Extracting Lexical Features from Dialects via Interpretable Dialect Classifiers

Roy Xie[♣♦] **Orevaoghene Ahia**[♠] **Yulia Tsvetkov**[♠] **Antonios Anastasopoulos**[♦]
[♣] Duke University
[♠] Paul G. Allen School of Computer Science & Engineering, University of Washington
[♦] Department of Computer Science, George Mason University
ruoyu.xie@duke.edu {oahia, yuliats}@cs.washington.edu antonis@gmu.edu

Abstract

Identifying linguistic differences between dialects of a language often requires expert knowledge and meticulous human analysis. This is largely due to the complexity and nuance involved in studying various dialects. We present a novel approach to extract distinguishing lexical features of dialects by utilizing interpretable dialect classifiers, even in the absence of human experts. We explore both post-hoc and intrinsic approaches to interpretability, conduct experiments on Mandarin, Italian, and Low Saxon, and experimentally demonstrate that our method successfully identifies key language-specific lexical features that contribute to dialectal variations.[1]

1 Introduction

Dialects and closely related languages exhibit subtle but significant variations, reflecting regional, social, and cultural differences (Chambers and Trudgill, 1998). Identifying and distinguishing differences between these dialects is of great importance in linguistics, language preservation, and natural language processing (NLP) research (Salameh et al., 2018; Goswami et al., 2020). Traditionally, identifying the specific linguistic features that distinguish dialects of a language has relied on manual analysis and expert knowledge (Cotterell and Callison-Burch, 2014), as the differences between these dialects could be subtle and hard to detect without linguistic expertise (Zaidan and Callison-Burch, 2011). This process is also time-consuming and is usually language-specific due to the peculiarities that different languages exhibit.

Extracting distinctive words of particular dialects is essential for studying dialectal variation, especially in dialectology (Chambers and Trudgill, 1980). In this work, we focus on extracting word-level distinguishing and salient features in dialects,

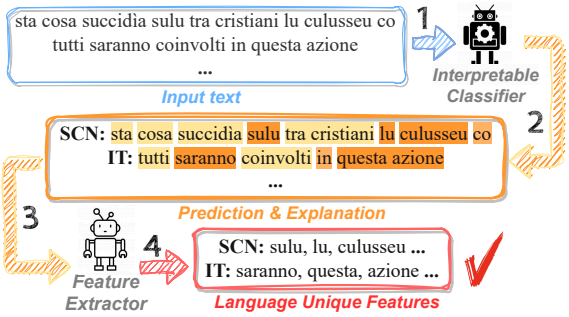

Figure 1: *(1)* given an input text; *(2)* the interpretable dialect classifier return labels (SCN and IT) and explanations; *(3)* the extractor takes the explanations and *(4)* outputs meaningful features to the languages.

also called 'shibboleths' (Prokić et al., 2012). We propose an automated workflow that could potentially assist researchers such as dialectologists and corpus linguists in their analysis of dialectal variations. To achieve this, we leverage strong neural classifiers of dialects (Aepli et al., 2023; Scherrer et al., 2023, 2022) and pair them with model interpretability techniques to extract these features.

We hypothesize that there are certain distinguishing features in dialects that the models learn during training, which enables them to make accurate predictions at test time. We utilize post-hoc (Simonyan et al., 2014; Ribeiro et al., 2016) and intrinsic (Alvarez-Melis and Jaakkola, 2018; Arik and Pfister, 2020) feature attribution explanation methods to extract these features from model interpretations, in the form of local explanations from dialect classifiers.

Our experiments focus on three language groups: Mandarin, Italian, Low Saxon, and their respective dialects. We demonstrate the effectiveness of our approach through automatic evaluation, human evaluation, and extensive analysis. We use known lexical features of some of the dialects we consider and show the viability of using explanation methods to extract unique dialectal features.

[1]Data and code are available: https://github.com/ruoyuxie/interpretable_dialect_classifier

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 54–69
June 16-21, 2024

2 Background and Related Work

The importance of interpreting model decisions has increasingly been recognized in recently (Belinkov et al., 2020; Koh and Liang, 2017; Mosca et al., 2022; Rajagopal et al., 2021), primarily due to its role in deciphering the inner workings of black-box models. Two main approaches include: *(i)* post-hoc methods (Lundberg and Lee, 2017; Ribeiro et al., 2016; Jin et al., 2019) which provide insights into the predictions of pre-existing models based on model internals; *(ii)* intrinsic methods (Rajagopal et al., 2021; Alvarez-Melis and Jaakkola, 2018; Rigotti et al., 2021) where interpretability is an integral feature that is optimized concurrently with the model's main primary task during training.

This paper also closely relates to dialect classification (Aepli et al., 2023; Scherrer et al., 2023, 2022; Jauhiainen et al., 2019). Some high-performing neural dialect classifiers are able to achieve $> 90\%$ accuracy even when there are multiple categories and levels of noise present or there are just very subtle differences in general (Srivastava and Chiang, 2023). Most languages worldwide are under-resourced (Joshi et al., 2020), and basic tasks like identifying verbs and nouns within a dialect can be challenging, given that syntactic parsing is primarily efficient for well-resourced languages (Hellwig et al., 2023; Hou et al., 2022). In this work, we focus on extracting lexical differences between dialects, as they show potential in distinguishing dialectal variations. The idea of automatically extracting linguistic features is not new (Brill, 1991; Demszky et al., 2021). However, we identify these features *through the lens of model explanations* by using feature-attribution methods. To the best of our knowledge, this is the first study to undertake such an endeavor.

3 Method

We present a simple yet effective method to extract lexical differences that distinguish dialects using explanations obtained from interpretable dialect classifiers as shown in Figure 1.

