# Linguistic Fingerprint in Transformer Models: How Language Variation Influences Parameter Selection in Irony Detection

**Michele Mastromattei**[1,2], **Fabio Massimo Zanzotto**[2]
[1] Campus Bio-Medico University of Rome, Italy, [2] University of Rome Tor Vergata, Italy
michele.mastromattei@{unicampus, uniroma2}.it

**Abstract**

This paper explores the correlation between linguistic diversity, sentiment analysis and transformer model architectures. We aim to investigate how different English variations impact transformer-based models for irony detection. To conduct our study, we used the EPIC corpus to extract five diverse English variation-specific datasets and applied the KEN pruning algorithm on five different architectures. Our results reveal several similarities between optimal subnetworks, which provide insights into the linguistic variations that share strong resemblances and those that exhibit greater dissimilarities. We discovered that optimal subnetworks across models share at least 60% of their parameters, emphasizing the significance of parameter values in capturing and interpreting linguistic variations. This study highlights the inherent structural similarities between models trained on different variants of the same language and also the critical role of parameter values in capturing these nuances.

**Keywords:** Explainable models, language variation, irony detection, model optimization

## 1. Introduction

Sentiment analysis datasets, particularly those annotated on crowdsourcing platforms, may contain biases due to the lack of information about the cultural backgrounds of the annotators. This can lead to machine learning models trained on this data amplifying these biases, affecting how people perceive and label sentiment. Although these models can capture general sentiment, they often fail to capture the nuances experienced by different groups.

This paper examines the impact of linguistic diversity on transformer models designed for irony detection. Using the EPIC corpus (Frenda et al., 2023), we created five subsets tailored to different variations of English. We trained different transformer models and used the KEN pruning algorithm (Mastromattei and Zanzotto, 2024) to extract the minimum subset of optimal parameters that maintain the original performance of the model. We conducted this experimental process across five transformer architectures, revealing a minimum parameter overlap of 60% among resulting subnetworks. We then performed a comprehensive analysis to identify subnetworks with the highest and lowest similarity. Additionally, we used $KEN_{viz}$ for a visual examination of pattern similarities. Our results show that the linguistic variation is closely related to the individual values of each parameter within the models. This suggests that the diversity among linguistic variation is not just a structural aspect, but is deeply rooted in the specific values contained in the model. These insights can help create models that better capture the richness of linguistic variation and address bias effectively.

## 2. Background and related work

Artificial intelligence (AI) models impact our daily lives in many ways. Some applications go beyond just processing data and strive to understand the intricate human elements and cultural nuances of our world. For instance, sentiment analysis requires a deeper understanding of implicit phrases and cultural differences to accurately interpret emotions (Tourimpampa et al., 2018; Sun et al., 2022). This is why rigorous studies are essential before deploying data and models in real-world settings. When creating data, it is crucial to incorporate different perspectives evaluation standards, such as "golden standards" (Basile et al., 2021), incorporating criteria for evaluating annotators (Miłkowski et al., 2021; Abercrombie et al., 2023; Mieleszczenko-Kowszewicz et al., 2023), grouping them according to potential bias factors (Fell et al., 2021) or using text visualization techniques to analyze annotated datasets (Havens et al., 2022). On the model level, explainable AI (XAI) techniques (Samek et al., 2017; Samek and Müller, 2019; Vilone and Longo, 2021) are being used to demystify complex models and ensure transparency. Many neural interpretability models rely on attention-based techniques (Bodria et al., 2020), utilizing auxiliary tasks (De Sousa Silveira et al., 2019), or external knowledge integration (Zhao and Yu, 2021). Moreover, attention-based models exhibit a grasp of the syntactic structure of analyzed sentences (Manning et al., 2020). Consequently, the role of syntax in model interpretation is being extensively studied across various domains, including irony (Cignarella et al., 2020) and hate speech (Mastromattei et al., 2022b,a). This multifaceted exploration contributes to a richer under-



