GePaDeSE: A New Resource for Clause-Level Aspect in German Parliamentary Debates

Julian Schlenker♣, Ines Rehbein♠, Lilly Brauner♡, Florian Ertz♢, Ines Reinig♣, Simone Paolo Ponzetto♣

♣University of Mannheim, ♠University of Münster,
♡Heidelberg University, ♢University of Göttingen
julian.schlenker@uni-mannheim.de

Abstract

This paper presents GePaDeSE, a new resource with annotations of clause-level aspect in German parliamentary debates, also known as Situation Entity types. The new resource includes 250 political speeches from the German Bundestag, given by 192 speakers, with over 220,000 tokens. In the paper, we first describe the new corpus and the annotation process. Then we present experiments on automatically classifying clause-level aspect and present an in-depth analysis where we show the potential of Situation Entities for the analysis of political discourse.

Keywords: Situation Entities, Clause-level Aspect, Parliamentary Debates, Political Text Analysis

1. Introduction

This paper presents a new resource for studying clause-level aspect in German political debates. While for English several corpora annotated with aspect features exist (Alikhani and Stone, 2019; Gantt et al., 2022; Friedrich and Palmer, 2014b), only few resources are available for languages other than English, especially for the political domain. Our work addresses this research gap by presenting a new, large dataset for German, annotated for Situation Entities (Smith, 2003). Smith's Situation Entity (SE) types provide a classification for clause-level aspect, covering eventualities (States, Events, Reports), general statives (Generic, Generalizing sentences) and abstract entities (Fact, Proposition; see Table 1 for illustrating examples and §2.1 for an introduction).

Applications that benefit from information on aspect include the processing of temporal information as well as Machine Translation where an incorrect translation of aspect can result in misleading readings of a text (see examples in Friedrich et al. (2023, p.619)). In addition, we argue that aspectual information is also relevant for applications in the area of political text analysis and other analyses in the social sciences, which are the main focus of our work. For illustration, consider the Situation Entity types Generic and Generalizing. Both have been discussed as linguistic devices associated with *stereotypes* (Geurts, 1985; Leslie, 2014; Radden, 2009; Novoa et al., 2023; Bosse, 2024; Ralston, 2024). The automatic extraction of Generic and Generalizing sentences, however, is known to be challenging and existing datasets for this task are limited. Our new dataset includes naturally occurring sentences, produced by humans in real-world settings in the political domain. We believe that our data provides a valuable resource for studies on the use of Generic sentences and its association with stereotyping in argumentative text.

As another example, consider the SE types Fact (as objects of knowledge) versus Proposition (as objects of belief). Recasens et al. (2013) coined the term *epistemological bias*, referring to "whether propositions that are presupposed or entailed in the text are uncontroversially accepted as true". Previous work on epistemological bias has highlighted the importance of the concept in the context of framing (i.e., presenting a proposition with different degrees of certainty, e.g., as universal truth or as an unverified assumption).

However, most work on the identification of epistemological bias so far was rule-based (Recasens et al., 2013; Patel and Pavlick, 2021; Rehbein et al., 2024), relying on lexicons of assertives and implicatives (Hooper, 1975; Karttunen, 1971). Through the annotation of abstract objects in our corpus, we provide test data for investigating the validity and coverage of lexicon-based extraction of epistemological bias. These two examples, stereotypes and epistemological bias, motivate the relevance of aspect for political text analysis.[1]

2. Related Work

We first introduce SE types and give an overview over existing datasets annotated for aspect. Then we review work on computational modeling of aspectual features, focusing on SE types.

[1] All code and data for this work are available at our GitHub repository: https://github.com/umanlp/gepadese.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3617–3632
11–16 May 2026.

Situations	SE types	Examples
Eventualities	STATE	Merkel is German chancellor.
	EVENT	Merkel opens the borders.
	EVENT-PERFECT-STATE	Merkel has opened the borders.
	REPORT	..., says Merkel.
General Statives	GENERIC	Politicians participate in policy-making processes.
	GENERALIZING	Merkel often met with Macron.
Speech acts	QUESTION	Does Merkel still rule Germany?
	IMPERATIVE	Don't forget to vote for Merkel!
Abstract Entities	FACT	I know that Merkel is a physicist.
	PROPOSITION	I believe that Merkel is a physicist.

Table 1: Examples for the different SE types, following Smith (2003).

2.1. Situation Entities in a Nutshell

SE have been introduced by Smith (2003) in her book "Modes of discourse" where the author identifies five distinct discourse modes: Narrative, Description, Report, Information, and Argument. Discourse modes can be seen as *linguistic* properties of text passages, thus contrasting a more pragmatic view that typically focuses on the genre of a text. According to Smith, discourse modes can be characterized by two features, (i) the situations they introduce (i.e., Events, States, General statives and Abstract Entities) and (ii) their manner of progression, either temporal or metaphorical (Palmer and Sporleder, 2009). Each discourse mode can be identified by its distinct pattern of SE types that are predominant in this particular mode.

SE operate on the clause level, evoked by the main verb and its arguments, and can refer to situations in the world as well as to more abstract descriptions of kinds or abstract individuals (SE type GENERIC). GENERALIZING sentences, on the other hand, express a pattern or regularity rather than a particular event or state (Krifka et al., 1995). In addition, there are abstract entities which include FACTS (objects of knowledge) and PROPOSITIONS (objects of belief). Those entities can also have an SE type, resulting in two labels for the same instance (see Example 1).

(1) I know (STATE)
that Merkel opened the borders.
(FACT, EVENT)

Finally, the SE type classification includes two speech acts, QUESTION and IMPERATIVE (See Table 1 for an overview of SE types and examples).

2.2. Corpora Annotated for Aspect

Aspect is a well-studied phenomenon in linguistics (see, e.g., Vendler (1967); Dowty (1979); Asher (1993); Smith (1983, 1991), among others) and several datasets with annotations of one or more aspectual features have been created, such as *stativity, habituality, punctuality, telicity, durativity* and *boundedness* (Friedrich and Palmer, 2014a; Friedrich et al., 2016; Friedrich and Gateva, 2017; Alikhani and Stone, 2019; Govindarajan et al., 2019; Kober et al., 2020; Gantt et al., 2022). The first dataset annotated for SE types has been presented by Palmer et al. (2007), including around 6,000 clauses from the English Brown corpus (Francis and Kučera, 1979) and MUC-6 (Grishman and Sundheim, 1996). Another, much larger, SE dataset has been created by Friedrich et al. (2016), with over 40,000 clauses extracted from the MASC corpus (Ide et al., 2008, 2010) and Wikipedia, covering 13 different text genres to enable studies of the interaction between SE types and discourse modes.

While all of the resources above have been created for English, only few datasets are available for other languages. Loáiciga and Grisot (2016) present a small-scale study on English-French parallel texts from the Europarl corpus where they aim at improving the results of a statistical Machine Translation system using *boundedness*, i.e., "whether or not a situation is described as having reached a temporary boundary" (Depraetere, 1995). Egg et al. (2019) create the SdeWac-Aspect corpus, a language resource with 4,200 German clauses annotated for *stativity, durativity* and *boundedness*.

Mavridou et al. (2015) present a small-scale study on English-German parallel data with roughly 2,500 clauses, confirming that the SE schema developed for English is also applicable to German with some minor adaptations. In follow-up work, the schema is applied to German texts from different genres, creating a dataset with roughly 18,000 annotated clauses (Becker et al., 2017). The distribution of SE types strongly deviates across genres, similar to what has been observed for the English SE corpus (Friedrich et al., 2016).