3.1 Interpretable Dialect Classifier

Our interpretable dialect classifier, built on top of transformer-based models, is designed to work with both post-hoc and intrinsic interpretable methods. For the scope of this work we focus on Leave-One-Out (LOO), a popular model-agnostic feature-attribution method, and SelfExplain (Rajagopal et al., 2021), an intrinsic interpretable method that learns to attribute text classification decisions to relevant parts in the input text. For the intrinsic approach, we incorporate the model encoder into the SelfExplain framework, exclusively extracting only local explanations. For the post-hoc approach, we train a separate classifier to calculate the LOO interpretations.

3.2 Explanation Methods

We start with a dialect classifier trained to take an input sentence X and predict its corresponding label $\mathbf{y}$. Let $\mathbf{u}_s$ be the final layer representation of the "[CLS]" token for $\mathbf{X}$, which is the sentence representation typically used to make a prediction.

Post-hoc Approach During inference, LOO estimates the attribution score of each token x_i in input $\mathbf{X}$ in relation to model's prediction $\hat{\mathbf{y}}$. To do so, $\mathbf{u}_s$ is passed through ReLU, affine, and softmax layers to yield a probability distribution over outputs. For each feature x_i, LOO calculates the change in probability when $\{x_i\}$ is removed from $\mathbf{X}$. Let $X \backslash \{x_i\}$ denote input $\mathbf{X}$ without feature x_i and u_i the final layer representation of the "[CLS]" token for $X \backslash \{x_i\}$. We term this the relevance score and expect that influential features/explanations in the input $\mathbf{X}$ will have higher scores.

$$\begin{aligned} \boldsymbol{\ell} &= \text{softmax}(\text{affine}(\text{ReLU}(\mathbf{u}_s))) \\ \boldsymbol{\ell_i} &= \text{softmax}(\text{affine}(\text{ReLU}(\mathbf{u_i}))) \\ \boldsymbol{\nabla_i} &= \boldsymbol{\ell} - \boldsymbol{\ell_i} \end{aligned} \tag{1}$$

Intrinsic Approach For our intrinsic approach using SelfExplain (Rajagopal et al., 2021), we augment the dialect classifier with a Local Interpretability Layer (LIL) during training. This layer quantifies the relevance of each feature x_i in input $\mathbf{X}$ to the final label distribution $\boldsymbol{\ell}$ via activation difference (Shrikumar et al., 2017), and is trained jointly with the final classifier layer. Taking $\boldsymbol{\ell}$ in Equation 1, the loss is the negative log probability, summed over all training instances:

$$L_{dialect-classifier} = -\sum_i \log \boldsymbol{\ell}[y_i^*]$$

where y_i^* is the correct label for instance i. To obtain the attribution score of each feature x_i in input $\mathbf{X}$, we first estimate the output label distribution without x_i by transforming the difference between $\mathbf{u}_s$ and $\mathbf{u}_j$, where $\mathbf{u}_j$ is the MLM representation of feature x_i:

$$\mathbf{s}_j = \text{softmax}(\text{affine}(\text{ReLU}(\mathbf{u}_s) - \text{ReLU}(\mathbf{u}_j)))$$

$$\text{loss} = L_{dialect-classifier} + \alpha_1 L_{LIL}$$

The relevance of each feature x_i can be defined as the change in probability of the correct label when x_i is included vs. excluded:

$$r_j = [\boldsymbol{\ell}]_{y_i^*} - [\mathbf{s}_j]_{y_i^*}$$

where higher r_j signifies more relevant features to the prediction, serving as better explanations.

Mapping Explanations to Lexical Features We extract explanations from the classifiers outlined above. Note, however, that these explanations are at the *sentence* level, but one ideally would need features that in general identify/describe one dialect in contrast to another at the *language* level, i.e., at the *corpus* level. To achieve this, we devise a corpus-level feature extraction method that takes sentence-level explanations as input and produces "global" features.[2]

Given a set of sentence-level explanations $E = \{e_1, e_2, \ldots, e_n\}$ from a classifier, we first filter out explanations from incorrect predictions or those that are not unique to a specific language variety. Let E' represent the filtered set of explanations:

$$E' = \{e \in E \mid \text{isCorrect}(e) \wedge \text{isUnique}(e)\}$$

Next, we apply Term Frequency-Inverse Document Frequency (TF-IDF) to E' to extract the most salient global features. Let F be the final set of extracted features. The TF-IDF score for a term t in a document d in a corpus D is given by:

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D)$$

We can then define the feature extraction as:

$$F = \{\text{TF-IDF}(t, d, E') \mid t \in d, d \in E'\}$$

4 Languages and Datasets

4.1 Languages

We discuss below each of the language continua, which have multiple dialects that vary distinctly across different regions. We study three distinct language continua and their respective dialects: Mandarin, Italian, and Low Saxon (Dutch and German). Our selection is largely influenced by cultural and typological diversity concerns, but also by the dearth of dialectal data for other languages. More information about the dialects can be found in Appendix A.

[2]In this study, we extract distinguishing lexemes (words) through unigrams. Nonetheless, our approach can be readily adapted to phrase-level analysis using ngrams, though such an extension falls beyond the purview of this work.

4.2 Datasets

We use FRMT dataset (Riley et al., 2022) for Mandarin dialects. FRMT provides sentence- and word-level translations of English Wikipedia text into Mainland Mandarin (CN) and Taiwan Mandarin (TW). For Italian dialects, we combine data from the Identification of Languages and Dialects of Italy (ITDI) shared task (Aepli et al., 2022) and Europarl v8 corpus (Koehn, 2005). ITDI provides 11 Italian dialects obtained from crawling Wikipedia, and we use them to mix with the same amount of data from standard Italian in Europarl. For Low Saxon dialects, we use LSDC dataset (Siewert et al., 2020), which consists of 16 West-Germanic Low Saxon dialects from Germany and the Netherlands. Appendix B shows the statistics for all datasets. We remove all punctuations and lowercase all Latin characters to focus on extracting lexical features.