*NLPerspectives 2024@LREC-COLING 2024*, pages 123–130

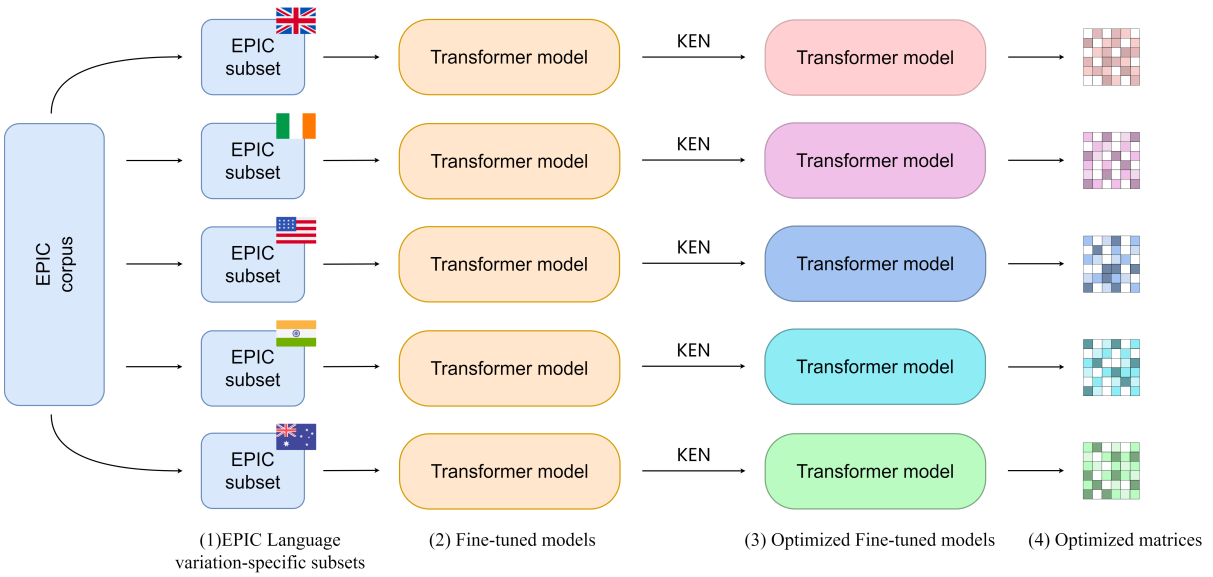



Figure 1: Workflow overview. Specific language variations are selected from the EPIC corpus (1). For each unique language subset, a dedicated transformer model is trained (2). This ensures that each model specializes in the intricacies of its assigned language variation. Finally, the KEN pruning algorithm is applied to optimize the trained models (3). This involves efficient and lightweight architectures for each language variant (4).

standing of the interplay between language, culture and model interpretability to achieve increasingly inclusive AI models.

## 3. Methods and Data

This section introduces the core components of our research: the EPIC corpus and the KEN pruning algorithm. Sec. 3.1 provides an in-depth exploration of the EPIC corpus, explaining its composition and the diverse language varieties it encompasses. Sec. 3.2 analyzes the KEN pruning algorithm, emphasizing its key role in transformer model optimization.

### 3.1. EPIC Corpus

The EPIC (Frenda et al., 2023) corpus consists of 3,000 conversations from social media platforms. It covers five different varieties of English, including Australian (AU), British (GB), Irish (IE), Indian (IN) and American (US). The corpus offers valuable insights into how cultural and linguistic factors shape the perception of irony, giving a comprehensive analysis of it from different perspectives.