2.3. Computational Modeling of SE

We now turn to related work on automatic prediction of SE types. Again, most work has been done for

English, based on the SE corpus of Friedrich et al. (2016). Initial work has used linguistic features to predict SE types, experimenting with Maximum Entropy models (Palmer et al., 2007), Random Forests and Conditional Random Fields (Friedrich et al., 2016) and using distributional features represented as Brown clusters (Brown et al., 1992). Experiments showed that providing more context to the model (e.g., by including predicted labels for previous clauses or utilizing genre features) is beneficial.

More recent work replaced linguistic features for modeling aspect with distributional representations (Heuschkel, 2016; Kober et al., 2020) and utilized deep learning architectures like RNNs (Becker et al., 2017) and (Bi)LSTMs and contextualized text representations (Dai and Huang, 2018), based on transformers (Vaswani et al., 2017; Devlin et al., 2019). Rezaee et al. (2021) combine a variational auto-encoder (VAE) model with contextualized BERT-based representations to encode the input and then employ a generative model, GPT-2 (Radford et al., 2019), to regenerate the input text. For computational approaches to modeling other types of aspectual features beyond SE types, see (Friedrich et al., 2023).

3. Annotation: Data and Process

This section describes the data and situation entity annotation process used to create GePaDeSE.

Data Our corpus includes 250 speeches from the German Bundestag,[2] mostly from the 19th legislative term (2017-2021), held by 192 speakers from 6 parties (CDU/CSU: 72, SPD: 55, AfD: 37, FDP: 31, The Left: 27, Greens: 24, non-attached: 4). The total size of the data is 222,387 tokens. For more detailed information on the data and sampling method, please refer to the GePaDeSE datasheet.[3]

Annotation scheme We follow the detailed annotation guidelines developed in the SE project for English and German.[4] The guidelines have been tested in a cross-linguistic corpus study comparing English and German annotations of Situation Entity types and developing guidelines for the annotation of German SEs (Mavridou, 2016). While the SE schema proved to be applicable across languages, small adaptations had to be made for German perfect tense clauses where the boundary between states and events is not always clear. To address this issue, a new class EVENT-PERFECT-STATE has been introduced (see example 2 for illustration).

(2) Wir haben die Grenzen bereits geschlossen.
We have the borders already closed.
"We have already closed the borders."

In German, the example can be interpreted as either the event of closing the borders or, in the second reading, can be understood as describing the result of the event (i.e., being in a state where the borders are closed now). We follow Mavridou (2016) and also mark clauses as EVENT-PERFECT-STATE that can not be unambiguously identified as either State or Event.

Annotators The annotations have been carried out by two advanced students of linguistics. All finite verbs in each speech have been independently annotated by each of the two coders. The students have been trained and received feedback throughout the annotation duration.

Annotation setup We noticed that the automatic segmentation of texts into clauses done in the SE project introduced some errors in the data. We therefore decided not to use a sequence labeling setup where annotators assign SE types to segments of text but, instead, to directly assign SE labels to each finite verb. To ease the annotation process, we preprocessed the data and assigned dummy labels to each finite verb, based on the predictions of the spaCy PoS tagger (Honnibal et al., 2020).[5] We instructed the annotators to replace the dummy labels with proper SE types and also to remove incorrectly predicted dummy tags and add missing tags where the tagger failed to identify the finite verb. For annotation, we use the INCEpTION platform (Klie et al., 2018).

Inter-annotator agreement We compute a Cohen's κ of 0.538, indicating moderate agreement (Landis and Koch, 1977). For comparison, Becker et al. (2016) report a Cohen's κ of 0.4 on argumentative microtexts and a slightly higher κ of 0.5 on texts from different genres that have been used for annotator training. The percentage agreement between our annotators is 73%.

Out of the 19,676 instances, annotators disagreed on 5,355 cases (27%), highlighting the inherent subjectivity in SE annotation. As in Becker et al. (2016), the STATE vs. GENERIC distinction is responsible for most of the disagreements (39%).

[2] The transcripts are available for download from the open data website of the Bundestag.

[3] The data, datasheet and annotation guidelines (partly in German) are available from our GitHub repository here.

[4] Available from the SE project website.

[5] We used the German `de_core_news_sm` model.

Corpus statistics 19,676 SE instances have been labeled by the two annotators. The most frequent SE type in GePaDeSE is STATE (12,321), followed by GENERIC sentences (2,360) and EVENT (2,145), showing an imbalanced SE type distribution in the corpus.[6]

This is different from the theoretical assumptions of Smith (2005) for argumentative text genres and also deviates from the empirical observations in Becker et al. (2016) for argumentative microtexts, where the majority of SE types are general statives. It indicates that the parliamentary debates cannot be described as "100%" argumentative but also serve other functions, such as providing information, or being used for self-representation through highlighting one's own achievements and by attacking the political opponent (also see Kondratenko et al. (2020) on the different functions of communication in parliamentary settings).

4. Experiments

Based on the GePaDeSE corpus, we present results for four language models[7] and two baselines to examine how well SE types in German parliamentary debates can be labeled automatically.

4.1. Setup

Similar to previous work (Plank, 2022; Cabitza et al., 2023), we acknowledge annotator disagreement and the subjectivity inherent to the task. Therefore, we cast SE classification as a multi-class, *multi-label* sequence classification task, where each instance receives one SE label in case of annotator agreement and two in case of disagreement. A label is predicted if its probability exceeds 0.5.

As instances are at the paragraph level, with one paragraph potentially containing multiple SE, we use `[FOCUS]` tokens[8] around the SE-evoking verb to mark the input position to attend to.

4.2. Models

Baselines As two naive baselines, we include a stratified random classifier, which predicts a randomly sampled label according to the class distribution, and a majority baseline, which always predicts the label of the majority class STATE.

GBERT We use the pre-trained German BERT model $\text{GBERT}_{\text{Large}}$ (henceforth, GBERT) which employs the original BERT architecture and has been trained exclusively on German data (Chan et al., 2020). GBERT comprises 335M parameters. We fine-tune[9] GBERT for our SE classification task and train a classification head on top of the pre-trained model to perform multi-label classification.

EuroBERT To examine the language-specific requirements of the SE classification task, we include EuroBERT (Boizard et al., 2025), a multilingual encoder model, covering 15 mostly European languages, with monolingual German corpora making up 6% of the overall pre-training data. We use the 610M variant of EuroBERT. Fine-tuning and evaluation are analogously performed to GBERT.

Llama 3.2 1B To investigate to what extend architectural differences impact the SE classification task, we include two Llama models employing a decoder-only architecture. The first one is Llama 3.2 (Grattafiori et al., 2024) which officially supports 8 languages, including German, yet exhibits a heavy English bias. Although exact numbers are not reported, Llama 3.2 builds on Llama 3.1, so a similar language distribution can be assumed; the latter contains 8% multilingual tokens. In order to keep scores comparable between models with different architectures, we fine-tune Llama 3.2 and train a classification head instead of applying in-context learning. To reduce computational cost, we use PEFT methods (Mangrulkar et al., 2022) and apply QLoRA (Dettmers et al., 2023; Hu et al., 2021).