5 Experiments

In this section, we present multiple experiments and demonstrate that our method results in distinguishing and meaningful feature extraction. For the main results, we focus on Mandarin (CN-TW) and Sicilian-Italian (SCN-IT). A total of four annotators are involved in the evaluation process.[3]

5.1 Models

Our dialect classifiers are built on XLM-RoBERTa base (Conneau et al., 2020). We maintain the hyperparameters and weights from the pre-training of the encoders and train the models for 5 epochs with a batch size 16 for each language pair. All experiments are done on an NVIDIA A100 GPU.

5.2 TF-IDF Baseline

As a baseline, we use TF-IDF to evaluate how effectively it extracts meaningful lexical features from CN-TW language pair (Table 1). We present annotators with the extracted lexical features using TF-IDF, asking them to determine whether these features are salient and unique to the language.[4]

5.3 Explanation Evaluation

We hypothesize that good explanations from highly performing dialect classifiers should be subsets of

[3]All annotators are proficient in the annotated language pairs. Three annotators work on CN-TW, and one for SCN-IT, as it is difficult to find annotators who are proficient in multiple dialects.

[4]The annotators are given options to select if a feature is likely to belong to one, both, or neither of the dialects.

Option (%)	CN	TW
CN	**70.0**	0.0
TW	0.0	**60.0**
Both	30.0	40.0
None	0.0	0.0

Table 1: Baseline results for capturing language-unique features for CN-TW.

the input that represent distinctive features of the respective dialect in which the input is written. We first evaluate the general robustness of explanations using **Sufficiency** (Jacovi et al., 2018; Yu et al., 2019) (Do explanations adequately represent the model predictions?) and **Plausibility** (Ehsan et al., 2019; Hayati et al., 2023) (Do explanations seem credible and comprehensible to humans?)

Sufficiency We train a separate classifier to perform the dialect classification task with only the explanations as input, and the predicted labels as target. Higher accuracy indicates that the explanations are more reflective of the model predictions. We train these models with the top ranking explanations of each sentence as input, and present the results in Table 2 for both explanation methods. Both methods achieve over 90% accuracy when k $\geq$ 3, which sets a reliable baseline for further evaluation. CN-TW and SCN-IT have an average sentence length of 18.8 and 16.2, respectively, which implies that our sufficiency scores are trustworthy, as we obtain them with less than 20-30% of an average sentence.

Methods	Dialects	$k=1$	$k=3$	$k=5$
SelfExp	CN-TW	76.5	96.7	97.8
	SCN-IT	87.8	95.8	97.9
LOO	CN-TW	81.3	93.2	97.2
	SCN-IT	87.4	95.4	96.6

Table 2: More explanations lead to higher sufficiency. Both explanation methods are over 90% accurate when $k \geq 3$, setting a reliable baseline for future evaluation.

Plausibility We give each annotator 25 sentences with the model's predictions and the top three explanations from LOO and SelfExplain. We randomly shuffle and anonymize the explanation methods and ask annotators *(i)* Should the model classify a given sentence in certain dialects based on the explanations? *(ii)* If the explanations do not adequately justify the model's prediction, what should the model's prediction be based on? The annotators are given options to select one, both, or neither explanation method. We present the percentage of instances that are adequately justified according to the annotators on Figure 2. Overall, LOO achieves a higher percentage of perceived adequate justification, compared to SelfExplain. Therefore, we will use LOO as the main explanation method for the rest of the studies in this work.

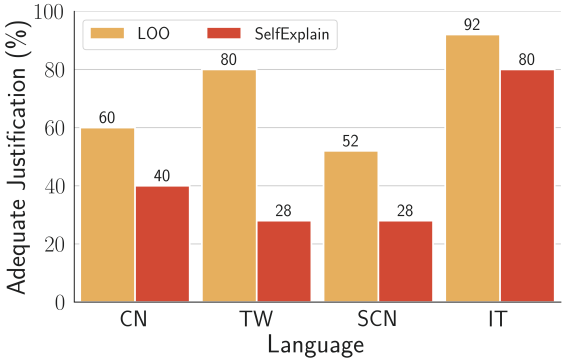

Figure 2: Adequate justification percentage for LOO and SelfExplain. Humans found LOO produces more justifiable explanations across all four dialects.

5.4 Can we extract lexical features by interpreting dialect classifiers?

Automatic Evaluation As a sanity check, we propose a simple automatic evaluation metric, PICK-UP RATE (PR), to evaluate the ability of explanation methods on capturing language-unique features. Based on comparative dialectology and other linguistic literature, we identify a small set of distinctive features for the languages we work with.[5] We define PR as the likelihood of language-unique features being captured in the input sample. The higher the PR is, the better the explanation method aligns with some ground truth features. Given e_g as the number of times that a language-unique feature g is used as an explanation e and c_g as the number of times that g appears in a corpus c, $PR(g) = \frac{\text{count}(e_g)}{\text{count}(c_g)}$.

FRMT provides a set of Mandarin word translations between CN and TW. We present the PR result for these translated words in Appendix C.1. For both explanation methods, the language-unique translations tend to only appear in their corresponding classes with reasonable pick-up rates. Note that script differences between CN and TW make its fairly easy for the classifiers to correctly classify them. Therefore, we conduct the same experiments

[5]After reviewing the literature, we identified features that can be easily measured, which allows us to potentially show that the features that humans described align with the outputs generated by the model.

on two additional Low Saxon dialect pairs for confirmation (Appendix C.2). All experiments showed similar patterns, indicating that language-unique lexical features can indeed be retrieved from the explanations. Further discussion about PR can be found in Appendix D.1 and D.2.

Human Evaluation Similar to the baseline in §5.2, we select the top 20 extracted features using our extraction method as described in §3.2. The inter-annotator agreement statistics can be found on Appendix G. We calculate the percentage of choices made in the each option relative to the total choices the and present the CN-TW and SCN-IT results in Table 3 and compare CN-TW results with baseline in Table 4.

Option (%)	CN	TW	Option (%)	SCN	IT
CN	**88.9**	10.5	SCN	**45.0**	0.0
TW	0.0	**89.5**	IT	15.0	**63.2**
Both	11.1	0.0	Both	40.0	36.8
None	0.0	0.0	None	0.0	0.0

Table 3: Our extraction method effectively extracts language-unique features in two language pairs, CN-TW and SCN-TW.

