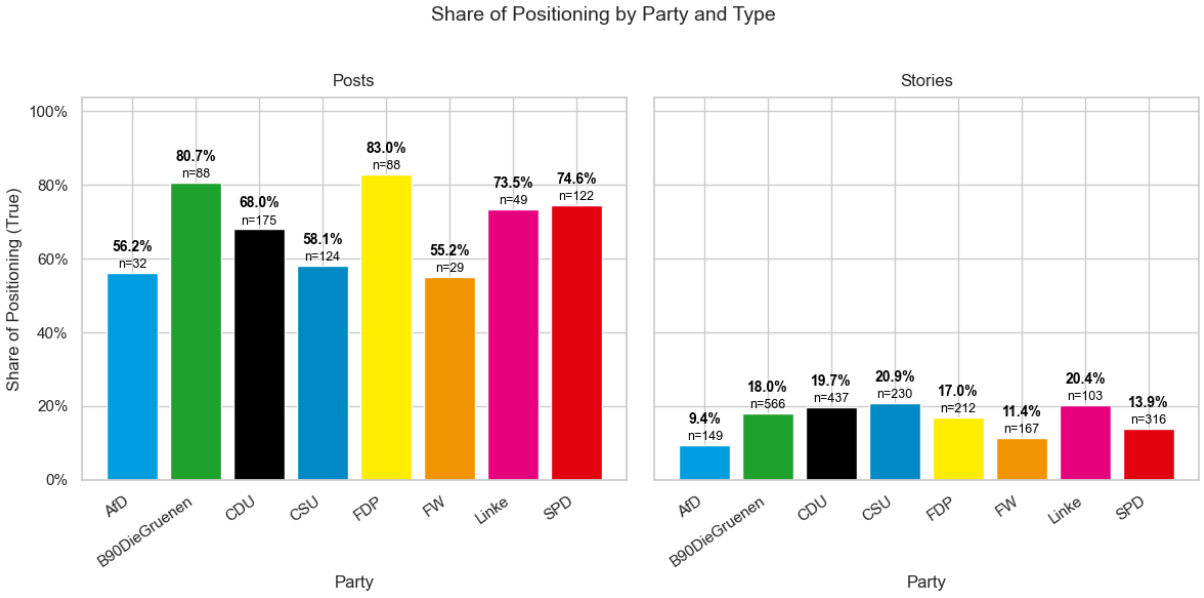

Figure 2: Share of posts and Stories containing policy-issue references by party. Posts show consistently high reference rates across parties, whereas issues in Stories remain uniformly low.

level, two broad patterns stand out: parties differ in issue breadth, and they vary in which issues they emphasise. The Greens display the most concentrated profile, dominated by climate and gender equality. The FDP similarly concentrates on a narrow set—digitalisation, education, and fiscal policy—but with a different orientation. The SPD shows one of the broadest profiles, spanning labour, housing, pensions, climate, and education, while the CDU/CSU exhibit a more economically centred agenda with additional attention to foreign and security policy. Smaller parties (Die Linke, AfD, Free Voters) contribute fewer interpretable texts; Die Linke shows some emphasis on labour and welfare topics, while the AfD and Free Voters are only marginally represented across categories.

In sum, both supervised and prompt-based models reliably detect policy-issue presence in heterogeneous Instagram campaign texts. The substantive analyses reveal that policy visibility varies far more by content format and text channel than by party, while the exploratory topic model confirms familiar patterns of issue emphasis within the subset of policy-relevant texts. The following discussion

situates these findings within the broader literature on automated issue classification and considers what the pronounced format asymmetry implies for understanding policy communication on visual social media platforms.

5. Discussion

The findings converge on three themes: (i) binary issue detection is tractable even in a heterogeneous Instagram text corpus, (ii) policy visibility is structured primarily by platform format rather than party identity, and (iii) when parties do communicate policy, their issue emphases largely reproduce familiar patterns.