LLäMmlein 1B Similar to the encoder models, we include a German-only decoder model to investigate language-specific requirements of the task. Specifically, we use LLäMmlein 1B (Pfister et al., 2025), a Llama-based model exclusively trained on German data. Fine-tuning is conducted analogously to Llama 3.2.

4.3. Evaluation Metrics

We report precision, recall and F1 (micro and macro) for the SE classification task. To account for label disagreement among annotators and to assess whether models produce similar uncertainties as the annotators, we additionally report the Jaccard index and the MASI score (Passonneau, 2006).[10] Both MASI and Jaccard compare sets

[6] See Table 5 in the Appendix for the overall SE type distribution.

[7] See Tables 6 and 7 for hyperparameter details and Table 8 for general model details.

[8] Alternatively, we used `token_type_ids` in preliminary experiments, yet these are specific to BERT models.

[9] Unless stated otherwise, fine-tuning denotes *full* fine-tuning, meaning that all model parameters are updated.

[10] While MASI is originally proposed as a distance metric, we reformulate it as a similarity metric to adequately compare it against Jaccard.

Model	Prec.	Rec.	F1 (micro)	F1 (macro)	Jaccard	MASI
majority	0.09	0.12	0.63	0.10	0.62	0.59
random	0.16	0.17	0.45	0.16	0.40	0.34
GBERT	**0.82** (±0.01)	**0.80** (±0.02)	**0.84** (±0.00)	**0.80** (±0.01)	**0.83** (±0.00)	**0.78** (±0.00)
EuroBERT	0.80 (±0.01)	0.73 (±0.02)	0.81 (±0.01)	0.75 (±0.01)	0.79 (±0.01)	0.74 (±0.01)
LLäMmlein	0.79 (±0.01)	0.74 (±0.01)	0.81 (±0.01)	0.76 (±0.01)	0.79 (±0.01)	0.74 (±0.01)
Llama 3.2	0.74 (±0.01)	0.62 (±0.02)	0.76 (±0.01)	0.66 (±0.01)	0.73 (±0.01)	0.68 (±0.01)

Table 2: Results for SitEnt prediction. Bold numbers indicate best scores across models. Scores are averaged over five random seeds. Standard deviation in parentheses. Prec. and Rec. report macro averaged precision and recall scores, respectively.

Class	Precision	Recall	F1
Generic	0.72 ±0.03	0.74 ±0.02	0.73 ±0.01
Generalizing	0.65 ±0.04	0.54 ±0.08	0.59 ±0.04
Event	0.84 ±0.02	0.72 ±0.06	0.77 ±0.03
State	0.94 ±0.01	0.91 ±0.01	0.92 ±0.00
Report	0.90 ±0.05	0.76 ±0.05	0.82 ±0.03
Question	**1.00** ±0.00	**0.94** ±0.02	**0.97** ±0.01
Imperative	0.93 ±0.02	**0.94** ±0.02	0.94 ±0.02
Event-perfect-state	0.61 ±0.02	0.82 ±0.06	0.70 ±0.01

Table 3: Precision, Recall and F1 scores (± standard deviation) per class for the SitEnt prediction task. Bold numbers indicate best scores across classes. Model: GBERT.

based on label overlap. While Jaccard only considers raw intersection over union, MASI extends Jaccard by incorporating subset relations via weighted penalties, with subset overlaps being less penalized than non-subset overlaps.[11] Since MASI builds on the Jaccard by adding penalty terms, it follows from their formulations that MASI $\leq$ Jaccard.

4.4. Results

Table 2 shows the results for the different models evaluated on the test set. All models are able to effectively learn SE as they all substantially outperform the naive baselines. Obtained Macro-F1 scores range from 0.66 (Llama 3.2) to 0.80 for our best performing model GBERT. This indicates that SE carry a learnable signal and the model were able to learn features that can discriminate between different SE classes.

We observe that monolingual, German-only models consistently surpass multilingual[12] models as GBERT and LläMmlein outperform multilingual variants EuroBERT and Llama 3.2 by +0.05 and +0.10, respectively (measured in absolute Macro-F1 gains). Notably, GBERT (335M) has roughly half as many parameters as EuroBERT (610M). This suggests that SE classification requires fine-grained language-specific understanding which is still limited in larger, multilingual models.

Regarding different model architectures, we find smaller, encoder-only models to outperform larger, decoder-only models. This aligns with previous work (Upravitelev et al., 2025; Bucher and Martini, 2024) showing that encoder-models benefit from bidirectional, sequence-level representations for classification tasks compared to unidirectional representations of decoder-only models optimized for next-token prediction. Future work should explore whether different prompting strategies with larger decoder-only models can close this gap, while carefully weighing the substantially higher computational costs and reduced explainability they entail.

In order to assess model performance in greater detail, we report the scores obtained for the individual SE classes for the best-performing model (GBERT) in Table 3. Despite being minority classes, Question and Imperative show the overall best scores, achieving F1 scores of 0.97 and 0.94, respectively. We mainly attribute this to salient surface-level cues strongly associated with these classes (question and exclamation marks) that models may simply exploit for their predictions.

Finally, we emphasize the subjectivity inherent in the SE classification task and report both Jaccard and MASI scores to account for label disagreement among annotators. As Table 2 shows, the MASI score closely aligns with the Jaccard index, indicating that model predictions largely overlap with, or constitute subsets of, the gold labels. Consequently, most misclassifications still include human-plausible labels, underscoring the reliability of our SE classifier.

To substantiate this, we analyzed model errors of our best-performing model (GBERT): Only 9% of test instances were overpredictions, with fewer than 1% including more than two predicted labels. The most frequent case of subset overprediction involved the combination Generic, State when

[11] See Appendix D for formulas and walkthrough examples for computing MASI and micro F1 in our multi-label setting.

[12] For readability, Llama 3.2 is considered multilingual despite its heavy English bias.

the gold label was STATE (33% of these cases). Conversely, 13% of instances were underpredictions, with fewer than 1% producing empty predictions. The most frequent case here involved missing GENERIC in instances labeled GENERIC, STATE in the gold data (22%).

Overall, these patterns show that most model errors reflect partial rather than divergent label assignments. The near-zero rates of excessive or empty predictions further confirm that the model's predictions mostly remain within the allowed range of one to two labels per instance.

Taken together, our results demonstrate the effectiveness of the multi-label modeling approach: SE types can be learned robustly, as evidenced by consistently high performance across models, and even misclassifications tend to reflect linguistically plausible alternatives.

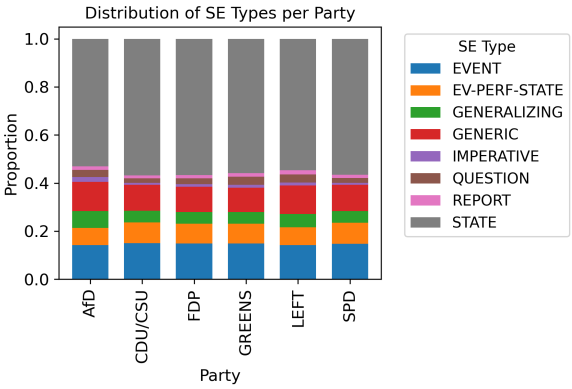


Figure 1: SE type distribution across parties for BT 16-21. Parties differ significantly ($p < .001$), though the effect size (Cramér's $V = .03$) indicates negligible practical differences.