We observe that for CN-TW, which is easily distinguishable, our method captures 88.9% and 89.5% of CN and TW features, respectively. Compare to the baseline in §5.2 (70% and 60%), our method is **27%** and **49%** higher. The numbers for SCN and IT are slightly lower (45% for SCN and 63.2% for IT) but it is important to note that the two languages do share a large percentage of their vocabulary, so we believe that these scores are indeed encouraging. The reason is that the method is confirmed to be rather precise: None of the suggested features for Italian would be appropriate for Sicilian, and only 15% of the suggested SCN features would not be appropriate for it.

	Baseline		Ours	
Option (%)	CN	TW	CN	TW
CN	70.0	0.0	**88.9**	10.5
TW	0.0	60.0	0.0	**89.5**
Both	30.0	40.0	11.1	0.0
None	0.0	0.0	0.0	0.0

Table 4: Comparison of the baseline and our extraction method on capturing language-unique features for CN-TW. Our method significantly outperforms the baseline, capturing nearly **90%** of the language-unique features for both languages.

5.5 Classification Accuracy

While our primary goal is to extract lexical features, ensuring high classification accuracy is also crucial as incorrect predictions could undermine explanations. We train 21 distinct models for binary classification for all dialect pairs and present their results on Appendix E. We observe our method achieves high accuracy across all language pairs, with an average of 98.7%. This ensures that the features extracted by the model are supported by reliable predictions, thereby enhancing the value and reliability of the explanations it provides.

6 Conclusion

In this work, we introduce a novel approach for capturing language-unique lexical features from dialects through interpretable dialect classifiers. We utilize both post-hoc and intrinsic explanation methods and experiment on three language groups - Mandarin, Italian, and Low Saxon, and their respective dialects, conducting extensive evaluation and analysis to showcase the effectiveness of our method. In the future, we plan to broaden this approach to address additional linguistic aspect beyond the lexical level (Appendix D.3).

More broadly, our paper takes a first step to assess how interpretability techniques can be used to unearth lexical and, potentially, other linguistic features. By doing so, we hope to provide a framework that future studies can build upon.

7 Ethics Statement

We envision a future where researchers, regardless of their expertise in a particular language or dialect, can leverage our method to gain insights of dialectal variations. While our primary intention is to promote dialectal inclusivity, we realize that, like any tool, it could be misused in ways that might lead to division or stereotyping. As dialects can be associated with specific ethnic groups or nationalities, the technology might be misused for profiling purposes. It might introduce misleading correlations and negatively impact certain groups.

8 Limitations

A limitation to our work is that we are working with binary classification and data that we *a priori* know to belong to a certain language variety, to provide a proof-of-concept. If one was applying this work in the real world, e.g. on data collected from multiple locations within a language's geographical area, we

could substitute our classification scheme to now predict the location or class of the data collection.

While our emphasis is on lexical aspects, it is crucial to acknowledge the broad spectrum of other linguistic elements that contribute to the richness and complexity of dialects, such as syntactic, phonetic, and semantic features. Our method demonstrates excellent performance in lexical feature extraction, it may not yet adeptly identify and analyze these additional facets of linguistic variation. Additionally, finding annotators proficient in multiple dialects is challenging. Therefore, some experiments were only conducted on a subset of languages due to the limited availability of data and annotators. In the future, we plan to extend our method to more languages and integrate modules that will focus specifically on these diverse linguistic features, making it a more comprehensive tool for dialect analysis.

9 Acknowledgements

The authors are thankful for the anonymous reviewers for their feedback. This work was generously supported by NSF Award IIS-2125466 as well as through a GMU OSCAR award for undergraduate research. This research is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract #2022-22072200004. This material is also funded by the DARPA Grant under Contract No. HR001120C0124. We also gratefully acknowledge support from NSF CAREER Grant No. IIS2142739, NSF Grants No. IIS2125201, IIS2203097, and the Alfred P. Sloan Foundation Fellowship. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Noëmi Aepli, Antonios Anastasopoulos, Adrian-Gabriel Chifu, William Domingues, Fahim Faisal, Mihaela Gaman, Radu Tudor Ionescu, and Yves Scherrer. 2022. Findings of the VarDial evaluation campaign 2022. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–13, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Noëmi Aepli, Çağrı Çöltekin, Rob Van Der Goot, Tommi Jauhiainen, Mourhaf Kazzaz, Nikola Ljubešić, Kai North, Barbara Plank, Yves Scherrer, and Marcos Zampieri. 2023. Findings of the VarDial evaluation campaign 2023. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 251–261, Dubrovnik, Croatia. Association for Computational Linguistics.

David Alvarez-Melis and Tommi S. Jaakkola. 2018. Towards robust interpretability with self-explaining neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 7786–7795, Red Hook, NY, USA. Curran Associates Inc.

Sercan Ömer Arik and Tomas Pfister. 2020. Protoattend: Attention-based prototypical learning. *J. Mach. Learn. Res.*, 21:210:1–210:35.

Yonatan Belinkov, Sebastian Gehrmann, and Ellie Pavlick. 2020. Interpretability and analysis in neural NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 1–5, Online. Association for Computational Linguistics.

Eric Brill. 1991. Discovering the lexical features of a language. In *29th Annual Meeting of the Association for Computational Linguistics*, pages 339–340.

Jack K Chambers and Peter Trudgill. 1998. *Dialectology*. Cambridge University Press.

J.K. Chambers and P. Trudgill. 1980. *Dialectology*. Cambridge Studies in Oral and Literate Culture. Cambridge University Press.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Ryan Cotterell and Chris Callison-Burch. 2014. A multi-dialect, multi-genre corpus of informal written Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 241–245, Reykjavik, Iceland. European Language Resources Association (ELRA).

Dorottya Demszky, Devyani Sharma, Jonathan Clark, Vinodkumar Prabhakaran, and Jacob Eisenstein. 2021. Learning to recognize dialect features. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2315–2338, Online. Association for Computational Linguistics.

Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O Riedl. 2019. Automated rationale generation: a technique for explainable ai and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 263–274.

Koustava Goswami, Rajdeep Sarkar, Bharathi Raja Chakravarthi, Theodorus Fransen, and John P. McCrae. 2020. Unsupervised deep language and dialect identification for short texts. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1606–1617, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Shirley Anugrah Hayati, Kyumin Park, Dheeraj Rajagopal, Lyle Ungar, and Dongyeop Kang. 2023. StyLEx: Explaining style using human lexical annotations. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2843–2856, Dubrovnik, Croatia. Association for Computational Linguistics.

Oliver Hellwig, Sebastian Nehrdich, and Sven Sellmer. 2023. Data-driven dependency parsing of vedic sanskrit. *Language Resources and Evaluation*, pages 1–34.

Shengyuan Hou, Jushi Kai, Haotian Xue, Bingyu Zhu, Bo Yuan, Longtao Huang, Xinbing Wang, and Zhouhan Lin. 2022. Syntax-guided localized self-attention by constituency syntactic distance. *arXiv preprint arXiv:2210.11759*.

Alon Jacovi, Oren Sar Shalom, and Yoav Goldberg. 2018. Understanding convolutional neural networks for text classification. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 56–65, Brussels, Belgium. Association for Computational Linguistics.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research*, 65:675–782.

Xisen Jin, Junyi Du, Zhongyu Wei, Xiangyang Xue, and Xiang Ren. 2019. Towards hierarchical importance attribution: Explaining compositional semantics for neural sequence models. *CoRR*, abs/1911.06194.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1885–1894. PMLR.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4765–4774.

Edoardo Mosca, Ferenc Szigeti, Stella Tragianni, Daniel Gallagher, and Georg Groh. 2022. SHAP-based explanation methods: A review for NLP interpretability. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4593–4603, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jelena Prokić, Çağrı Çöltekin, and John Nerbonne. 2012. Detecting shibboleths. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 72–80.

Dheeraj Rajagopal, Vidhisha Balachandran, Eduard H Hovy, and Yulia Tsvetkov. 2021. SELFEXPLAIN: A self-explaining architecture for neural text classifiers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 836–850, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938.

Mattia Rigotti, Christoph Miksovic, Ioana Giurgiu, Thomas Gschwind, and Paolo Scotton. 2021. Attention-based interpretability with concept transformers. In *International Conference on Learning Representations*.

Parker Riley, Timothy Dozat, Jan A. Botha, Xavier Garcia, Dan Garrette, Jason Riesa, Orhan Firat, and Noah Constant. 2022. FRMT: A benchmark for few-shot region-aware machine translation. *CoRR*, abs/2210.00193.

Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2018. Fine-grained Arabic dialect identification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1332–1344, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yves Scherrer, Tommi Jauhiainen, Nikola Ljubešić, Preslav Nakov, Jörg Tiedemann, and Marcos Zampieri. 2022. Proceedings of the ninth workshop on nlp for similar languages, varieties and dialects. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*.

Yves Scherrer, Tommi Jauhiainen, Nikola Ljubešić, Preslav Nakov, Jörg Tiedemann, and Marcos Zampieri. 2023. Tenth workshop on nlp for similar languages, varieties and dialects (vardial 2023). In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3145–3153. JMLR.org.

Janine Siewert, Yves Scherrer, Martijn Wieling, and Jörg Tiedemann. 2020. LSDC - a comprehensive dataset for low Saxon dialect classification. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 25–35, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034.

Aarohi Srivastava and David Chiang. 2023. Fine-tuning BERT with character-level noise for zero-shot transfer to dialects and closely-related languages. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 152–162, Dubrovnik, Croatia. Association for Computational Linguistics.

Mo Yu, Shiyu Chang, Yang Zhang, and Tommi Jaakkola. 2019. Rethinking cooperative rationalization: Introspective extraction and complement control. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4094–4103, Hong Kong, China. Association for Computational Linguistics.

Omar F. Zaidan and Chris Callison-Burch. 2011. The Arabic online commentary dataset: an annotated dataset of informal Arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 37–41, Portland, Oregon, USA. Association for Computational Linguistics.

A Dialects

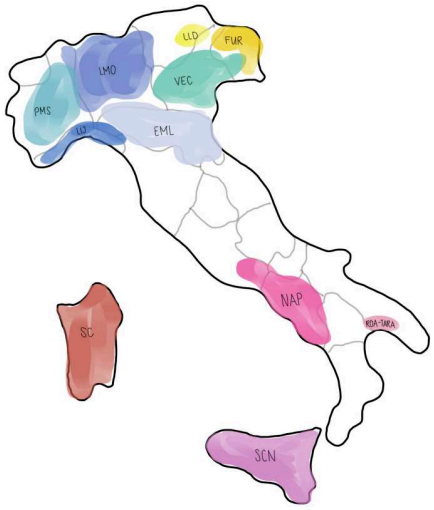

Figure 3: A general overview of the geographical areas in Italy for the 11 languages and dialects. While the map's vague due to the complexity of the situation, it provides a rough idea of where in Italy to locate the varieties. The map is sourced from Aepli et al. (2022).

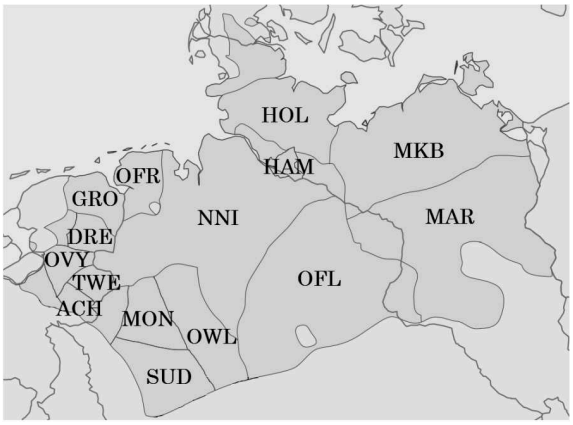

Figure 4: Rough regions where the 16 considered Low Saxon languages and dialects are spoken. This map is taken from Siewert et al. (2020).