To ensure the authenticity of the data, EPIC sources its content from Twitter and Reddit, capturing informal communication across different regions and demographic areas. Rigorous data curation guarantees the inclusion of potential ironies while maintaining a balanced distribution across language varieties, mitigating selection bias. Native speakers from each country independently label instances as ironic or non-ironic, using a multi-perspective annotation process. This ensures a robust and nuanced understanding of cultural humor. Annotators possess robust language skills and familiarity with online communication styles, reinforcing the reliability of their judgments. The inclusive approach in both data collection and annotation facilitates the development of *perspective-aware* models (Akhtar et al., 2021) that account for cultural and linguistic variations.

### 3.2. KEN algorithm

KEN (**K**ernel density **E**stimator for **N**eural network compression) (Mastromattei and Zanzotto, 2024), is a pruning algorithm designed to extract the most essential subnetwork from transformer models. It exploits the *winning ticket lottery hypothesis* (Frankle and Carbin, 2018), according to which an optimal subset of fine-tuned parameters maintains the same performance as the original one.

KEN leverages Kernel Density Estimations (KDEs) to generalize point distributions for each row of a transformer matrix, resulting in a streamlined version of the original fine-tuned model. By pinpointing the $k$ most representative parameters within each distribution, KEN effectively prunes the network, preserving them while reverting the remaining parameters to their pre-trained state. KEN archives minimum parameter reduction between 25% and 60% for specific models, maintaining equivalent or better performance than their unpruned counterparts. The resultant subnetwork

| Model | AU | GB | IE | IN | US |
|---|---|---|---|---|---|
| Bert | 47.54 | 58.03 | 58.03 | 58.03 | 58.03 |
| DistilBert | 56.26 | 34.39 | 50.79 | 50.79 | 56.26 |
| DeBerta | 44.88 | 55.91 | 55.91 | 55.91 | 55.91 |
| Ernie | 58.03 | 47.54 | 58.03 | 58.03 | 58.03 |
| Electra | 91.18 | 91.18 | 64.75 | 91.18 | 82.37 |

(a) Pertentage of parameter reset after the KEN pruning step for all the models on each language variation subsets analyzed. The percentage indicates the number of parameters reset to their pre-trained value in the entire model

| Model | AU | GB | IE | IN | US |
|---|---|---|---|---|---|
| Bert | +2.0 | +2.1 | +5.5 | +4.6 | +0.0 |
| DistilBert | +0.6 | +0.0 | +3.5 | +2.4 | +0.0 |
| DeBerta | +1.3 | +2.9 | +7.2 | +1.4 | +0.0 |
| Ernie | +0.0 | +0.0 | +0.0 | +13.5 | +0.0 |
| Electra | +5.2 | +0.7 | +1.5 | +0.1 | +2.1 |

(b) Variation of the F1-weighted measure across all the language variation subsets after the KEN pruning step. Positive values indicate a score improvement compared to the unpruned version

Table 1: Result obtained during our experiment: Tab. 1a shows the percentage of parameter reset of each model in all language variation subsets analyzed while Tab. 1b presents per F1-weighted performance variation obtained.

can be seamlessly archived and reintegrated into its pre-trained configuration for diverse downstream applications. This approach not only significantly reduces model size but also enhances efficiency and flexibility across various tasks.

## 4. Experiments

This section provides a detailed explanation of the entire process we followed during our experiment. The process began with the variant-specific datasets extraction to the optimal subnetworks search and the transformer architecture tested.

The EPIC corpus contains approximately 3,000 sentences annotated by multiple annotators, resulting in 14,172 records. To create language-variant-specific datasets, we distilled unique sentences from the corpus and applied majority voting based on annotations, with ties resolved by labeling records as "irony." This meticulous process yielded well-balanced datasets, each comprising approximately 600 records.