5. Analysis

We now present an illustrative analysis of how SE types manifest in political discourse, using a corpus of parliamentary debates from the German Bundestag (BT) covering speeches from 2005 to 2025. This analysis is intended to complement our two primary contributions–the GePaDeSE corpus and the SE classifiers–by demonstrating their potential for research in political science. It is exploratory in nature and methodologically limited, serving primarily as a starting point for future, more comprehensive analyses.

Given their linguistic properties, we do not expect to have certain SE types being strongly associated with a specific party (see Figure 1 exemplifying the SE distribution across parties), but rather to observe differences in the use of *specific* SE types in political discourse. Hence, we view SE as linguistic filters. In the following, we exemplify this function through the SE type EVENT.

Our analysis is divided into two parts. In Part I (§5.3), we demonstrate the effectiveness of SE as a linguistic filter by comparing EVENTS against other SE types. Based on the results, we investigate in Part II (§5.4) how EVENTS manifest in political discourse and analyze the use of EVENTS across parties and over time. We conclude by proposing concrete research directions in the field of political science using SE in §5.5. We first outline the data and overall methodology before turning to specific analyses and results for each analysis part.

5.1. Data

Corpus The analyses are based on an extensive, unlabeled corpus of parliamentary debates from the German Bundestag, covering 274,876 speeches from the last six legislative periods (BT 16-21), ranging from 2005 to 2025. We use the raw transcripts from the GermaParl Corpus of Plenary Protocols (Blaette, 2017), which span the period from 1949 to 2023, and augment the data with more recent transcripts downloaded from the Open Data website of the German Bundestag.[13]

Since parts of the BT 19 subcorpus were used to train our classifier, we exclude those speeches from subsequent analyses to avoid potential bias from instances seen during training.

Preprocessing To apply our SE classifier to unlabeled data, instances must be enriched with `[FOCUS]` tokens wrapped around the main verbs. Main verbs are automatically identified using spaCy.[14]

5.2. Classifier Application & Filtering

We apply our best-performing classifier to obtain SE predictions for all instances in the BT 16-21 corpus. Since the classifier is trained in a multi-label fashion, each instance can be assigned to multiple SE types. We therefore perform a loose filtering step and extract all instances for which the model predicts EVENT, possibly alongside other SE.

5.3. Part I: SE as Linguistic Filters

Linguistically, EVENTS can be broadly characterized by a *specific* main referent and by a *dynamic* main verb. Consider the following EVENT example from our corpus, where "we" is the main referent and "dismantling" the dynamic verb:

(3) Wir bauen energisch die Bürokratie ab.
We build energetically the bureaucracy up.
'We are energetically dismantling bureaucracy.'

[13] https://www.bundestag.de/services/opendata

[14] We use the German model `de_core_news_sm`.

Based on the properties of EVENTS, we now assess whether SE work as linguistic filters by comparing the proportions of dynamic verbs and specific main referents in EVENTS against other SE types. Proportions are expected to be highest for EVENTS.

Setting We compare EVENT clauses with the two other major SE types in our corpus, STATE and GENERIC. To approximate the proportions of dynamic verbs and specific main referents, we employ manually constructed lexicons.[15] Dynamic verbs are identified by first compiling a list of common German stative verbs (e.g., *sein*, *haben*) and treating all non-stative verbs as dynamic. To determine specific main referents, we restrict our lexicon to political actors occurring in subject position, including terms such as *federal chancellor* and *federal government* as well as pronouns like *we* and *I*. Again, we use spaCy to automatically identify subjects. To reduce noise, we restrict our corpus to instances where spaCy identified a single subject for the SE-evoking verb.

Results Across all EVENTS in BT 16-21, 83% contain a dynamic main verb, and among these, approximately 48% have a political actor as the subject. In comparison, STATE clauses contain 57% dynamic verbs with 30% political subjects, while GENERIC clauses include 61% dynamic verbs but only 9% political subjects.

These results indicate that different SE types are used to express distinct semantics. In particular, EVENTS in parliamentary debates are largely action-denoting and frequently describe actions performed by identifiable political actors, in the following referred to as *actor-action pairs*. Thus, our analysis suggests that filtering for the SE type EVENT offers a linguistically grounded approximation for identifying actor-action pairs in political discourse.

5.4. Part II: EVENTS in Parliamentary Debates

Building on the previous analysis, which demonstrated that EVENTS can serve as a linguistic approximation for extracting actor-action pairs, we now examine how such EVENTS manifest in political discourse. Specifically, we suggest that they function as instruments of *self-affirmation* – highlighting the speaker's or party's own achievements – or of *other-critique* – attributing responsibility or blame to political opponents.

This analysis is grounded in our political actor lexicon, whose entries can be broadly categorized into two groups: self-references (actors such as *I*, *we*, or the speaker's party name) and other-references (actors such as *you*, *they*, references to opposing parties or proper names). Linking these reference types with their discourse function – self-affirmation and other-critique, respectively – allows us to approximate how political actors use EVENTS to either affirm their own actions or criticize those of others. We analyze the proportions of both discourse functions across parties and over time.

Setting To estimate the proportions of *self-affirmation* and *other-critique*, we combine our political actor lexicon with rule-based heuristics indicating whether a referenced actor in a given actor-action pair is affirmed, criticized, or neutral. For example, actors beginning with *Bundes** (federal) are treated as potential targets of critique only if the speaker's party is not part of the government during the respective legislative period.

For the temporal analysis, we divide our corpus into six-month bins. For each bin, we calculate the proportion of EVENTS expressing *other-critique* or *self-affirmation* relative to the total number of EVENTS. This procedure is applied separately for each party.

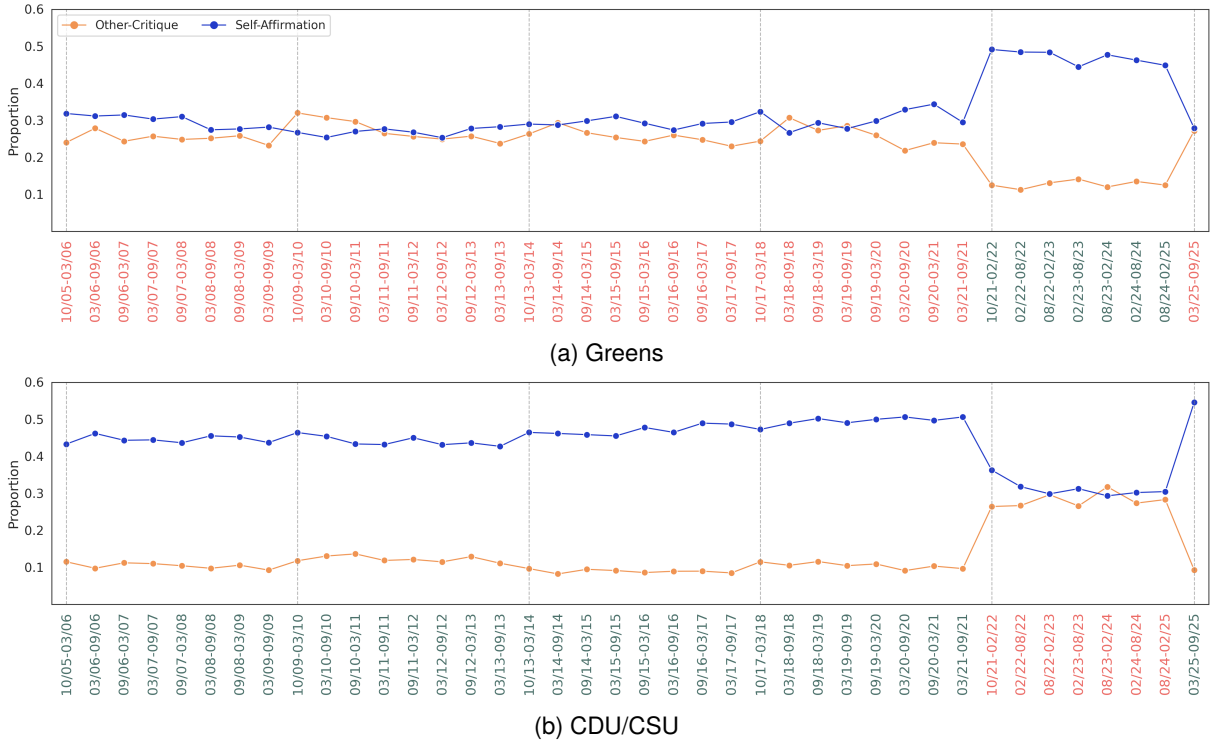
Results Figure 2 shows the proportions of *other-critique* and *self-affirmation* for the Greens and the CDU/CSU[16] across the last six legislative periods. During this time, the two parties have alternated between government and opposition: when the CDU/CSU governed, the Greens were in opposition, and vice versa.