Binary Issue Detection as a Feasible Foundation. Both GBERT and GPT-o3 reliably detected policy-issue presence across Instagram's heterogeneous text types. With cross-validated macro F_1 scores of 0.90 and 0.88 respectively, performance compares favourably to prior work on social-media issue classification, where dictionary-based CAP tools achieve low recall (Praet et al., 2021), supervised classifiers degrade across text types (Burscher et al., 2015), and even dedicated binary issue detectors report F_1 between 0.75 and 0.93 depending on platform and model (Stromer-Galley and Rossini, 2024).

Notably, these results were achieved with a relatively small gold standard of 1,357 annotated documents. That both a fine-tuned encoder and a prompted LLM reach competitive performance under these conditions underscores the viability of either approach for researchers working with limited annotation budgets. Fine-tuned models offer reproducibility and low inference cost; LLMs require minimal engineering and no task-specific training data, making them a versatile option for future analyses where gold standards are unavailable or infeasible. The choice between them is likely to depend on practical considerations rather than fundamental performance differences. Taken together, the results indicate that the main bottleneck for measuring issue visibility in Instagram campaigns may be less model capability than data access and modality coverage (e.g., video transcripts, OCR, and Stories).

Moderate intercoder agreement ($\alpha = 0.67$) places a natural ceiling on model performance, reflecting genuine conceptual ambiguity about where policy content ends. Reported scores should therefore be read as conservative estimates.

Policy Visibility and Format. The most distinctive finding concerns the structuring role of format. Posts carried the vast majority of substantive policy content, whereas Stories—across all parties—rarely did. This gap overshadows cross-party variation, suggesting that the post–Story distinction reflects different communicative logics rather than idiosyncratic editorial choices. On its own, our analysis can only establish this asymmetry, not explain it.

The differences across textual channels add further nuance. Captions, Instagram's only native text channel, exhibit the highest rate of policy references, followed by speech transcripts, with OCR-extracted text. Methodologically, this means that analyses restricted to captions risk undercounting policy communication when video content is prevalent, while ignoring Stories inflates the apparent share of issue-oriented output. Prior work on German Instagram campaigning has documented policy-issue rates in permanent posts (Haßler et al., 2021, 2023), our results extend these findings by showing that ephemeral formats constitute a large and substantively distinct layer of campaign activity that systematic analyses cannot afford to overlook.

Issue Emphasis and the Persistence of Ownership Patterns. Within the subset of policy-relevant texts, the exploratory topic model provides a coarse but informative map of the issue landscape. The Greens concentrated on environment and climate; the FDP emphasised digitalisation, education, and fiscal policy; the SPD and the Union parties distributed references across a broader range of domains. These profiles are broadly consistent with established issue-ownership expectations (Walgrave et al., 2015) and mirror cross-party variation documented in earlier Instagram (Haßler et al., 2021) and Twitter analyses (Praet et al., 2021), where ideologically distinct parties exhibit sharper issue profiles than centrist ones. The persistence of these patterns across platforms and election cycles suggests that parties' core issue agendas are relatively stable communicative commitments rather than platform-contingent adaptations. At the same time, the constraints of the topic model mean that the topics likely capture only the most salient policy communication.

Limitations and Future Work

Since GBERT is comparatively insensitive to compressed or sloganized expressions—which are especially common in OCR—reported policy-issue rates should be understood as conservative estimates. A manual evaluation of OCR quality on 50% of all Stories in the corpus found a character-level error rate of approximately 3%, suggesting that the low policy-issue rate in OCR-extracted text is unlikely to be driven primarily by extraction artifacts.

The LLM evaluation carries two design-related

constraints. The sequential comparison of prompting strategies and models means we cannot confirm that few-shot prompting is equally effective across all tested models; a full factorial design would be needed to establish this. Moreover, all LLM experiments relied on proprietary models that cannot be fully versioned or reproduced, further motivating our choice of GBERT for the substantive analyses. The exploratory topic model offered only coarse issue clusters, constraining fine-grained interpretation of party emphases. More broadly, our empirical scope is limited to one election, one platform, and one linguistic context. That said, the format asymmetry between posts and Stories may reflect broader conventions of platform vernaculars (Gibbs et al., 2015) rather than election-specific dynamics, whereas absolute policy-issue rates and party-level emphasis patterns are more plausibly tied to the specific competitive constellation of the 2021 campaign.

