Linguistic Features in German BERT: The Role of Morphology, Syntax, and Semantics in Multi-Class Text Classification

Henrike Beyer[1] **& Diego Frassinelli**[2]
[1]Centre for Argument Technology, University of Dundee, UK
[2]Center for Information and Language Processing, LMU Munich, Germany
2579207@dundee.ac.uk, frassinelli@cis.lmu.de

Abstract

Most studies on the linguistic information encoded by BERT primarily focus on English. Our study examines a monolingual German BERT model using a semantic classification task on newspaper articles, analysing the linguistic features influencing classification decisions through SHAP values. We use the TüBa-D/Z corpus, a resource with gold-standard annotations for a set of linguistic features, including POS, inflectional morphology, phrasal, clausal, and dependency structures. Semantic features of nouns are evaluated via the GermaNet ontology using shared hypernyms. Our results indicate that the features identified in English also affect classification in German but suggests important language- and task-specific features as well.

1 Introduction

Even today, with large language models (LLMs) like GPT-4 (OpenAI et al., 2023), Llama (Touvron et al., 2023), or Mistral (Jiang et al., 2023) representing the de facto state-of-the-art systems for most NLP tasks in English, the exploration of BERT-like models still provides extremely useful insights for low-resource and non-English scenarios (Brookshire and Reiter, 2024; Sivanaiah et al., 2024; Bressem et al., 2024), often offering more efficient and lightweight solutions.

Despite the extensive research evaluating the linguistic knowledge encoded in various English versions of BERT (Devlin et al., 2019) using interpretative methods like attention analysis (Jawahar et al., 2019; Goldberg, 2019; Kalouli et al., 2022), monolingual models pre-trained on languages other than English have received considerably less attention. Given that languages can differ quite significantly in their morphological, syntactic, and semantic complexity, it is crucial to identify which behaviours observed for English translate to other languages and which, instead, are language-specific.

For example, Jawahar et al. (2019) found that different types of linguistic information are distributed across different layers of English BERT; surface-level information like phrasal structure is processed by layers closer to the input, syntactic information by the middle layers, and semantic information by the layers closer to the output. The ability of BERT-like models to process syntactic information has been evaluated by assessing their performance on subject-verb agreement in English (Goldberg, 2019). More recently, Kalouli et al. (2022) assessed the quality of the semantic representations for general function words (e.g. negations, coordinating conjunctions, and quantification terms) in these models. Their findings suggest that BERT-like models struggle to accurately complete sentences based on these function words alone, often relying on other indicators, like Named Entities (NEs), for their predictions.

Our work investigates which morphological, syntactic, and semantic features are the strongest predictors in an eight-class text classification task for a German BERT model. Building on evidence from English, we analyse similarities and differences, particularly exploring how the richer inflectional morphology of German (Eisenberg, 2020) affects model performance. Former studies on German have analysed morphological or syntactic features separately (Zaczynska et al., 2020; Guarasci et al., 2021). Claeser (2022) conducts a study on the same corpus we use in this work, but considers only the influence of morphology with regard to CNNs. Our study covers a larger selection of morphological, syntactic, and semantic features and focuses on BERT.[1]

[1]Additional information for reproducibility can be found at: https://github.com/CoPsyN/ling-in-German-BERT

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop), pages 28–39

2 Materials and Methodology

2.1 Corpus Selection

For our analysis, we use the *Tübinger Baumdatenbank Deutsch/Zeitungskorpus* (TüBa-D/Z; Telljohann et al. (2004)). This corpus contains 3,642 newspaper articles (1,782,129 tokens; 153,990 types) from the German newspaper *Die Tageszeitung* and includes gold-standard annotations for inflectional morphology, part-of-speech tags, and syntax, along with automatically generated dependency structures. In addition, we use the semantic annotation layer by Claeser (2022), that categorizes the articles into eight topics with varying levels of coverage across the corpus: culture (kultur; 24%), politics (politik; 22%), miscellaneous (panorama; 17%), conflicts abroad (konfliktausland; 11%), economy (wirtschaft; 9%), crime (kriminalität; 8%), sport (sport; 5%), and environment (umwelt, 4%). This corpus offers consistent, rich, high-quality annotations on all layers. In addition, the text classification task covers a broad range of topics, allowing for good generalisability. To make the text compatible with BERT, we split the available text into 6,674 chunks of approximately 500 tokens each, ensuring that only full sentences are included. We ensured that there is no different in performance between chunks of the same text throughout our experiments.

2.2 Model Fine-Tuning

We fine-tune a monolingual BERT-base German-cased model (Chan et al., 2020) for 5 epochs with a batch-size of 8, a learning-rate of 5e-5 and AdamW as optimizer on the 8-way classification task mentioned above. We use a 10-fold-cross-validation design on the multi-class classification task described above. Due to the corpus' relatively small size and class imbalance, the fine-tuning of each fold is repeated five times with a new random initialization of the model. The classifier achieves an average F1-score of 0.72±0.01. Table 1 reports the accuracy scores per class, showing considerable differences (0.59 for "environment" vs. 0.90 for "sport"). Such differences should be considered when interpreting the results in the following analyses.

2.3 SHAP Value Calculation

To determine the importance of specific words in the classification task, we use the KernelSHAP algorithm (Lundberg and Lee, 2017) through the TransSHAP library (Kokalj et al., 2021). SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017) have been successfully applied to various NLP tasks (Chakravarthi et al., 2023; Jang et al., 2023; Tang et al., 2024; Rizinski et al., 2024). For our analysis, we calculate the SHAP values, which reflect the importance of each token in a text to the classification decision for the whole text.

class	percentage	accuracy
culture	24%	$0.82 \pm .06$
politics	22%	$0.70 \pm .04$
miscellaneous	17%	$0.60 \pm .03$
conflicts abroad	11%	$0.68 \pm .06$
economy	9%	$0.65 \pm .09$
crime	8%	$0.73 \pm .06$
sport	5%	$0.90 \pm .06$
environment	4%	$0.59 \pm .11$

Table 1: Distribution of semantic text classification categories in the 500-word chunk version of the corpus and the validation accuracy.