Clear patterns emerge: when in opposition, parties employ EVENTS more frequently as instruments of *other-critique* than when in government. Conversely, when in government, the same parties use EVENTS more often for *self-affirmation*. For example, the CDU/CSU exhibits an *other-critique* share of 28% while in opposition (BT 20) and 10% while in government (avg. BTs 16-19, 21). The Greens show a comparable pattern, with 26% in opposition (BTs 16-19, 21) and 13% in government (BT 20). This pattern generalizes across parties represented in the German Bundestag over the past 20 years that have participated in government at least once (SPD and FDP), as shown in Appendix F. In contrast, parties that have remained in opposition exhibit less clearly discernible patterns. This applies to The Left throughout the entire period and to the AfD since its entry into the Bundestag in 2017.

Our results indicate that the discourse function of EVENTS is systematically linked to a party's parliamentary role: government parties predominantly engage in *self-affirmation*, whereas opposition parties rely more on *other-critique*.

[15] Available at our GitHub repo here.

[16] The CDU and CSU form a joint parliamentary group in the German Bundestag. Therefore, we treat them as



(a) Greens

(b) CDU/CSU

Figure 2: Proportion of other-critique and self-affirmation for the Greens and CDU/CSU across the last six legislative periods. Green labels on the x-axis indicate times in government, red labels indicate times in opposition. Vertical lines mark the start of each new legislative period.

ID	Subject	Subtree (truncated)
12	we	*We are increasing investments in the rail network.*
12	we	*We are strengthening the social market economy.*
12	we	*We are investing in a modern state and a digital administration.*
69	you	*You are cutting international climate financing.*
69	you	*You are jeopardizing Europe's 2040 climate target.*
69	Vice Chancellor Klingbeil	*Vice Chancellor Klingbeil is funding climate-damaging subsidies.*

Table 4: Examples of EVENT subtrees in BT 21 for two cluster IDs, illustrating the *self-referential* vs. *other-referential* pattern in actor-action pairs. SE-evoking verbs are underlined.

Qualitative Analysis To complement the quantitative results, we qualitatively examine semantic patterns in EVENTS across different parliamentary roles through clustering. To obtain clusters, we represent each EVENT by the syntactic subtree[17] of its SE-evoking verb,[18] which we encode using the multilingual model EmbeddingGemma (Schechter Vera et al., 2025). We then apply UMAP (McInnes et al., 2018) for dimensionality reduction and HDBSCAN (McInnes et al., 2017) to group semantically similar EVENTS. Unclustered instances are reassigned to the nearest cluster centroid based on cosine similarity to ensure broad coverage.[19]

We subsequently focus on clusters dominated by either government or opposition parties to illustrate typical semantic patterns of EVENT usage across parliamentary roles, concentrating on BT 21. Table 4 presents examples for these clusters. In Cluster 12, government parties emphasize their policy objectives in the context of economic growth, whereas in Cluster 69, opposition parties criticize the government for insufficient action in the climate debate. These examples illustrate how EVENT clauses are employed to express self-affirmation and other-critique in concrete policy areas.

Taken together, in Part I, we demonstrated that SE can serve as an effective linguistic filter and that EVENTS in the parliamentary discourse predominantly yield actor-action pairs. In Part II, we extended this analysis by distinguishing these actor-action pairs according to their discourse function,

a single party.

[17] Extracted via spaCy's `subtree` method.

[18] For infinite verbs, we extract the subtree of the governing verb.

[19] See Table 10 in the Appendix for further details.

showing that government parties primarily use EVENTS for self-affirmation, whereas opposition parties tend to employ them for other-critique.

5.5. Potential Use Cases

Our analyses show that our GePaDeSE corpus and the SE classifier provide a solid foundation for addressing research questions in political science. The EVENT-linked *other-critique* and *self-affirmation* patterns strongly resonate with the *blame avoidance theory* (Weaver, 1986) which posits that political actors seek to deflect responsibility for unpopular outcomes while claiming credit for success. GENERICS and GENERALIZING sentences, on the other hand, may be used to identify and examine stereotypes in political discourse.

6. Conclusions

In this paper, we presented GePaDeSE, a new resource annotated with SE types, modeling clause-level aspect in German parliamentary debates with more than 19k manually annotated SE types. We demonstrated that SE types can be effectively learned by fine-tuning BERT- and Llama-style models, yielding F1 scores over 0.8. Moreover, we showed that model misclassifications mostly constitute linguistically plausible alternatives to the gold label, highlighting the subjectivity inherent to the task. Finally, we illustrated how SE can serve as a linguistic filter to support political text analysis at scale, finding for EVENT clauses on a large-scale corpus of German parliamentary debates that government parties tend to use them to highlight achievements, whereas opposition parties employ them primarily to criticize government actions. All resources described in the paper are made available to the research community.

7. Limitations

We acknowledge that automatic evaluation metrics for multi-label classification offer only a limited view of model performance in subjective tasks like SE classification. Although we included metrics that account for subset relations and conducted an error analysis of model misclassifications, more nuanced analyses are needed to assess how well model uncertainty aligns with annotator disagreement.

In addition, our experiments did not include large-scale LLMs due to computational constraints and the specific focus of this study. Future work should therefore explore various prompting strategies and model scales.

Regarding our corpus analyses, we particularly highlight the limitations of our lexicon-based approach for identifying dynamic verbs and political actors. Future work could employ NLP techniques such as NER and coreference resolution to more accurately capture political actors, complemented by targeted stance detection models or few-shot LLM approaches to more reliably identify evaluative language (instances of *self-affirmation* and *other-critique*).

Finally, the generalizability of our approach beyond parliamentary debates remains to be tested.

Acknowledgments

The work presented in this paper is funded by the German Research Foundation (DFG) under the UNCOVER project (PO1900/7-1 and RE3536/3-1).

We would like to thank the anonymous reviewers for their constructive feedback.

8. Bibliographical References

Malihe Alikhani and Matthew Stone. 2019. "caption" as a coherence relation: Evidence and implications. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, pages 58–67, Minneapolis, Minnesota. Association for Computational Linguistics.