Several directions follow. Sentence-level annotation could reduce ambiguity by isolating individual claims, and iterative codebook refinement would lay the groundwork for multi-class CAP-level classification. Replicating the LLM evaluation with open-weight models (e.g., Llama, Mistral) would strengthen reproducibility and reduce dependence on proprietary APIs whose behaviour may change over time. Finally, Stories have become a cross-platform format—now also available on Facebook, WhatsApp, and more—and applying the binary detector across platforms and election cycles would clarify whether the low policy density observed here is Instagram-specific or a more general consequence of ephemeral format logics.

Ethical Considerations and Data Availability

We collected only publicly available content from verified party and front-runner accounts; no further user-generated data (e.g., comments) is included. Following Venturini and Rogers (2019), we acknowledge the ethical tensions inherent in scraping but consider it warranted here, as the data comprise campaign communications that institutional political actors intentionally published for broad public reach.

To support transparency and reproducibility, we release annotation guidelines, code, prompts, and annotations. The text content itself has been redacted to comply with the platform's terms of service. All data is available at `https://github.com/michaelachmann/political-nlp-issue-detection`.

6. Bibliographical References

Michael Achmann and Christian Wolff. 2023. Policy issues vs. Documentation: Using BERTopic to gain insight in the political communication in Instagram stories and posts during the 2021 German Federal election campaign. In *DHNB Publications*, volume 5, pages 11–28, Oslo. University of Oslo Library.

Quinn Albaugh, Julie Sevenans, Stuart Soroka, and Peter John Loewen. 2013. The Automated Coding of Policy Agendas: A Dictionary-Based Approach.

Jennifer Bast. 2021. Politicians, Parties, and Government Representatives on Instagram: A Review of Research Approaches, Usage Patterns, and Effects. *Review of Communication Research*, 9.

Frank R Baumgartner, Christian Breunig, and Emiliano Grossman. 2019. The Comparative Agendas Project : Intellectual Roots and Current Developments. In *Comparative Policy Agendas*, pages 3–16. Oxford University Press.

Márton Bene, Melanie Magin, and Jörg Haßler. 2024. Political issues in social media campaigns for national elections: A plea for comparative research. *Politics and governance*, 12(0).

Michael Bossetta. 2018. The Digital Architectures of Social Media: Comparing Political Campaigning on Facebook, Twitter, Instagram, and Snapchat in the 2016 U.S. Election. *Journalism & mass communication quarterly*, 95(2):471–496.

Bjorn Burscher, Rens Vliegenthart, and Claes H De Vreese. 2015. Using Supervised Machine Learning to Code Policy Issues: Can Classifiers Generalize across Contexts? *The ANNALS of the American Academy of Political and Social Science*, 659(1):122–131.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Stroudsburg, PA, USA. International Committee on Computational Linguistics.

Martin Gibbs, James Meese, Michael Arnold, Bjorn Nansen, and Marcus Carter. 2015. #Funeral and Instagram: death, social media, and platform vernacular. *Information, Communication and Society*, 18(3):255–268.

Fabrizio Gilardi, Theresa Gessler, Maël Kubli, and Stefan Müller. 2022. Social media and political agenda setting. *Political communication*, 39(1):39–60.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv [cs.CL]*.

Erkan Gunes and Christoffer Koch Florczak. 2025. Replacing or enhancing the human coder? Multiclass classification of policy documents with large language models. *Journal of Computational Social Science*, 8(2):31.

Jörg Haßler, Anna Sophie Kümpel, and Jessica Keller. 2021. Instagram and political campaigning in the 2017 German federal election. A quantitative content analysis of German top politicians' and parliamentary parties' posts. *Information, Communication and Society*, pages 1–21.