Mandarin Dialects Mandarin, also known as Standard Chinese or Putonghua, is part of the Sinitic language family and one of the most widely spoken languages worldwide. We focus on two variations of Mandarin: Mainland Mandarin (CN) and Taiwan Mandarin (TW). The two varieties are closely related and they share a core vocabulary, but there are variations in the usage of certain words and phrases. In writing, Mainland Mandarin also has adopted simplified Chinese characters, while Taiwan uses traditional characters.

Italian Varieties Italian, a Romance language, consists of a diverse range of dialects across different regions of Italy. While these dialects have evolved from Latin and share many common words, they exhibit variations in phonetics, vocabulary, and grammar (Figure 3). For example, Sicilian, spoken in Sicily and the southern regions of Italy, presents distinct phonological features and a rich vocabulary influenced by Arabic and other languages, and as a result it has its own grammar, vocabulary, and pronunciation rules, to the point that it can be difficult for Italian speakers to understand. We note that a lot of the Italian vernaculars are categorized as distinct languages with their own ISO codes (e.g. Venetian, Neapolitan, Sicilian, to name a few). But nevertheless it is undeniable that all of them fall within the same branch of Italic languages and in practice for a diverse language continuum.

Low Saxon Dialects Low Saxon is a West Germanic language that encompasses a range of dialects spoken in the northern regions of the Netherlands and Germany (Figure 4). There are multiple dialects that vary distinctly across different regions. For example, the dialects spoken in the Netherlands, such as Gronings and Twents, have distinct vowel sounds and consonant variations compared to the German dialects, such as Plattdeutsch.

B Dataset Statistics

Language	Dialect Region	Train	Test
Mandarin	Mainland (CN)	3802	467
	Taiwan (TW)	3807	488
Italian	Piemonte (PMS)	3305	414
	Veneto (VEC)	11249	1447
	Sicilia (SCN)	3250	399
	Campania (NAP)	2012	254
	Emilia Romagna (EML)	1648	222
	Taranto (ROA TARA)	716	90
	Sardegna (SC)	810	99
	Liguria (LIJ)	4575	558
	Friuli (FUR)	2990	368
	Veneto (LID)	-	-
	Lombardia (LMO)	5846	733
Low Saxon	Achterhoek (ACH)	791	106
	Drenthe (DRE)	5322	634
	Groningen (GRO)	27	2
	Hamburg (HAM)	5559	705
	Holstein (HOL)	10381	1293
	Mark-Brandenburg (MAR)	177	20
	Mecklenburg-Vorpommern (MKB)	15654	1913
	Munsterland (MON)	589	86
	Northern Lower Saxony (NNI)	649	80
	Lower Prussia (NPR)	298	33
	Eastphalia (OFL)	7512	952
	East Frisia (OFR)	197	25
	Overijssel (OVY)	1063	113
	Eastern Westphalia (OWL)	11480	1396
	Sauerland (SUD)	13747	7425
	Twente (TWE)	547	59

Table 5: Dataset Statistics

C PR Results

C.1 PR on CN-TW

		LOO		SelfExp.	
Label	Feature	CN PR	TW PR	CN PR	TW PR
CN	菠萝	12.5	0	12.5	0
	鼠标	38.5	0	15.4	0
	新西兰	50	0	66.7	0
	打印机	33.3	0	33.3	0
	站台	0	0	0	0
	过山车	57.1	0	14.3	0
	三文鱼	100	0	100	0
	洗发水	28.6	0	28.6	0
	软件	60	0	0	0
	悉尼	0	0	50	0
	回形针	50	0	33.3	0
Avg. PR		***39.1***	***0***	***35***	***0***
TW	鳳梨	0	33.3	0	55.6
	滑鼠	0	16.7	0	0
	紐西蘭	0	42.9	0	14.3
	印表機	0	83.3	0	50
	月台	0	50	0	16.7
	雲霄飛車	0	33.3	0	33.3
	鮭魚	0	0	0	60
	洗髮精	0	0	0	100
	軟體	0	30	0	20
	雪梨	0	0	0	0
	迴紋針	0	75	0	25
Avg. PR		***0***	***33.1***	***0***	***34.1***

Table 6: Each language-unique feature predominantly appears only in its own class, implying that explanation methods are capable to extract language-unique features. For example, the CN word '菠萝' is only used in CN's explanation, which is never used as TW's explanation, and vice versa for its translation '鳳梨'.

C.2 PR on Low Saxon Dialects

We conduct two additional PR experiments in Low Saxon dialects: *(i)* The first and second singular pronouns in Eastphalian (OFL) are 'mik/dik', compared to the rest of the Low Saxon dialects ('mi/di') (Siewert et al., 2020) (Table 7); *(ii)* The differences between all German and Dutch varieties' orthography for 'house' ('Huus' vs 'hoes') and 'for' ('för' vs 'veur') (Table 8). In both experiments, we find similar general patterns where language-unique features tend to appear in their corresponding classes, further enhancing our finding that such dialectal lexical features can be extracted by interpreting dialect classifiers.

Count	OFL		Non-OFL	
	mik	*dik*	*mi*	*di*
OFL Exp.	17	9	0	0
Non-OFL Exp.	1	1	156	57
Text	34	19	577	155
OFL PR (%)	50.0	47.4	0.0	0.0
Non-OFL PR (%)	2.9	5.3	27.0	38.0

Table 7: The OFL words *'mik/dik'* are observed in OFL's explanations about half the time (PR=50%), whereas the Non-OFL words *'mi/di'* mainly appear in Non-OFL's explanations, indicating that explanation methods can effectively capture language-specific features.