Five models, each specializing in a single language variant, were trained using the same transformer architecture. After fine-tuning, we used the KEN pruning algorithm to extract the smallest and most efficient subnetwork in each model. This process involves incrementally increasing the number of fine-tuning parameters retained and decreasing those restored to pre-training values, starting from a minimal subset of parameters and expanding it until the pruned model performance matches or exceeds its unpruned counterpart. Using these optimized subnetworks, we analyzed the internal structures of the models and measured the similarities between the optimized subnetworks across different language variants. For each layer, we extracted the corresponding matrices and conducted a meticulous analysis of the positions of the optimal parameters within each optimal subnetwork. This involved an *"in-breadth" analysis*, which identified the parameters present in all optimal models examined and *pairwise comparisons* between models to identify the language variants with the greatest and least similarity, regardless of the model architecture. We conducted these analyses for each architecture under examination on the layers that constitute the attention mechanism or similar structures, as these layers concentrate most of the arithmetic operations of the model and are a strength of the transformer model core structure.

We replicate this experiment across five distinct transformer model architectures, including Bert (Devlin et al., 2018), DistilBERT (Sanh et al., 2019), DeBERTa (He et al., 2020), Ernie (Sun et al., 2020) and Electra (Clark et al., 2020). The provided Fig. 1 visually depicts the entire workflow, starting with language variety subset extraction to the resulting optimized subnetworks obtained.

## 5. Results

The KEN algorithm is an effective method for selecting the best model parameters for each language variation. The rate at which these parameters are reset varies across different architectures, as shown in Tab. 1a. However, this resetting rate consistently exceeds 50% on average. Surprisingly, despite the substantial resetting, performance actually improves in most cases, as demonstrated by the F1-weighted scores in Tab. 1b. Notably, these results were achieved through tuning steps on relatively small data sets, with only 600 examples per variation. It is essential to note that our primary goal was not to establish new state-of-the-art (SoTa) models, but rather to investigate the impact of language variations on model parameters within each architecture examined. From this perspective, the results are encouraging and demonstrate a positive impact. Additionally, the varying percentages of parameter resets among linguistic variations using the same architecture contribute to a more nuanced understanding of the optimal subnetworks and their comparison.

After examining subnetwork structures, it was discovered that two optimal subnetworks share at

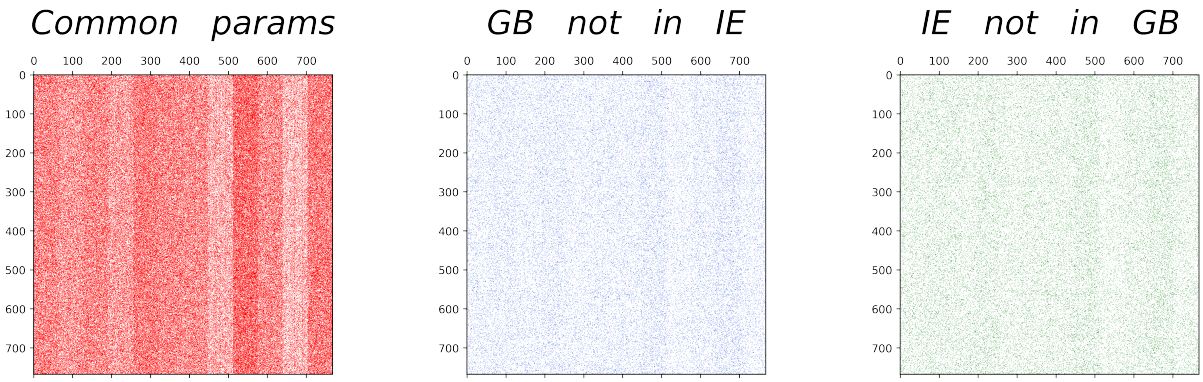



Figure 2: Comparison of the optimal subnetworks of two DeBERTa models (layer 0, attention output matrix) trained on British (GB) and Irish (IE) linguistic variation, respectively. The matrix on the left shows the number of common parameters between the two matrices (subnetwork overlap), while the middle one shows the location of the optimal parameters of the GB subnetwork not present in IE, and on the right the exact opposite. Blank values refer to the values not belonging to the optimal network and thus the collection of points that the KEN algorithm has reset to their pre-training value. Additional results are shown in Apx. A

least 60% of their parameters. This percentage, however, does not take into account parameters reset by KEN, which could significantly impact the final result. Tab. 2 indicates that Indian (IN) and American (US) variations have the highest overlap, with more than 90% in three out of five models. British (GB) and Irish (IE) also have considerable overlap across all models, which is highly desirable. Despite extensive analysis, identifying the most distinct variants remains challenging, as the percentage difference between pairs of language variations across all models is relatively small.