3 Results and Discussion

All steps in the following analysis focus on the top 10% tokens with positive SHAP values in correctly classified texts; in this way we inspect only words that positively contribute to the correct classification decision. In addition to the usual quantitative analysis of SHAP values, we run a statistical analysis to identify which features significantly affect model performance. We make this decision to ensure that all reported effects are significant above chance, which is especially important for less frequent features. As null hypothesis, we assume that the distribution of each SHAP feature within a category matches its original distribution in the corpus for the same category. Positive contributions are reported when values significantly exceed the null hypothesis, and negative contributions when they are significantly lower (refer to Appendix A for more details). To reduce data sparsity and increase generalizability, we group all linguistic features into coarse-grained categories (e.g., "verbs" would include all types of verbs; refer to Appendix B for the detailed mapping).

3.1 POS Analysis

The outcome of the POS analysis in Figure 1 depicts the null hypothesis as the red central line,

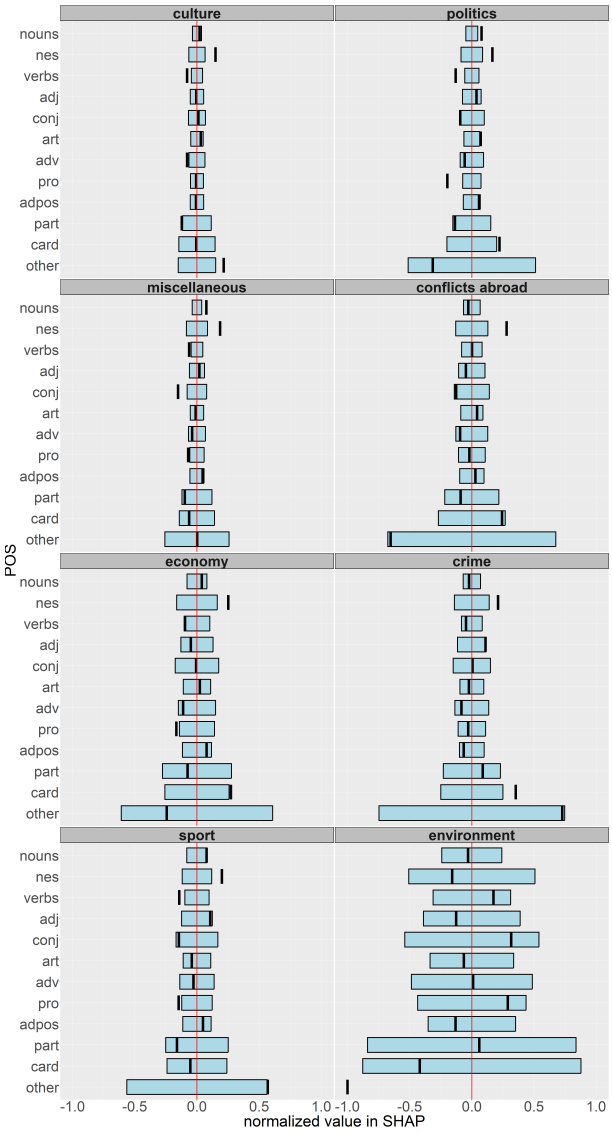

Figure 1: Distribution of POS per classification category, normalized against its category distribution. POS label groups are explained in Table 3 in Appendix B.

with a two standard deviations confidence interval.[2] The black vertical line represents the observed frequency of each POS among SHAP values. Values to the right of the red line indicate that a specific feature has a positive contribution in the SHAP values compared to its corpus distribution, while those to the left that it is has a negative contribution. Significance is reached when the black line is outside the confidence interval.

Among the noun POS-tags analysed, named entities (nes) have a significant positive impact on the predictions across all categories except for "environment". This finding perfectly mirrors the results by Kalouli et al. (2022) on English. Additionally, cardinal numbers (card) strongly predict categories related to factual content, such as "politics", "conflicts abroad", "economy", and "crime". Likely due to data sparsity (see Section 2.1), "environment" is the only category where no features reach significance.

3.2 Inflectional Morphology

We analyse morphological features for nouns, adjectives, and verbs. Figure 2 shows that for the classes "politics" and "conflicts abroad", nouns in nominative are significantly more present than in the overall distribution, while nouns in accusative are significantly less present. For "miscellaneous", only accusative reaches significance as a negative predictor. These differences are surprising given the many syncretisms between nominative and accusative in German, leading to identical surface forms. Possibly, the distinction mainly comes from their roles as subjects and objects.

For the number feature, plural is a negative predictor in "economy" and a positive one in "environment". This result is not straightforward to interpret and may hint to category-specific preferences.

For adjectives only the underspecification of case reaches positive significance in "politics", "miscellaneous", "culture", "crime", and "sport". In addition, accusative is a significant negative predictor in "politics" and genitive in "miscellaneous". For number, there are no significant predictors.

For verbs, the subjunctive is a significant negative predictor in "politics", "miscellaneous", "economy", and "crime". Given that the German subjunctive differs strongly in its morphology from

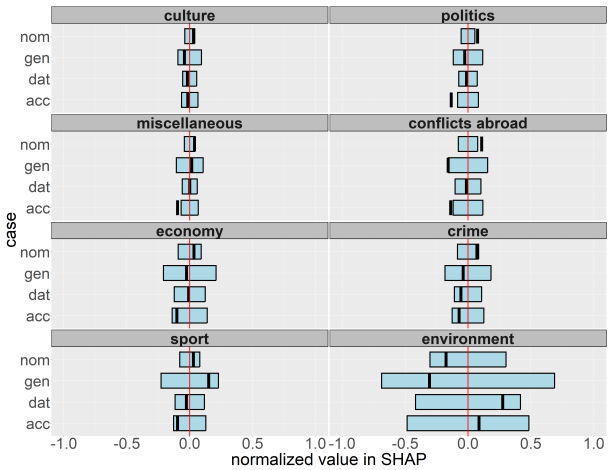

Figure 2: Normalized distribution of **case** (nominative, genitive, accusative, dative) for nouns.