Nicholas Asher. 1993. *Reference to Abstract Objects in Discourse*. Kluwer.

Maria Becker, Alexis Palmer, and Anette Frank. 2016. Argumentative texts and clause types. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 21–30, Berlin, Germany. Association for Computational Linguistics.

Maria Becker, Michael Staniek, Vivi Nastase, Alexis Palmer, and Anette Frank. 2017. Classifying semantic clause types: Modeling context and genre characteristics with recurrent neural networks and attention. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 230–240, Vancouver, Canada. Association for Computational Linguistics.

Nicolas Boizard, Hippolyte Gisserot-Boukhlef, Duarte M. Alves, André Martins, Ayoub Hammal, Caio Corro, Céline Hudelot, Emmanuel Malherbe, Etienne Malaboeuf, Fanny Jourdan, Gabriel Hautreux, João Alves, Kevin El-Haddad, Manuel Faysse, Maxime Peyrard, Nuno M. Guerreiro, Patrick Fernandes, Ricardo Rei, and Pierre Colombo. 2025. Eurobert: Scaling multilingual encoders for european languages.

Anne Bosse. 2024. Stereotyping and generics. *Inquiry: An Interdisciplinary Journal of Philosophy*, 67(10):3876–3892.

Peter F. Brown, Peter V. Desouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Martin Juan José Bucher and Marco Martini. 2024. Fine-tuned 'small' llms (still) significantly outperform zero-shot generative ai models in text classification.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model.

Zeyu Dai and Ruihong Huang. 2018. Building context-aware clause representations for situation entity type classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3305–3315, Brussels, Belgium. Association for Computational Linguistics.

Ilse Depraetere. 1995. On the necessity of distinguishing between (un)boundedness and (a)telicity. *Linguistics and Philosophy*, 18(1):1–19.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

David R. Dowty. 1979. *Word meaning and Montague grammar: the semantics of verbs and times in generative semantics and in Montague's PTQ*. Number v. 7 in Synthese language library. D. Reidel Pub. Co, Dordrecht ; Boston.

Markus Egg, Helena Prepens, and Will Roberts. 2019. Annotation and automatic classification of aspectual categories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3335–3341, Florence, Italy. Association for Computational Linguistics.

W. Nelson Francis and Henry Kučera. 1979. *Brown corpus manual*.

Annemarie Friedrich and Damyana Gateva. 2017. Classification of telicity using cross-linguistic annotation projection. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2559–2565, Copenhagen, Denmark. Association for Computational Linguistics.

Annemarie Friedrich and Alexis Palmer. 2014a. Automatic prediction of aspectual class of verbs in context. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 517–523, Baltimore, Maryland. Association for Computational Linguistics.

Annemarie Friedrich and Alexis Palmer. 2014b. Situation entity annotation. In *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*, pages 149–158, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Annemarie Friedrich, Alexis Palmer, and Manfred Pinkal. 2016. Situation entity types: automatic classification of clause-level aspect. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1757–1768, Berlin, Germany. Association for Computational Linguistics.

Annemarie Friedrich, Nianwen Xue, and Alexis Palmer. 2023. A kind introduction to lexical and grammatical aspect, with a survey of computational approaches. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 599–622, Dubrovnik, Croatia. Association for Computational Linguistics.

William Gantt, Lelia Glass, and Aaron Steven White. 2022. Decomposing and recomposing event structure. *Transactions of the Association for Computational Linguistics*, 10:17–34.

Barts Geurts. 1985. Generics. *Journal of Semantics*, 3(4):247–255.

Venkata Govindarajan, Benjamin Van Durme, and Aaron Steven White. 2019. Decomposing generalization: Models of generic, habitual, and episodic statements. *Transactions of the Association for Computational Linguistics*, 7:501–517.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, ..., and Zhiyu Ma. 2024. The llama 3 herd of models.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Liesa Heuschkel. 2016. Automatic classification of lexical aspectual class using distributional and rule-based methods. Master's thesis, Saarland University.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Joan B. Hooper. 1975. On assertive predicates. *Syntax and Semantics*, 4:91–124.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Nancy Ide, Collin Baker, Christiane Fellbaum, Charles Fillmore, and Rebecca Passonneau. 2008. MASC: the manually annotated sub-corpus of American English. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Nancy Ide, Collin Baker, Christiane Fellbaum, and Rebecca Passonneau. 2010. The manually annotated sub-corpus: A community resource for and by the people. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 68–73, Uppsala, Sweden. Association for Computational Linguistics.

Lauri Karttunen. 1971. Implicative verbs. *Language*, pages 340–358.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics. Event Title: The 27th International Conference on Computational Linguistics (COLING 2018).

Thomas Kober, Malihe Alikhani, Matthew Stone, and Mark Steedman. 2020. Aspectuality across genre: A distributional semantics approach. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4546–4562, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Natalia V. Kondratenko, Anastasiia A. Kiselova, and Liubov V. Zavalska. 2020. Strategies and tactics of communication in parliamentary discourse. *studies about languages*, 36:17–29.

Manfred Krifka, Francis J Pelletier, Gregory N Carlson, Gennaro Chierchia, Godehard Link, and Alice ter Meulen. 1995. Genericity: An introduction. In *The Generic Book*, pages 1–24. University of Chicago Press, Chicago and London.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Sarah-Jane Leslie. 2014. Carving up the social world with generics. *Oxford studies in experimental philosophy*, 1:208–232.

Sharid Loáiciga and Cristina Grisot. 2016. Predicting and using a pragmatic component of lexical aspect of simple past verbal tenses for improving English-to-French machine translation. *Linguistic Issues in Language Technology*, 13(3).

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. PEFT: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Kleio-Isidora Mavridou. 2016. Situation entity types: a cross-linguistic corpus study and a comparison of automatic classifiers. Master's thesis, Institute for Computational Linguistics at Saarland University, Germany.

Kleio-Isidora Mavridou, Annemarie Friedrich, Melissa Peate Sørensen, Alexis Palmer, and Manfred Pinkal. 2015. Linking discourse modes and situation entity types in a cross-linguistic corpus study. In *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics*, pages 12–21, Lisbon, Portugal. Association for Computational Linguistics.

Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

Gustavo Novoa, Margaret Echelbarger, Andrew Gelman, and Susan A. Gelman. 2023. Generically partisan: Polarization in political communication. *Proceedings of the National Academy of Sciences*, 120(47):e2309361120.

Alexis Palmer, Elias Ponvert, Jason Baldridge, and Carlota Smith. 2007. A sequencing model for situation entity classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 896–903, Prague, Czech Republic. Association for Computational Linguistics.

Alexis Palmer and Caroline Sporleder. 2009. Situation entities and genre distinctions in the Penn Discourse TreeBank (abstract).

Rebecca Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Roma Patel and Ellie Pavlick. 2021. "Was it "stated" or was it "claimed"?: How linguistic bias affects generative language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10080–10095, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jan Pfister, Julia Wunderle, and Andreas Hotho. 2025. LLäMmlein: Transparent, compact and competitive German-only language models from scratch. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2227–2246, Vienna, Austria. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Günter Radden. 2009. Generic reference in English: A metonymic and conceptual blending analysis. In Klaus-Uwe Panther, Linda L. Thornburg, and Antonio Barcelona, editors, *Metonymy and Metaphor in Grammar*, pages 199–228. John Benjamins Publishing Company.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Tom Ralston. 2024. Reconceptualising the psychological theory of generics. *Philosophical Studies*, 181:2973–2995.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659, Sofia, Bulgaria. Association for Computational Linguistics.