Count	DE Feature		NL Feature	
	Huus	för	hoes	veur
DE Exp.	17	42	0	0
NL Exp.	0	0	0	23
Text	42	175	0	63
DE PR (%)	40.5	24.0	0.0	0.0
NL PR (%)	0.0	0.0	0.0	36.5

Table 8: In Low Saxon, German (DE) dialects *'house'* are written *'Huus'* and *'for'* is written as *'för'*, whereas in Dutch (NL) are written as *'hoes'* and *'veur'* (Siewert et al., 2020). Note that *hoes* does not appear on the NL corpus.

D Additional Discussion

D.1 Why is Pick-Up Rate Low?

In the CN-TW experiments, the language-unique lexical features (see Table 6) are always correctly identified, which means that our method has high precision. The fairly low pick-up rates, though, imply that our method has somewhat low recall. To test if this is indeed the case, we explore whether there exist *other* lexical features that may also be language-unique but which are not part of our original feature list. To do so, we find all features that co-occur with the features in our list for both varieties, and rank them based on their counts. We then present this updated list of possible language-unique features to our annotators (in a manner similar to §5.4). We find that the majority of them are indeed good candidates for language-unique features as well. In particular, 74.2% of the ones selected for CN and 68.3% of the ones for TW are indeed unique to the respective language. This implies that the apparent low recall of our method is simply due to the presence of many good options in the data – and the feature lists in Table 6 contain only a subset of the possible explanations.

D.2 Explanations from the Training Sets

We explore how expanding the scale of data points influences the explanation methods' ability to capture features. Therefore, we run the classifier on the training set rather than the test set, to collect more explanations. We conduct experiments on LSDC dataset, similar to §5.4. Comparing the results, presented in Table 9 with the original test set results in Table 7, we find a substantial increase in the number of data points along with noticeable fluctuations in PRs. Despite these variations, the language-specific features continue to mainly appear within their respective predicted classes. This observation reaffirms our findings that the language-unique lexical features can indeed be captured with high precision by explanations.

Count	OFL		Non-OFL	
	mik	*dik*	*mi*	*di*
OFL Exp.	92	44	0	0
Non-OFL Exp.	2	9	1422	493
Text	263	150	4895	1310
OFL PR (%)	35.0	29.3	0.0	0.0
Non-OFL PR (%)	0.76	6.0	29.0	38.0

Table 9: Despite an increase in data points and fluctuations in PR values, language-specific features consistently appear predominantly within their respective predicted classes. This observation strengthens our previous findings, reaffirming that the explanation method can indeed effectively capture language-unique features, regardless of the scale of data points.

Count	DE Feature		NL Feature	
	Huus	för	hoes	veur
DE Exp.	101	388	3	0
NL Exp.	2	0	0	182
Text	327	1492	5	559
DE PR (%)	30.1	26.0	6.0	0.0
NL PR (%)	0.6	0.0	0.0	33.0

Table 10: Evaluating explanations from the training sets for DE vs NL.

D.3 Other Features

While our primary focus is extracting lexical features from dialects, we also explore extracting sub-word features. We conducted a study using the LSDC dataset (Siewert et al., 2020), which contains examples where the plural suffix of verbs in the present tense differs among dialects. In dialects MKB, MAR, NPR, OFR, and GRO (class 1), the plural suffix is *-(e)n*, while in the rest of the dialects in the dataset (class 0), it is *-(e)t*. We counted the occurrences of these two suffixes in the text and used PR to evaluate whether explanation models can recognize these subtle, language-specific features. Table 11 illustrates the results. For class 1 the models do indeed return features with its unique feature *-(e)n* with a relatively high PR (20%) and it (correctly) does not return features for class 0 (only 1% PR).

The other type of ending (*-(e)t*), on the other hand, is not returned as part of the model explanations for class 0. We hypothesize that this discrepancy is due to feature overload: several other words in these dialects, which are not present-tense verbs, have the same ending. To accurately capture these sub-word features, further investigation is necessary, along with the development of morphological and morphosyntactic analysis tools for these dialects, which extends beyond the scope of our current work.

Count	*-et* (C.0)	*-en* (C.1)
C.0 Exp.	39	209
C.1 Exp.	904	4046
Text	3772	21112
C.0 PR (%)	1.0	1.0
C.1 PR (%)	24.0	19.2

Table 11: The PR for the class 1 unique feature *-en* is higher within its class, but the model's recognition of the subtle morphological distinction is unclear, given the PR for *-et* is 1% within its class. This could be attributed to the inclusion of non-present-tense verbs with *-et* endings.

E Classification Results

	Dialect	Accuracy	
		Baseline	Ours
Italian	EML	98.3	99.1
	FUR	99.7	99.8
	LIJ	99.7	100.0
	LMO	99.7	99.3
	NAP	100.0	99.2
	PMS	100.0	99.7
	ROA	100.0	100.0
	SC	98.5	97.4
	SCN	99.7	99.2
	VEC	99.7	99.4
Low Saxon	ACH	93.6	99.1
	DRE	97.7	98.6
	HAM	94.1	95.0
	HOL	96.0	93.8
	MAR	83.3	99.8
	MKB	96.5	96.8
	MON	85.7	99.1
	NPR	87.0	99.6
	OFL	97.7	98.3
	OVY	87.0	99.5
Mandarin	TW	99.3	99.4

Table 12: Pairwise test set accuracy of baseline classifiers versus interpretable classifiers. Our method achieve generally high accuracy for all language pairs. We see equally high performance on Italian and Mandarin dialects. However we see a disparity in performance across Low Saxon dialects.

F Feature Counts for CN-TW PR

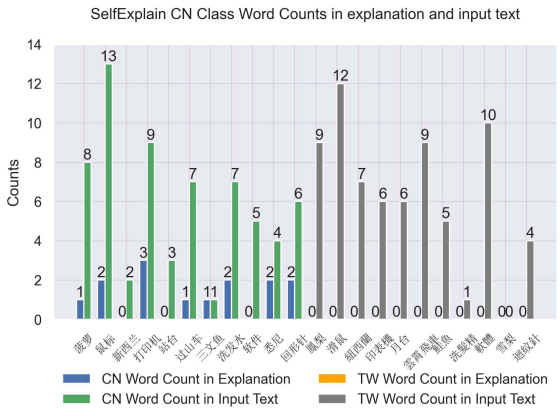

Figure 5: SelfExplain CN Class feature counts in explanation and input text.