[2]The main text includes only the most relevant figures for each step of the analysis, while the full set of plots supporting the discussion is provided in Appendix C.

the more commonly used indicative forms, BERT likely considers these less frequent forms as less important for the classification decision. In addition, it needs to be mentioned that the subjunctive in German has two morphological forms. Among the two, the *subjunctive 1* is used more commonly in German newspaper texts (as present in the TüBa corpus) compared to more casual forms of written German as it can be found in social media posts or the rest of the internet.

For the inflectional degree, tense, and person, only one significant negative predictor is found for infinitive and past forms in "conflicts abroad" and 1st person in "culture". For number only plural is a significant positive predictor for "environment". These observations hint to a class-specific phenomenon rather than a generalisable behaviour of the assessed model.

3.3 Syntactic Analysis

We study phrase, clause, complement, and dependency relations between the words in a sentence based on the annotation layers in the TüBa-corpus.

In the analysis of phrases, noun phrases and determiner phrases are significant positive predictors in "miscellaneous", while prepositional phrases are in "sport". Significant negative predictors are finite verb phrases in "miscellaneous", finite verb phrases in "politics", "miscellaneous", "conflicts abroad", "crime", and "sport", and determiner phrases in "politics". The analysis on the distribution of nouns across different phrase types reveals no significant results. Since our analysis considers the full 12 layer models, the results of the phrasal analysis do not contradict Jawahar et al. (2019), who claims that phrasal information tends to be more diluted in the lower layers of BERT.

With regard to the higher-order phrase levels, relative clauses (R-SIMPX) are significant negative predictors for most categories (except "economy"). Other subordinate types are not very important for the classification task.

The analysis of complements shows subjects as significant positive predictors in "politics" and "miscellaneous", while objects are significant negative predictors in "culture". This perfectly aligns with our previous discussion on the nominative case. We do not observe a clear preference for any other complement tags.

Dependency relations[3] provide a perspective on the relations between the words in a sentence. Figure 3 shows the result for dependencies.

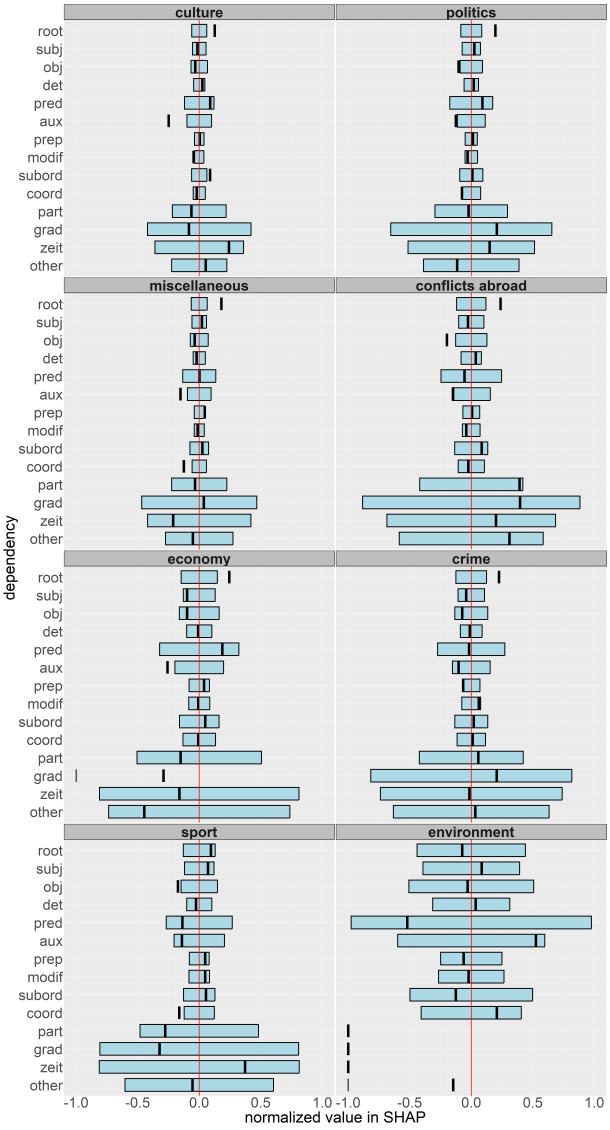

Figure 3: Normalized distribution of dependency labels for a dependency-grammar perspective on syntax. The labels are grouped according to Table 5 in Appendix B.

The multi-head attention mechanism in BERT allows the model to establish direct inter-token relationships similar to dependency relations. Similarly to the analysis of complements, the dependency analysis indicates that objects are strong negative predictors for "conflicts abroad", "politics", and "sport". However, this is not true for subjects. This variability has two possible explanations. First, the analysis relies on semi-automatically generated labels, increasing the probability of annotation errors. Second, complements and dependency relations differ in their theoretical definitions, leading

[3]Based on semi-automatically generated Hamburg Dependency Treebank (Foth et al., 2014) annotations from the 2010 version of the TüBa corpus.

to slight differences in the resulting annotation.

The initial word in a dependency structure ("root" tag) is a significant positive predictor for "culture", "politics", "miscellaneous", "conflicts abroad", "economy", and "crime". The additional results for significant positive and negative predictors are rather wide-spread over categories and relations, indicating possibly class-specific preferences.

Overall, the results from the dependency analysis are able to reproduce objects as a significant negative predictor, found in clausal representations and suggested in the morphological analysis of case for nouns. Further, the model seems to prefer the initial word in a dependency structure.