Ines Rehbein, Josef Ruppenhofer, Annelen Brunner, and Simone Paolo Ponzetto. 2024. Out of the mouths of MPs: Speaker attribution in parliamentary debates. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12553–12563, Torino, Italia. ELRA and ICCL.

Mehdi Rezaee, Kasra Darvish, Gaoussou Youssouf Kebe, and Francis Ferraro. 2021. Discriminative and generative transformer-based models for situation entity classification.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, ..., and Mojtaba Seyedhosseini. 2025. Embeddinggemma: Powerful and lightweight text representations.

Carlota S. Smith. 1983. A theory of aspectual choice. *Language*, 59(3):479–501.

Carlota S. Smith. 1991. *The Parameter of Aspect*. Kluwer.

Carlota S. Smith. 2003. *Modes of Discourse: The Local Structure of Texts*. Cambridge Studies in Linguistics. Cambridge University Press.

Carlota S Smith. 2005. *Aspectual entities and tense in discourse*, pages 223–237. Springer.

Max Upravitelev, Nicolau Duran-Silva, Christian Woerle, Giuseppe Guarino, Salar Mohtaj, Jing Yang, Veronika Solopova, and Vera Schmitt. 2025. Comparing LLMs and BERT-based classifiers for resource-sensitive claim verification in social media. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*, pages 281–287, Vienna, Austria. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Zeno Vendler. 1967. *Linguistics in Philosophy*. Cornell University Press.

R. Kent Weaver. 1986. The politics of blame avoidance. *Journal of Public Policy*, 6(4):371–398.

9. Language Resource References

Andreas Blaette. 2017. *GermaParl. Corpus of Plenary Protocols of the German Bundestag*. PolMine. 2023.

A. SE Type Distribution

Entity	Situation Entities A1	A2	Avg.
State	12,987	11,655	12,321.0
Generic	2,012	2,708	2,360.0
Event	1,406	2,885	2,145.5
Generalizing	942	1,062	1,002.0
Event-Perfect-State	1,286	259	772.5
Question	447	444	445.5
Imperative	345	335	340.0
Report	251	328	289.5
Total	19,676	19,676	19,676
Entity	**Abstract Entities** A1	A2	Avg.
Proposition	315	222	268.5
Fact	301	123	212.0

Table 5: Distribution of SE types in the GePaDeSE corpus. The columns A1, A2 show the number of instances for annotator 1 and 2, respectively. The last column displays the average number of instances for each SE type.

B. Training Details & Hyperparameters

B.1. General

Hyper-parameter	GBERT	EuroBERT	Llama 3.2 / LläMmlein
n epochs	4	4	4
Batch size	32	16	16
Learning rate	1e-5	5e-5	2e-4
Weight decay	0.01	0.01	0.01
Warm-up ratio	0.1	0.1	0.1
Optimizer	AdamW	AdamW	AdamW

Table 6: Training hyperparameters for different models for the SE classification task. Unspecified hyperparameters were set to the default values as provided in the `transformers` library. All models are evaluated using the best checkpoint according to the evaluation loss.

B.2. QLoRA

Hyperparameter	Value
Rank (r)	16
α (scaling factor)	8
Dropout	0.05
Target modules	[q_proj, k_proj, v_proj, o_proj]

Table 7: QLoRA hyperparameters used for Llama 3.2 and LläMmlein. 4-bit quantization is applied using NF4 with double quantization.

	GBERT	EuroBERT	Llama 3.2	LLäMmlein
Architecture	Encoder-only	Encoder-only	Decoder-only	Decoder-only
Model Size (# parameters)	337M	610M	1B	1B
Pre-training data size (# tokens)	163.5GB*	5T	15T	3T
German data (%)	100	6	<8**	100
Model Checkpoint	https://huggingface.co/deepset/gbert-large	https://huggingface.co/EuroBERT	https://huggingface.co/meta-llama/Llama-3.2-1B	https://huggingface.co/LSX-UniWue/LLaMmlein_1B

Table 8: Comparison between all models used for training and evaluation. *Only the raw size in GB is provided. **8% correspond to the amount of multilingual tokens in Llama 3.1.

C. Models

Table 8 provides details about all models used for training and evaluating our SE classification task.

D. Metrics

D.1. Micro F1

Formulas To compute Micro F1 in our multi-label setting, we binarize and flatten all predictions and gold labels across all samples and labels, treating the task as a single binary classification problem over all (sample, label) pairs.

Specifically, given binarized predictions $y_{\text{pred}} \in \{0,1\}^{N \times L}$ and gold labels $y_{\text{true}} \in \{0,1\}^{N \times L}$, where N denotes the number of instances and L the number of possible labels, i.e., the different SE types, the counts of true positives (TP), false positives (FP), and false negatives (FN) are accumulated over all N samples and L labels:

$$\text{TP} = \sum_{i,l} [y_{\text{true}}^{(i,l)} = 1 \wedge y_{\text{pred}}^{(i,l)} = 1],$$

$$\text{FP} = \sum_{i,l} [y_{\text{true}}^{(i,l)} = 0 \wedge y_{\text{pred}}^{(i,l)} = 1],$$

$$\text{FN} = \sum_{i,l} [y_{\text{true}}^{(i,l)} = 1 \wedge y_{\text{pred}}^{(i,l)} = 0].$$

Micro F1 is then defined as:

$$F1_{\text{micro}} = \frac{2\,\text{TP}}{2\,\text{TP} + \text{FP} + \text{FN}}$$

Unlike macro-averaging, which averages per-label F1 scores, micro-averaging treats every (instance, label) pair equally, making it effective when overall prediction quality is of interest.

Example Assume our SE label set with $L = 8$. For two instances ($N = 2$), we have:

$$y_{true} = [\{\text{Event, Event-Perfect-State}\}, \{\text{State}\}],$$
$$y_{pred} = [\{\text{Event}\}, \{\text{State}\}]$$

Binarized:

$$y_{true} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$
$$y_{pred} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

Across all $N \times L = 16$ label decisions, we obtain TP $= 2$, FP $= 0$, FN $= 1$, resulting in

$$F1_{\text{micro}} = \frac{2 \times 2}{2 \times 2 + 0 + 1} = 0.8.$$

D.2. Jaccard & MASI

Formulas Let A and B refer to the set of gold labels and model predictions of a single instance, respectively. A is aggregated over two annotators, thus, A is of length 2 if the annotators disagree, i.e., provide different labels, and of length 1 if the annotators agreee, i.e., provide the same label. Using the sets A and B, Jaccard and MASI are computed as follows:

$$\text{Jaccard: } J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

$$\text{MASI: } MASI(A,B) = M(A,B) \cdot J(A,B)$$

$$M(A,B) = \begin{cases} 1 & \text{if } A = B \\ \frac{2}{3} & \text{if } A \subset B \text{ or } B \subset A \\ \frac{1}{3} & \text{if } A \cap B \neq \emptyset \text{ and neither is a subset} \\ 0 & \text{if } A \cap B = \emptyset \end{cases}$$

Thus, MASI is stricter than Jaccard and penalizes non-subset overlaps more than subset overlaps. This property is desired, as illustrated in the following example.