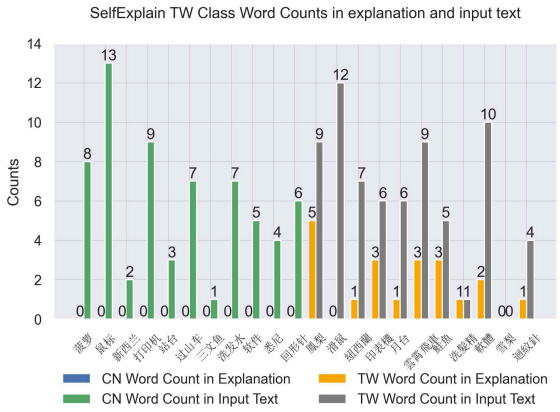

Figure 6: SelfExplain TW Class feature counts in explanation and input text.

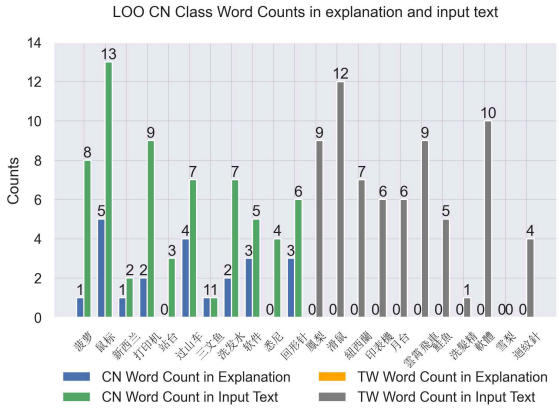

Figure 7: LOO CN Class feature counts in explanation and input text.

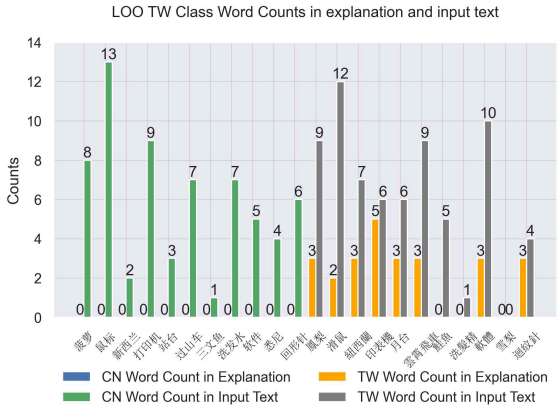

Figure 8: LOO TW Class feature counts in explanation and input text.

G Inter-annotator Agreement Statistics

To minimize potential biases, we mixed the features between both classes.

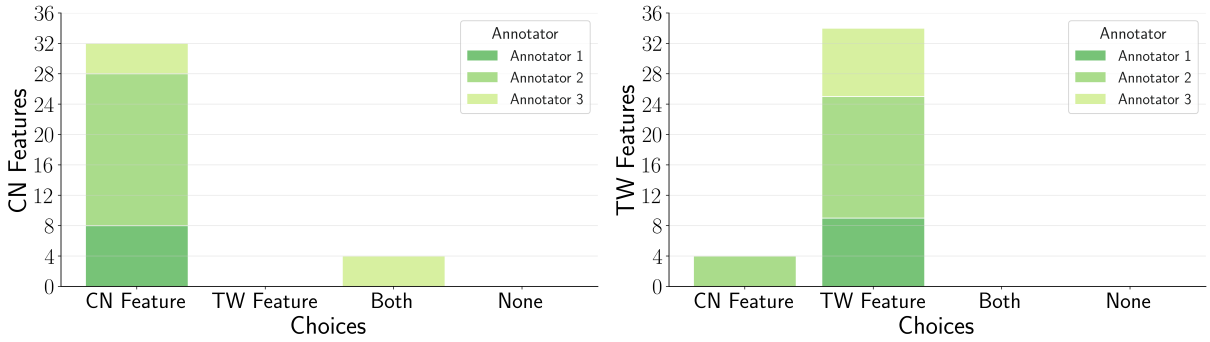

Figure 9: Inter-annotator agreement statistics on extracted CN features (left) and TW features (right). Most extracted features align with human annotators.

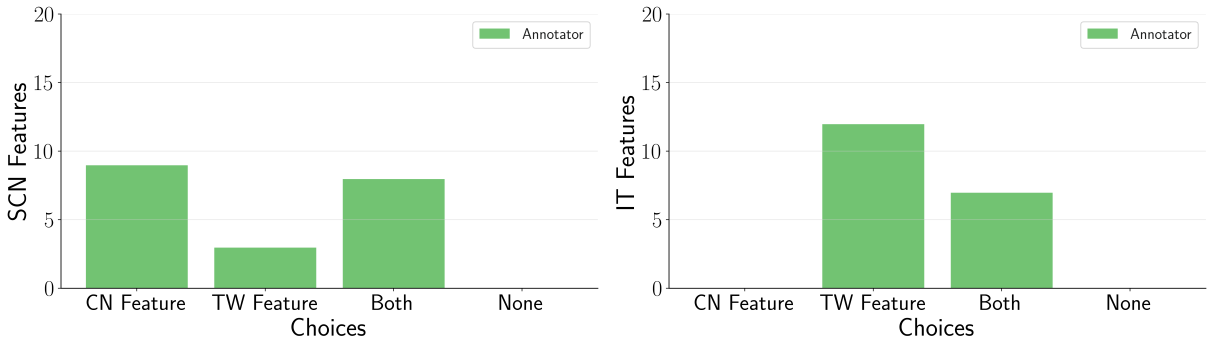

Figure 10: Inter-annotator agreement statistics on SCN (left) and IT (right) features. Note that there is only one annotator in SCN-IT experiment.