| Subnet A | Subnet B | BERT | DeBERTa | DistilBERT | Ernie | Electra |
|---|---|---|---|---|---|---|
| AU | GB | 69.73 | 69.94 | 61.69 | 69.81 | 89.49 |
| AU | IE | 69.79 | 69.94 | 75.22 | 82.72 | 23.15 |
| AU | IN | 69.73 | 69.94 | 75.17 | 83.22 | 87.6 |
| AU | US | 69.73 | 69.94 | 83.42 | 83.22 | 29.09 |
| GB | IE | 83.02 | 82.74 | 69.38 | 69.76 | 23.15 |
| GB | IN | 82.59 | 82.71 | 69.38 | 69.81 | 86.95 |
| GB | US | 82.59 | 82.71 | 61.66 | 69.81 | 29.06 |
| IE | IN | 82.6 | 82.86 | 85.85 | 82.39 | 23.15 |
| IE | US | 82.6 | 82.86 | 75.17 | 82.39 | 69.68 |
| IN | US | >90.0 | >90.0 | 76.22 | >90.0 | 29.45 |

Table 2: Similarity percentages between subnetworks specific to language variation. Percentages are obtained by comparing for each model the number of non-reset parameters within each attention (or similarity) layers

In addition to tabular descriptions, we have graphically presented the results obtained. Through $KEN_{viz}$, three different types of results are visualized: (1) the subnetwork overlap of two language variations within the same selected matrix layer, (2) fine-tuned parameters chosen for the linguistic variation A but not for B and (3) the reverse. Fig. 2 showcases one of the obtained results, while Apx. A provides more case studies by analyzing results across all models in their last attention layer for specific linguistic variations. These graphical representations offer insights into the precise placement of optimal parameters and the shared or differing structures between models.

## 6. Conclusion

This study conducted a thorough analysis of different transformer models to discover their divergences in detecting irony when trained on different linguistic variants. We uesd the EPIC corpus and created language-variant-specific datasets for five English variations (American, British, Indian, Irish and Australian). Using the KEN pruning algorithm, we extracted optimal subnetworks from five transformer architectures (BERT, DistilBERT, DeBERTa, Ernie and Electra) tailored to each language variation. Our study revealed that different linguistic variations share a remarkable number of parameters, regardless of the architecture used. We provided insights into the similarity of each pair of optimized subnetwork linguistic variations by reporting the percentage of common parameters. However, we found it challenging to rank the dissimilarity since the shared parameter percentage remained consistently high in all cases. To enhance our understanding of how linguistic diversity manifests in the models, we used $KEN_{viz}$ to provide a graphical view of the specific locations of shared and distinct parameters across models and language variations.

Although there are limitations such as the size of the dataset, our study demonstrates that training transformer models and adapting them to linguistic variations yield highly similar output models demonstrating how their difference is intrinsic to their parameter values.

Gavin Abercrombie, Verena Rieser, and Dirk Hovy. 2023. Consistency is key: Disentangling label variation in natural language processing with intra-annotator agreement. *arXiv preprint arXiv:2301.10684*.

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2021. Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection. *arXiv preprint arXiv:2106.15896*.

Valerio Basile, Federico Cabitza, Andrea Campagner, and Michael Fell. 2021. Toward a perspectivist turn in ground truthing for predictive computing. *arXiv preprint arXiv:2109.04270*.