In a last step, we test whether the model is more likely to consider a pair of tokens because they are in a dependency relation. Since we cannot measure a reference value from the full corpus data because the corpus contains already the full dependency structure, we estimate the expected value for this observation to be $\approx 1,392$ tokens (calculated based on Equation 2 in Appendix A). This step considers for the set of top 10% of positive SHAP values the binary decision of a token and its governing token, which is approximately Poisson distributed. The observed value of 5,360 is nearly four times higher than expected, indicating the model gives disproportionate importance to words connected by a dependency relation.

A closer look at the types of dependencies linking these tokens reveals some general (yet non significant; probably due to data sparsity) tendencies for tokens connected by a subject, subordinate, particle or modifier dependency relation.

Overall, these results suggest that the model does not favour a specific type of dependency. Instead, it appears to group different tokens based on their connections through specific dependency relations, such as subject, particle, modifier, or subordinate relations. This could indicate that the model considers words within smaller syntactic clusters, linking them according to their dependency relations.

3.4 Semantic Analysis

To assess the influence of semantic features on the classification task, we use GermaNet.[4] The semantic ontology includes information on verbs, adjectives and nouns. Since the results for POS on both verbs and adjectives do not yield significant results, the following analysis focuses uniquely on the same nouns as considered in the previous sections. The semantic analysis exploits the tree-like structure of the ontology. We pair each noun[5] in the category with every other noun in the same category and identify the closest shared hypernym for each pair. We then count the number of shared hypernyms in each category and normalize this by the frequency of each hypernym in the corpus. This allows us to identify hypernyms that are generally uncommon across categories, but very distinctive to a specific one. Finally, we rank these hypernyms by category, analyse the top 20, and manually select only those that are associated to the category. Table 2 reports the count of selected hypernyms appearing in the top 5, 10, 15, and 20 most frequent hypernyms and the percentage in the top 20 for each category.

class	top 5	top 10	top 15	top 20	percent
culture	1	2	4	5	10%
politics	0	0	0	1	4%
miscellaneous	0	0	0	0	0%
conflicts abroad	0	0	0	0	0%
economy	0	0	0	0	0%
crime	0	1	1	1	4%
sport	0	2	4	7	33%
environment	0	0	0	1	100%

Table 2: Analysis of class related hypernyms in top 5-20 hypernyms with the highest weighted mean. The last column reports the percentage of all class-related hypernyms present in the top 20.

In this ranking of counts of shared hypernyms (weighted by general hypernym frequency), we can assess which class has an exceptionally high number of hypernyms related to its topic. For, "culture" and "sport", a higher percentage of class-related concepts in the most frequent shared hypernyms correlates with a higher classification accuracy (cf. Table 1). The high number of class-related shared hypernyms indicates that the words that are important to the classification decision are more likely to be hyponyms of rather class-specific concepts. This outcome indicates a closer semantic cluster making it easier to for the model to discriminate the category.

Overall, this suggests that not only the class-size is decisive for the classification accuracy, but also that a smaller category may benefit from a higher semantic proximity of its characteristic words.

[4] Accessed using the provided Germanetpy API (https://github.com/Germanet-sfs/germanetpy).

[5] Pairs of identical nouns and named entities were excluded.

4 Conclusion

This work investigates the role of linguistic information in a monolingual German BERT model for a multi-class classification task. It replicates prior findings on the dilution of phrasal information in a full 12-layer model (Jawahar et al., 2019) and BERT's preference for NEs (Kalouli et al., 2022).

The results suggest that German BERT's syntactic representation prioritizes dependency relations over clausal or phrasal ones by focusing on word clusters in dependency relations, showing opportunities for further research. Additionally, German noun inflection has a minor influence, with a preference for nominative over accusative, possibly due to the syntactic function outweighing its morphological form.

The semantic analysis shows that classification accuracy depends not only on class size but also on smaller categories forming a coherent semantic space, and consequently, increasing their distinctiveness.

Overall, this study indicates some cross-linguistic consistency in BERT's linguistic representations while emphasizing the need for further analyses of language-specific phenomena, especially in low-resource contexts.

5 Limitations

When interpreting the results of this study, it is important to note that only one model (BERT), one corpus domain (news), and one specific semantic classification task was analysed. Therefore, the findings may reflect the specific distributions of the assessed corpus and task; yet high generalisability is expected given the broad nature of the chosen task. Some results are not straightforward to interpret, but we offer explanations based on the most reasonable assumptions.

Finally, the study does not specifically analyse the full complexity of inflectional morphology and syntax. A more detailed analysis of nouns could provide further insights into the model's preferences. Similarly, more research is needed to understand how structural simplifications impact syntactic complexity and the contribution of specific words to this process.

6 Ethics statement

We do not anticipate any ethical concerns with this work. We used open-sourced data and models, which have been appropriately cited.

7 Acknowledgements

We would like to thank the Fraunhofer Institute for Communication, Information Processing and Ergonomics (FKIE) for supporting this research, particularly Daniel Claeser and Albert Pritzkau for their valuable guidance. Additionally, we acknowledge the Volkswagen Stiftung for partially funding this work under grant 98 543. We also appreciate the resources provided by the Department of Linguistics at the University of Konstanz, which facilitated some of our experiments. Finally, we thank Jonathan Pampel for insightful discussions on the mathematical aspects of this paper.

References

Keno K. Bressem, Jens-Michalis Papaioannou, Paul Grundmann, Florian Borchert, Lisa C. Adams, Leonhard Liu, Felix Busch, Lina Xu, Jan P. Loyen, Stefan M. Niehues, Moritz Augustin, Lennart Grosser, Marcus R. Makowski, Hugo J.W.L. Aerts, and Alexander Löser. 2024. medbert.de: A comprehensive german bert model for the medical domain. *Expert Systems with Applications*, 237:121598.