Gold	Generic, Generalizing	
Pred 1	Generic, Generalizing	$M = 1$
Pred 2	Generic	$M = \frac{2}{3}$
Pred 3	Generic, State	$M = \frac{1}{3}$
Pred 4	State	$M = 0$

Table 9: Example to illustrate MASI factor.

Example MASI penalizes *Pred 3* more than *Pred 2* as the former is no subset of *Gold*. This is desired as we consider *Pred 2* to lie closer to the *Gold* as it only includes human-probable labels, whereas *Pred 3* includes labels diverging from the annotation.

In general, the formulas for Jaccard and MASI entail that $MASI(A, B) \leq J(A, B)$. The closer MASI lies to Jaccard, the better as this implies that model predictions contain less non-subset relations.

Note MASI's subset relations are symmetric, meaning that for two given label sets A and B, $A \subset B$ and $B \subset A$ are treated equally. Let A and B again refer to the gold label set and model prediction set of a single instance, respectively, then under-predictions (less predicted labels than gold labels: $B \subset A$) and over-predictions (more predicted labels than gold labels: $A \subset B$) are penalized equally.

In this study, we keep this property for simplicity, yet we note that weighing under- and over-predictions differently may be desired for a more nuanced model prediction analysis. E.g., penalizing under-predictions less than over-predictions may be a desired property when assuming that a single labels tend to sufficiently capture a potentially multi-label ground truth.

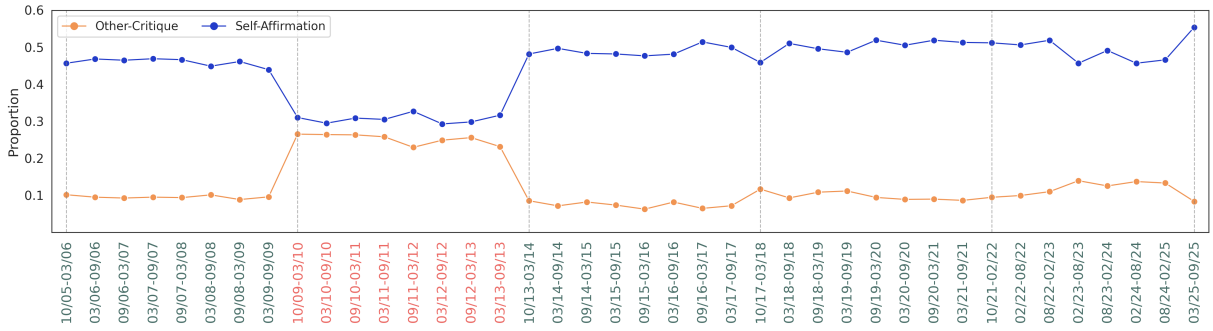
E. UMAP & Clustering

Hyperparameter	Value
UMAP	
Number of components	15
Number of neighbors	30
Minimum distance	0.00
Distance metric	cosine
Random state	42
HDBSCAN	
Minimum cluster size	10
Minimum samples	3
Reassignment threshold	0.8

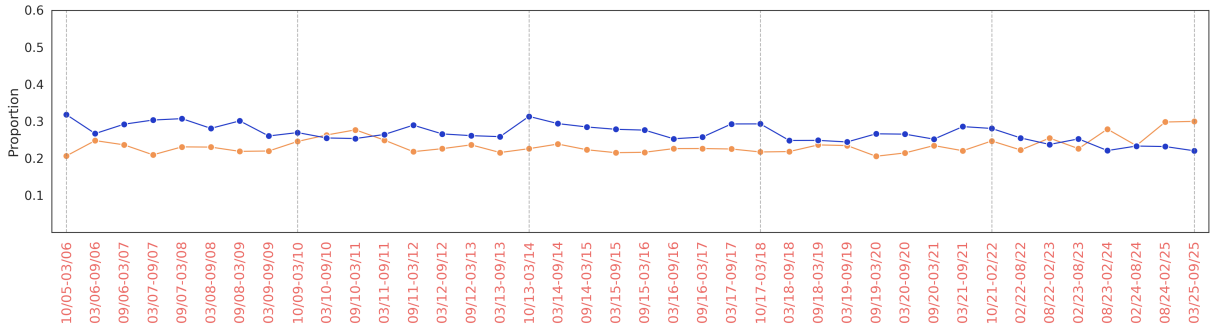
Table 10: Hyperparameter configuration for UMAP and HDBSCAN clustering.

F. Additional Plots

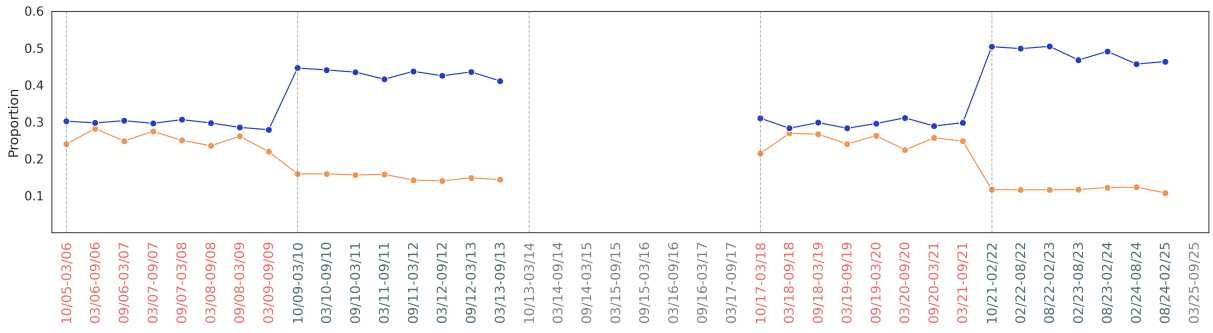
Figure 3 provides the proportions of self-affirmation and other-critique for SPD, The Left, FDP, and AfD.



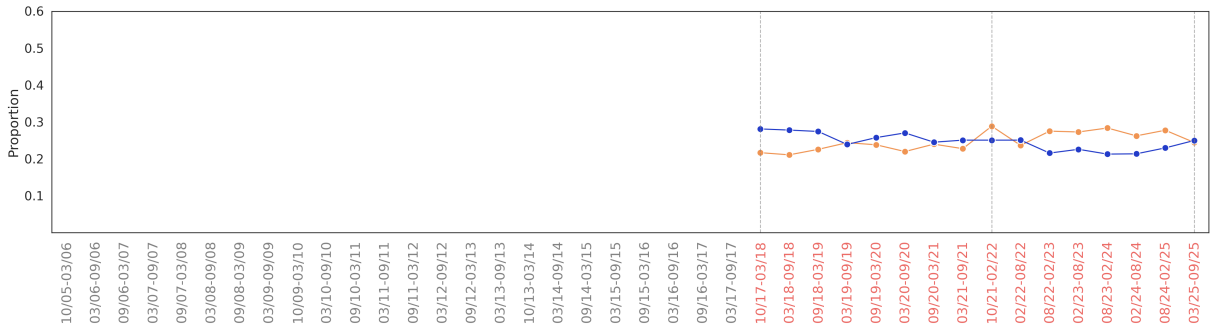
(a) SPD



(b) Left



(c) FDP



(d) AfD

Figure 3: Proportion of other-critique and self-affirmation for SPD, The Left, FDP, and AfD across the last six legislative periods. Green labels on the x-axis indicate times in government, red labels indicate times in opposition, gray labels indicate times not represented in the parliament. Vertical lines mark the start of each new legislative period.