Francesco Bodria, André Panisson, Alan Perotti, Simone Piaggesi, et al. 2020. Explainability methods for natural language processing: Applications to sentiment analysis. In *CEUR Workshop Proceedings*, volume 2646, pages 100–107. CEUR-WS.

Alessandra Teresa Cignarella, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, Paolo Rosso, and Farah Benamara. 2020. Multilingual irony detection with dependency syntax and neural models. *arXiv preprint arXiv:2011.05706*.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Thiago De Sousa Silveira, Hans Uszkoreit, and Renlong Ai. 2019. Using aspect-based analysis for explainable sentiment predictions. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 617–627. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Michael Fell, Sohail Akhtar, and Valerio Basile. 2021. Mining annotator perspectives from hate speech corpora. In *NL4AI@ AI* IA*.

Jonathan Frankle and Michael Carbin. 2018. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*.

Simona Frenda, Alessandro Pedrani, Valerio Basile, Soda Marem Lo, Alessandra Teresa Cignarella, Raffaella Panizzon, Cristina Marco, Bianca Scarlini, Viviana Patti, Cristina Bosco, and Davide Bernardi. 2023. EPIC: Multi-perspective annotation of a corpus of irony. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13844–13857, Toronto, Canada. Association for Computational Linguistics.

Lucy Havens, Benjamin Bach, Melissa Terras, and Beatrice Alex. 2022. Beyond explanation: A case for exploratory text visualizations of non-aggregated, annotated datasets. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, pages 73–82.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Christopher D Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054.

Michele Mastromattei, Valerio Basile, Fabio Massimo Zanzotto, et al. 2022a. Change my mind: how syntax-based hate speech recognizer can uncover hidden motivations based on different viewpoints. In *1st Workshop on Perspectivist Approaches to Disagreement in NLP, NLPerspectives 2022 as part of Language Resources and Evaluation Conference, LREC 2022 Workshop*, pages 117–125. European Language Resources Association (ELRA).

Michele Mastromattei, Leonardo Ranaldi, Francesca Fallucchi, and Fabio Massimo Zanzotto. 2022b. Syntax and prejudice: Ethically-charged biases of a syntax-based hate speech recognizer unveiled. *PeerJ Computer Science*, 8:e859.

Michele Mastromattei and Fabio Massimo Zanzotto. 2024. Less is ken: a universal and simple non-parametric pruning algorithm for large language models. *arXiv preprint arXiv:2402.03142*.

Wiktoria Mieleszczenko-Kowszewicz, Kamil Kanclerz, Julita Bielaniewicz, Marcin Oleksy, Marcin Gruza, Stanisław Wozniak, Ewa Dziecioł, Przemysław Kazienko, and Jan Kocon. 2023. Capturing human perspectives in nlp: Questionnaires, annotations, and biases. In *The ECAI 2023 2nd Workshop on Perspectivist Approaches to NLP. CEUR Workshop Proceedings*.

Piotr Miłkowski, Marcin Gruza, Kamil Kanclerz, Przemyslaw Kazienko, Damian Grimling, and

Jan Kocoń. 2021. Personal bias in prediction of emotions elicited by textual opinions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 248–259.

Wojciech Samek and Klaus-Robert Müller. 2019. Towards explainable artificial intelligence. *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 5–22.

Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Xu Sun, Xiaosong Zhou, Qingfeng Wang, and Sarah Sharples. 2022. Investigating the impact of emotions on perceiving serendipitous information encountering. *Journal of the Association for Information Science and Technology*, 73(1):3–18.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8968–8975.

Aglaia Tourimpampa, Athanasios Drigas, Alexandra Economou, and Petros Roussos. 2018. Perception and text comprehension. it'sa matter of perception! *International Journal of Emerging Technologies in Learning (Online)*, 13(7):228.