Patrick Brookshire and Nils Reiter. 2024. Modeling moravian memoirs: Ternary sentiment analysis in a low resource setting. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 91–100, St. Julians, Malta. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Shubanker Banerjee, Manoj Balaji Jagadeeshan, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Sean Benhur, and John Philip McCrae. 2023. Detecting abusive comments at a fine-grained level in a low-resource language. *Natural Language Processing Journal*, 3:100006.

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, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Daniel Benedikt Claeser. 2022. *Zur Rolle der Flexionsmorphologie in der automatischen Klassifikation deutschsprachiger Textdokumente*. Ph.D. thesis, Rheinische Friedrich-Wilhelms-Universität Bonn.

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.

Peter Eisenberg. 2020. *Grundriss der deutschen Grammatik: Das Wort*. J.B. Metzler.

Kilian A. Foth, Arne Köhn, Niels Beuck, and Wolfgang Menzel. 2014. Because Size Does Matter: The Hamburg Dependency Treebank. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2326–2333, Reykjavik, Iceland. European Language Resources Association (ELRA).

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *CoRR*, abs/1901.05287.

Raffaele Guarasci, Stefano Silvestri, Giuseppe De Pietro, Hamido Fujita, and Massimo Esposito. 2021. Assessing bert's ability to learn italian syntax: a study on null-subject and agreement phenomena. *Journal of Ambient Intelligence and Humanized Computing*, 14(1):289–303.

Hyewon Jang, Qi Yu, and Diego Frassinelli. 2023. Figurative Language Processing: A Linguistically Informed Feature Analysis of the Behavior of Language Models and Humans. *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9816–9832.

Ganesh Jawahar, Benoit Sagot, and Djamé Seddah. 2019. What Does BERT Learn about the Structure of Language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint*.

Aikaterini-Lida Kalouli, Rita Sevastjanova, Christin Beck, and Maribel Romero. 2022. Negation, Coordination, and Quantifiers in Contextualized Language Models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3074–3085, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Enja Kokalj, Blaž Škrlj, Nada Lavrač, Senja Pollak, and Marko Robnik-Šikonja. 2021. BERT meets Shapely: Extending SHAP Explanations to Transformer based Classifiers. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 16–21. Association for Computational Linguistics.

Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

OpenAI, Josh Achiam, et al. 2023. Gpt-4 technical report. *arXiv preprint*.

Maryan Rizinski, Hristijan Peshov, Kostadin Mishev, Milos Jovanovik, and Dimitar Trajanov. 2024. Sentiment analysis in finance: From transformers back to explainable lexicons (xlex). *IEEE Access*, 12:7170–7198.

Rajalakshmi Sivanaiah, Subhankar Suresh, Sushmithaa Pandian, and Angel Deborah Suseelan. 2024. Bridging the language gap: Transformer-based bert for fake news detection in low-resource settings. In *Speech and Language Technologies for Low-Resource Languages*, pages 398–411, Cham. Springer Nature Switzerland.

Xiaoyi Tang, Hongwei Chen, Daoyu Lin, and Kexin Li. 2024. Incorporating fine-grained linguistic features and explainable ai into multi-dimensional automated writing assessment. *Applied Sciences*, 14(10).

Heike Telljohann, Erhard Hinrichs, and Sandra Kübler. 2004. The Tüba-D/Z Treebank: Annotating German with a Context-Free Backbone. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Heike Telljohann, Erhard W. Hinrichs, Sandra Kübler, Heike Zinsmeister, and Kathrin Beck. Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv preprint*.

Karolina Zaczynska, Nils Feldhus, Robert Schwarzenberg, Aleksandra Gabryszak, and Sebastian Möller. 2020. Evaluating german transformer language models with syntactic agreement tests.

A Statistical Analysis

The statistical analysis in this paper highlights the significance of specific linguistic features in the text classification task. We compare the distribution of each feature in the SHAP values to its distribution in the same category within the TüBa corpus. Both distributions are normalized based on the total number of words per category in the SHAP values and in the entire corpus, respectively. As shown in Equation 1, the null hypothesis ($H0$) assumes that the two distributions are identical:

$$H0 : \frac{k}{n} = \frac{K}{N} \tag{1}$$

where k is the count of the feature in the category in the SHAP values, n is the total number of words in the top 10% positive SHAP values for the category, K is the count of the feature in the category within the corpus, and N is the total number of words in the category in the corpus.

The statistical analysis (two-sided t-test) identifies features in SHAP that have a significantly higher or lower contribution to model performance compared to their actual distribution in the corpus.

The analysis for the co-occurrence of tokens with their governing token in a dependency relation requires to estimate an expected value as reference under the assumption that the dependency-related tokens end up in the SHAP values based on a random selection. A random selection assumes in this case that there is a binary criterion of a token and its governing token being in the SHAP values or not. Under this assumption, Equation 2 approximates the question whether two dependency-related tokens end up in the SHAP values as a Poisson Distribution.

$$\begin{aligned} A &:= P_{\text{gov. token of last word in SHAPS}} \\ &= \frac{n_{SHAPs}}{n_{Tueba}} \\ &= \frac{46043}{1523384} \\ E_{\text{gov. tokens in SHAPs}} &= n_{SHAPs} \cdot A \\ &= \frac{n_{SHAPs}^2}{n_{Tueba}} \\ &= \frac{46043^2}{1523384} \\ &\approx \underline{\underline{1392}} \end{aligned} \tag{2}$$

B Labels Mapping

Here, we document the mapping between the fine-grained labels in the corpus, based on the Stuttgart-Tübingen-Tagset (STTS), complement labels and the dependency labels according to the Hamburg Dependency Treebank. Table 3 documents POS tags, Table 4 complement labels, and Table 5 the dependency labels.