Jörg Haßler, Anna-Katharina Wurst, and Katharina Pohl. 2023. Politicians over issues? Visual personalization in three Instagram election campaigns. *Information, Communication and Society*, pages 1–21.

Nils Constantin Hellwig, Jakob Fehle, Markus Bink, Thomas Schmidt, and Christian Wolff. 2024. Exploring Twitter discourse with BERTopic: topic modeling of tweets related to the major German parties during the 2021 German federal election. *International journal of speech technology*, 27(4):901–921.

Libby Hemphill and Angela M Schöpke-Gonzalez. 2020. Two Computational Models for Analyzing Political Attention in Social Media. *Proceedings of the International AAAI Conference on Web and Social Media*, 14:260–271.

Dustin Hillard, Stephen Purpura, and John Wilkerson. 2008. Computer-assisted topic classification for mixed-methods social science research. *Journal of information technology & politics*, 4(4):31–46.

Christopher Klamm, Ines Rehbein, and Simone Paolo Ponzetto. 2022. FrameASt: A framework for second-level agenda setting in parliamentary debates through the lense of comparative agenda topics. In *2022 Workshop on Creating, Enriching and Using Parliamentary Corpora, ParlaCLARIN III 2022*, pages 92–100, Paris. European Language Resources Association (ELRA).

Maxyn Rose Leitner, Rebecca Dorn, Fred Morstatter, and Kristina Lerman. 2025. Characterizing network structure of anti-trans actors on TikTok. In *Proceedings of the 17th ACM Web Science Conference 2025*, pages 472–483, New York, NY, USA. ACM.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep Learning–based Text Classification: A Comprehensive Review. *ACM Comput. Surv.*, 54(3):1–40.

John R Petrocik. 1996. Issue ownership in presidential elections, with a 1980 case study. *American Journal of Political Science*, 40(3):825.

Stiene Praet, Peter Van Aelst, Walter Daelemans, Tim Kreutz, Jeroen Peeters, Stefaan Walgrave, and David Martens. 2021. Comparing automated content analysis methods to distinguish issue communication by political parties on Twitter. *Computational Communication Research*, 3(2):1–27.

Stephen Purpura and Dustin Hillard. 2006. Automated classification of congressional legislation. In *Proceedings of the 2006 international conference on Digital government research*, New York, New York, USA. Digital Government Society of North America.

Christopher Michael Rytting, Taylor Sorensen, Lisa Argyle, Ethan Busby, Nancy Fulda, Joshua Gubler, and David Wingate. 2023. Towards coding social science datasets with language models. *arXiv [cs.AI]*.

Miklós Sebők, Ákos Máté, Orsolya Ring, Viktor Kovács, and Richárd Lehoczki. 2025. Leveraging open large language models for multilingual policy topic classification: The Babel Machine approach. *Social science computer review*, 43(2):295–317.

Julie Sevenans, Quinn Albaugh, Tal Shahaf, Stuart Soroka, and Stefaan Walgrave. 2014. The Automated Coding of Policy Agendas: A Dictionary Based Approach.

Jennifer Stromer-Galley and Patricia Rossini. 2024. Categorizing political campaign messages on social media using supervised machine learning. *Journal of information technology & politics*, 21(4):410–423.

Terri L Towner and Caroline L Muñoz. 2024. Tell Me an Instagram Story: Ephemeral Communication and the 2018 Gubernatorial Elections. *Social science computer review*, page 08944393241227554.

Terri L Towner and Caroline Lego Muñoz. 2022. A Long Story Short: An Analysis of Instagram Stories during the 2020 Campaigns. *Journal of Political Marketing*, pages 1–14.

Tommaso Venturini and Richard Rogers. 2019. "API-Based Research" or How can Digital Sociology and Journalism Studies Learn from the Facebook and Cambridge Analytica Data Breach. *Digital Journalism*, 7(4):532–540.

Stefaan Walgrave, Anke Tresch, and Jonas Lefevere. 2015. The conceptualisation and measurement of issue ownership. *West European politics*, 38(4):778–796.