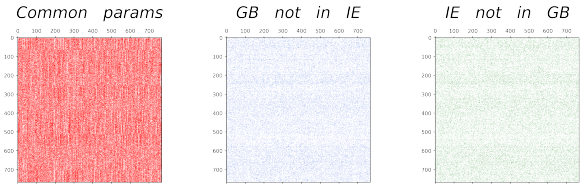
Giulia Vilone and Luca Longo. 2021. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106.

Anping Zhao and Yu Yu. 2021. Knowledge-enabled bert for aspect-based sentiment analysis. *Knowledge-Based Systems*, 227:107220.

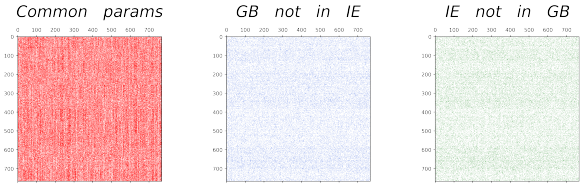
## A. $KEN_{viz}$ outputs

In this appendix, we present some graphical results obtained using $KEN_{viz}$ by analyzing the output of attention matrices in the last levels for each model analyzed. We selected several pairs of linguistic variations for each model that showed the most interesting results based on the findings in Tab.2. These visual results highlight the commonalities found within the optimal subnetworks and show the difficulty of finding differences between them. However, we can observe that in some cases, parameter selection focuses more on certain areas than others.