Grouped Tag	Abbreviation	STTS Tag
nouns	nouns	NN
named entities (NEs)	nes	NE
adjectives	adj	ADJA, ADJD
cardinal numbers	card	CARD
verbs	verbs	VMFIN, VAFIN, VVFIN, VAIMP, VVIMP, VVINF, VAINF, VMINF, VVIZU, VVPP, VMPP, VAPP
articles	art	ART
pronouns	pro	PPER, PRF, PPOSAT, PPOSS, PDAT, PDS, PIAT, PIDAT, PIS, PRELAT, PRELS, PWAT, PWS, PWAV, PAV
adverbs	adv	ADV
conjunctions	conj	KOUI, KOUS, KON, KOKOM
particle	part	PTKZU, PTKNEG, PTKVZ, PTKA, PTKANT
other	other	ITJ, TRUNC, XY, FM

Table 3: Mapping between fine-grained STTS labels and the coarse-grained labels used in the **POS** analysis.

Grouped Tag	Abbreviation	Complement Tag
subject	subj	ON
object	obj	OD, OA, OG, OS, OPP, OADVP, OADJP
predicate	pred	PRED
verbal objects	ov	OV
facultative prepositional object	fopp	FOPP
apposition	app	APP

Table 4: Grouping of **complement labels** for the analysis. For the original labels, see (Telljohann et al.).

Grouped Tag	Abbreviation	Dependency Tag
root node of dependency structure	root	ROOT
subject	subj	SUBJ, SUBJC, EXPL
object	obj	OBJA, OBJI, OBJG, OBJC, OBJD
subordination	subord	APP, NEB, REL, PAR, S, gmod-app
determiner	det	DET
predicative complement	pred	PRED
auxiliary	aux	AUX
prepositions	prep	PP, PN, OBJP
modifier	modif	ADV, ATTR, GMOD, PART, KOM
subordordination	subord	REL, NEB, PAR
coordination	coord	CJ, KONJ, KON, koord
participles	part	AVZ, PART
time information	zeit	ZEIT
gradual (indicating a measure)	grad	GRAD
other	other	left over punctuation signs, -UNKNOWN-

Table 5: Grouping of **dependency labels** for the analysis based on the labels from the Hamburg Dependency Treebank (Foth et al., 2014).

C Feature Analysis: Additional Plots

Below, we present the additional plots supporting the complete feature analysis as documented in the main text for the morphological features of nouns (Figure 4), adjectives (Figures 5 and 6), and verbs (Figures 7, 9, 8, 10, and 11), followed by the syntactic analysis for phrasal (Figures 12, 13, and 14), clausal (Figure 15), and dependency (Figure 16) features. As discussed in Section 3.1, the red central line indicates the null hypothesis surrounded by the $2\sigma(95\%)$ confidence interval in light blue. The black vertical line represents the observed frequency of each feature among SHAP values. Values to the right of the red line indicate that a specific feature is over-represented in the SHAP values compared to its corpus distribution, while those to the left that it is under-represented. Significance is reached when the black line is outside the confidence interval.

C.1 Morphology Plots

C.2 Morphology: Nouns

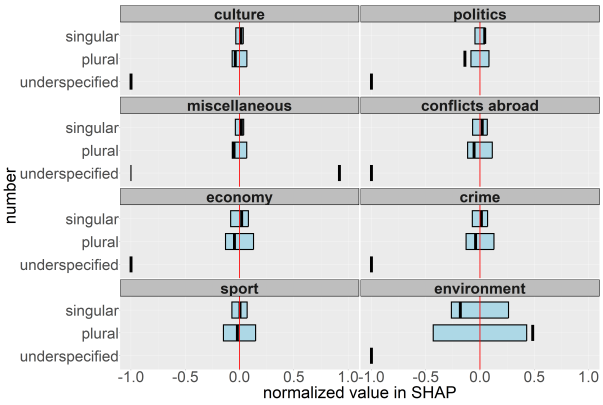

Figure 4: Normalized distribution of **number** (singular, plural, underspecified) for nouns.

C.2.1 Morphology: Adjectives

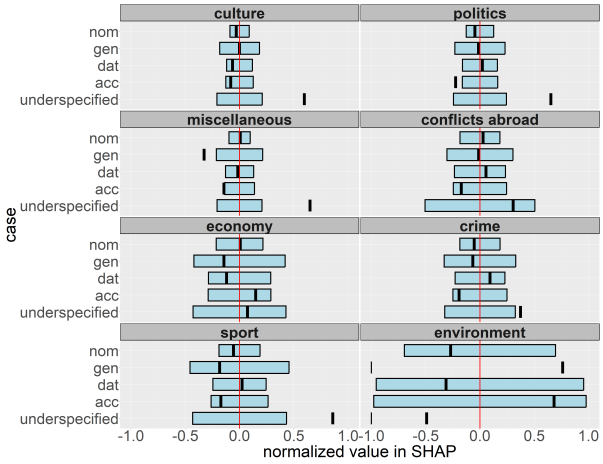

Figure 5: Normalized distribution of **case** (nominative, genitive, dative, accusative, underspecified) for adjectives.

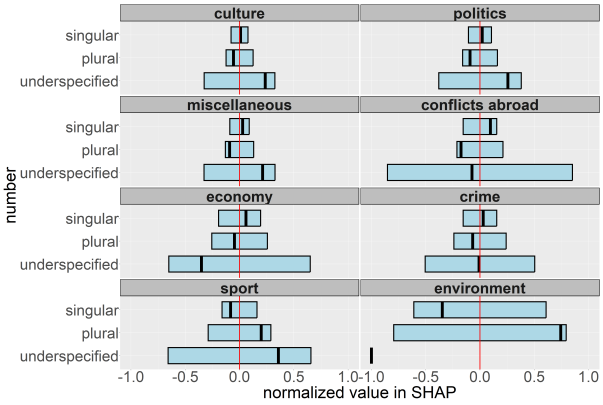

Figure 6: Normalized distribution of **number** (singular, plural, underspecified) for adjectives.