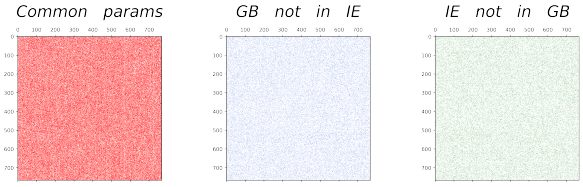
## A.1. Results in DeBerta model



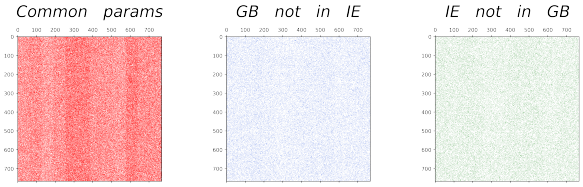

(a) `q_proj` matrices




(b) `pos_proj` matrices



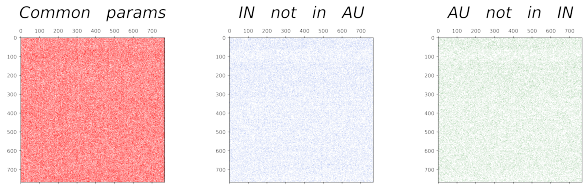

(c) `in_proj` matrices



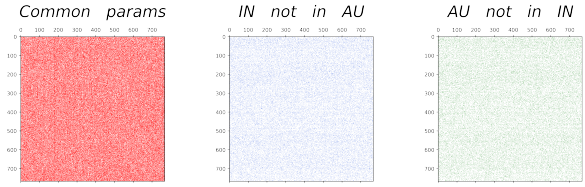

(d) Output matrices

Figure 3: Layer 12

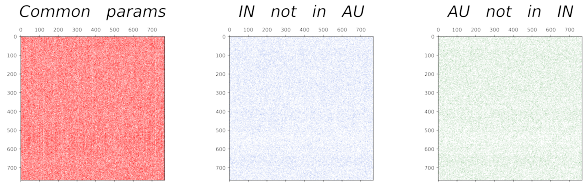
## A.2. Results on Ernie model



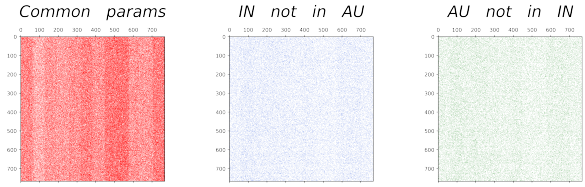

(a) Key matrices




(b) Query matrices



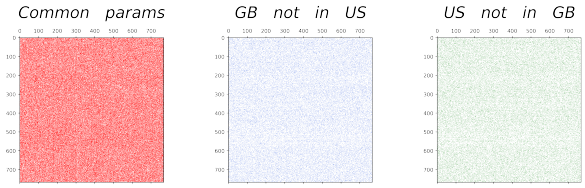

(c) Value matrices



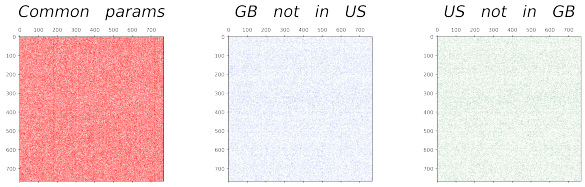

(d) Output matrices

Figure 4: Layer 11

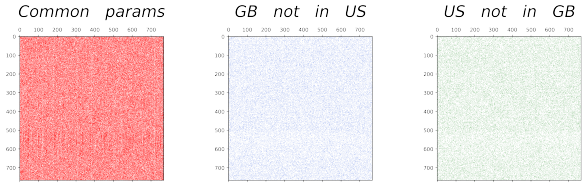
## A.3. Results on BERT model




(a) Key matrices



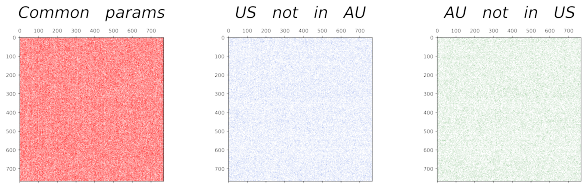

(b) Query matrices



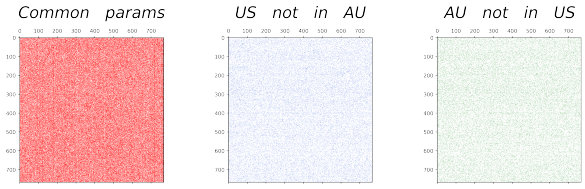

(c) Value matrices

Figure 5: Layer 12

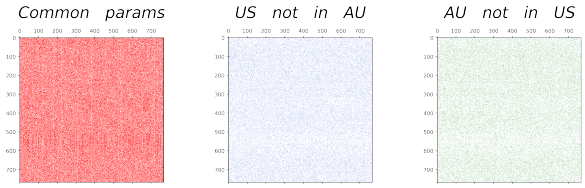
## A.4. Results on DistilBERT model




(a) `k_lin` matrices



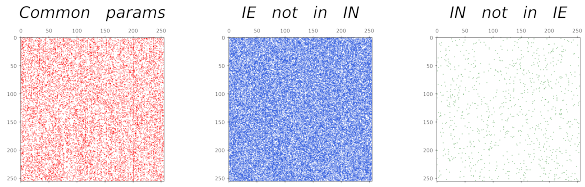

(b) `q_lin` matrices



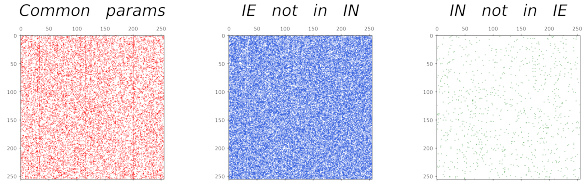

(c) `v_lin` matrices

Figure 6: Layer 5

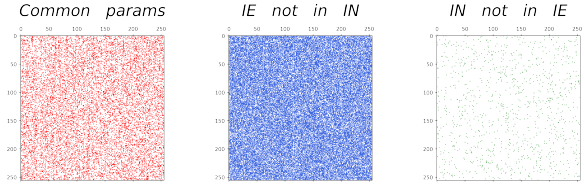
## A.5. Results on Electra model




(a) Key matrices




(b) Query matrices




(c) Value matrices

Figure 7: Layer 12