C.2.2 Morphology: Verbs

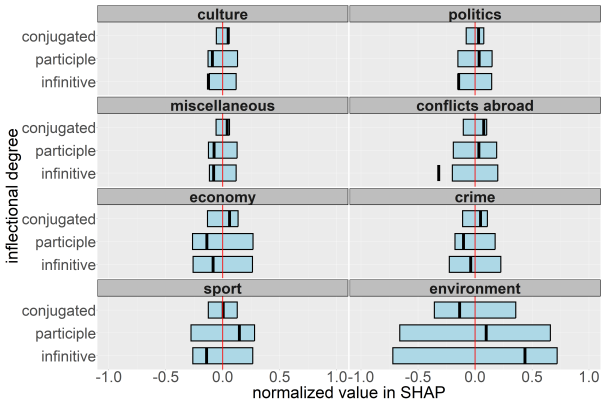

Figure 7: Normalized distribution of **inflectional degree** (infinitive, participle, inflected) for verbs.

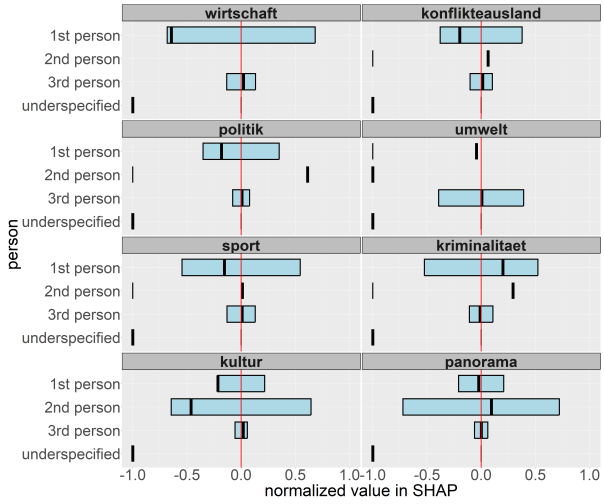

Figure 8: Normalized distribution of **grammatical person** (1st, 2nd, 3rd, underspecified) for verbs.

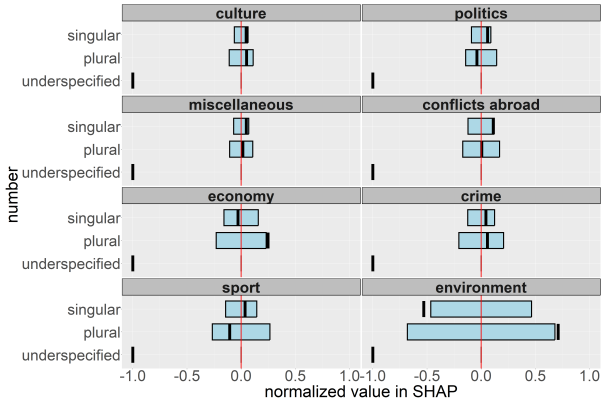

Figure 9: Normalized distribution of **number** (singular, plural, underspecified) for verbs.

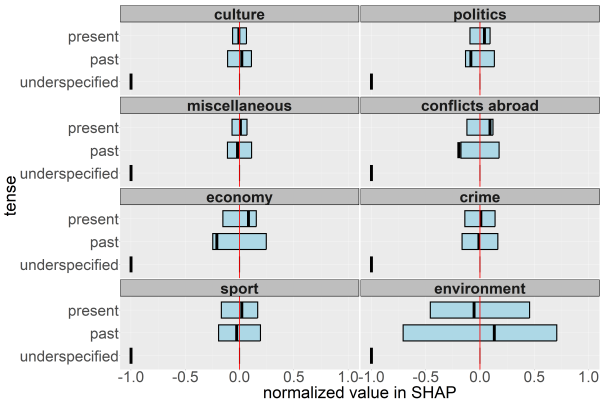

Figure 10: Normalized distribution of **tense** (present, past, underspecified) for verbs.

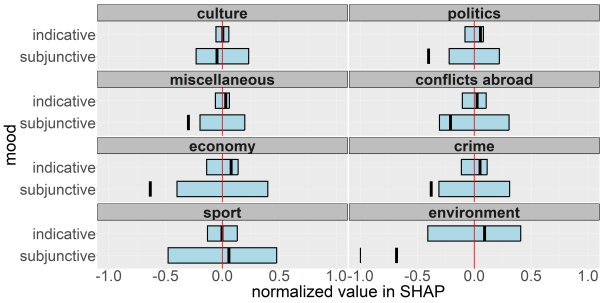

Figure 11: Normalized distribution of **mood** (indicative and subjunctive (*Konjunktiv*)) for verbs.

C.3 Syntactic analysis

C.3.1 Phrasal analysis

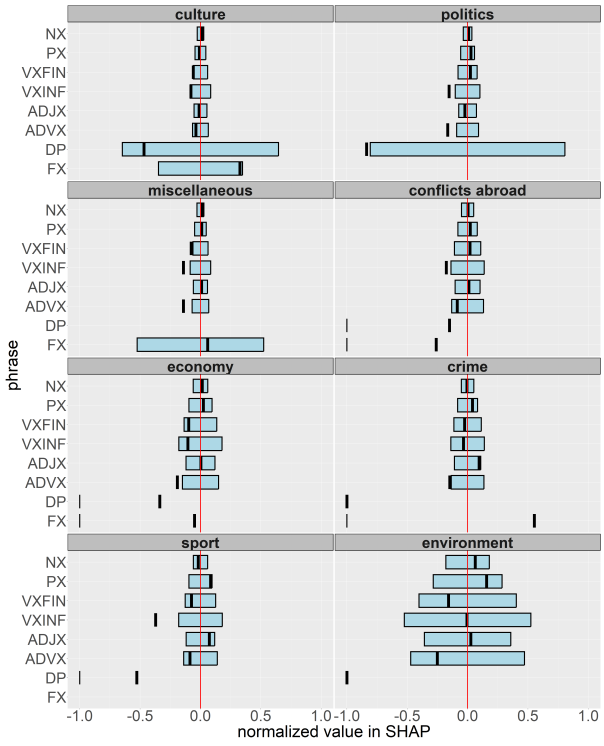

Figure 12: Normalized distribution of phrase labels.

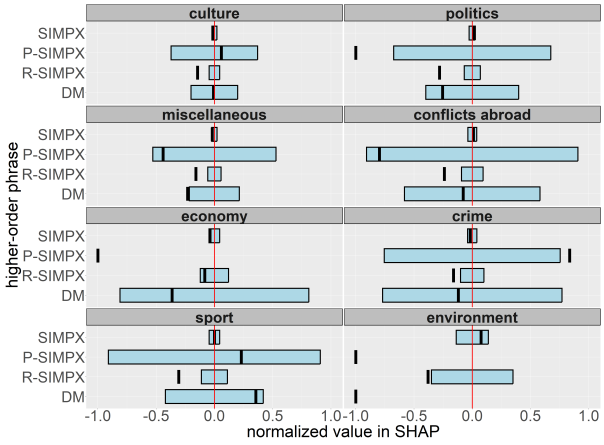

Figure 13: Normalized distribution of higher-order phrase labels.

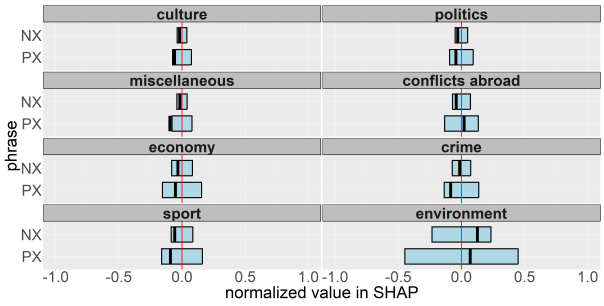

Figure 14: Normalized distribution of phrase labels for nouns to assess whether the phrasal attachment of a noun influences the classification.

C.3.2 Clausal Analysis

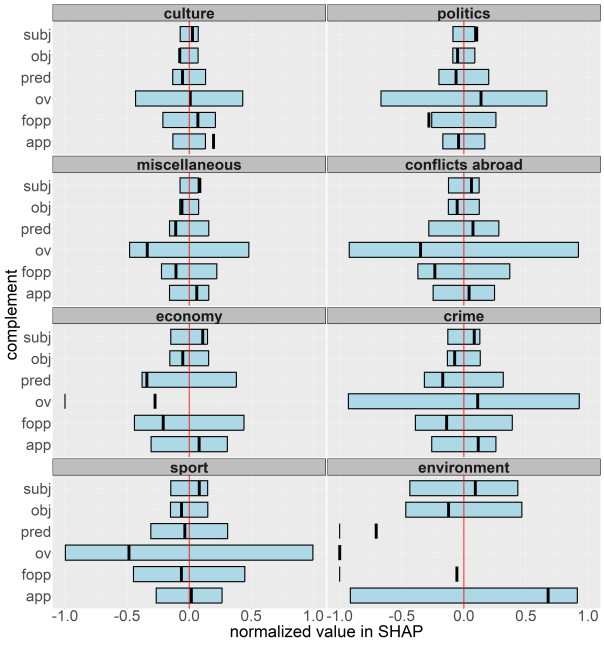

Figure 15: Normalized distribution of complement labels. The grouping of the labels can be found in 4.

C.3.3 Dependency Analysis

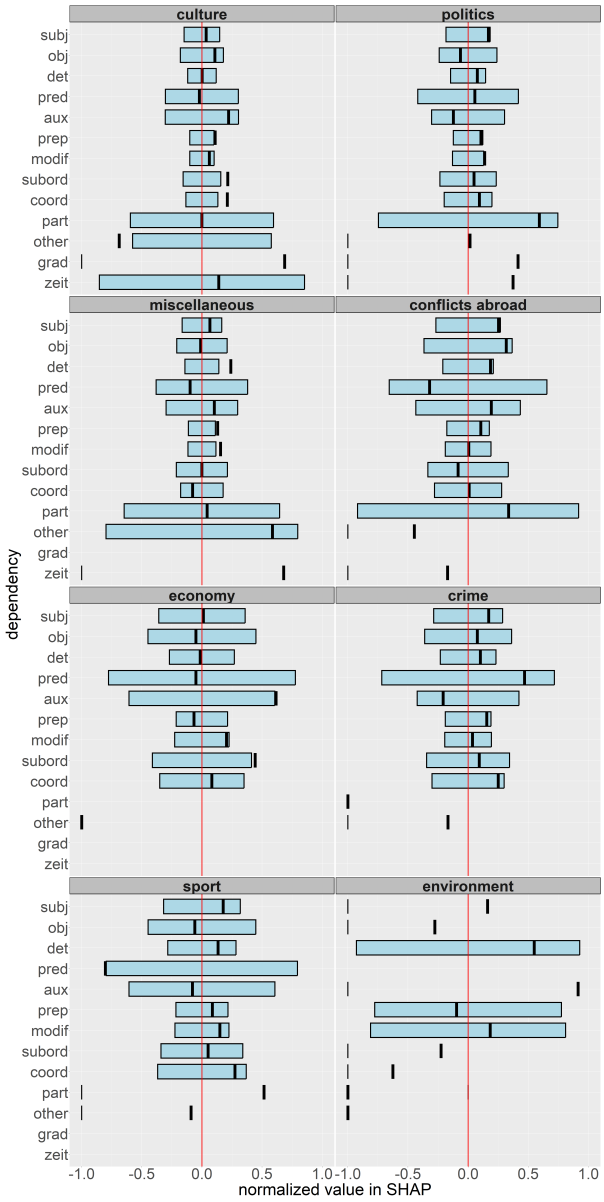

Figure 16: Normalized distribution of dependency relations between tokens, where both tokens appear in the SHAP values. This aims to reveal whether specific tokens are important to the classification task due to their dependency relations.