Lost in Activations: A Neuron-level Analysis of Encoders for Cross-Lingual Emotion Detection

Anonymous ACL submission

Abstract

The rapid advancement of multilingual pre-trained transformers has fueled significant progress in natural language understanding across diverse languages. Yet, their inner workings remain opaque, especially with regard to how individual neurons encode and generalize semantic and affective features across languages. This paper presents an interpretability study of a fine-tuned XLM-R model for multilingual emotion classification. Using neuron-level activation analysis, we investigate the variance of neurons across labels, cross-lingual alignment of activations, and the existence of "polyglot" versus language-specific neurons. Our results reveal that while certain neurons consistently encode emotion-related concepts across languages, others show strong monolingual specialization.

1 Introduction

The last half-decade has witnessed rapid advances in large-scale multilingual pre-trained models such as mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), and multilingual variants of T5 (Xue et al., 2021). While such models have achieved high performance on a number of downstream tasks, our understanding of their internal decision-making is still quite limited. López-Otal et al. (2025) survey multilingual interpretability studies and note that while some neurons encode shared morpho-syntactic features, others show clear typological specialization. In decoder-based multilingual transformers, similar findings emerge: typological or script-related differences cause neurons to form distinct clusters, suggesting that transfer depends on partial overlap between shared and language-specific neurons (Choenni and Shutova, 2022; Kojima et al., 2024). These mixed results highlight a central question in multilingual interpretability: to what extent does a model's cross-lingual generalization stems from universal semantic encoding versus overlapping but language-dependent mechanisms?

Neuron-level interpretability has become a promising direction for understanding deep models. Rather than treating layers as holistic black boxes, this approach aims to assign functional meaning to individual neurons or small neuron groups. N2G (Neuron-to-Graph) (Foote et al., 2023), a scalable method represents neuron behavior as interpretable graphs over input features, reducing manual inspection costs. Similarly, surveys such as Resck et al. (2025) review interpretability methods for multilingual transformers, concluding that most cross-lingual work focuses on probing and attention visualization, while direct neuron-level comparison remains rare.

Our work contributes to this discussion by examining emotion classification, a semantic and affective task rather than a purely syntactic one. By analyzing neuron activations in a fine-tuned XLM-R model across Arabic, Slovenian, and Dutch, we explore whether affective semantics (e.g., *joy*, *anger*, *fear*) are encoded in shared polyglot neurons or in language-specific representations. More specifically, this study aims to (1) identify neurons that differentiate between emotion labels through analysis of the activation variance, (2) compare neuron activations across Arabic, Slovenian, and Dutch to evaluate cross-lingual consistency, and (3) categorize neurons as polyglot (shared across languages) or monolingual (language-specific).

2 Methodology

2.1 Model and Training

We employ the XLM-Roberta (XLM-R) base model (Conneau et al., 2020), a transformer-based multilingual encoder pre-trained on 100 languages with the CC100 corpus (Wenzek et al., 2020). For this study, we use a version of XLM-R that has been fine-tuned for emotion classification. The model

was trained on the English portion of the EXALT dataset (Maladry et al., 2024), a multilingual benchmark containing emotion-annotated sentences spanning multiple typological families. Fine-tuning is performed using a cross-entropy loss function over six emotion labels: *Neutral*, *Joy*, *Love*, *Anger*, *Fear*, and *Sadness*. The final classifier head consists of a single feed-forward layer mapping the [CLS] embedding to the label space. The model achieves a macro-F1 of 0.433, 0.346, and 0.439 on the Dutch, Arabic, and Slovene test sets, respectively.

For interpretability, we focus on the penultimate hidden layer (layer 11), as previous studies have shown that semantic and affective features tend to emerge in the deeper layers of transformers (Rogers et al., 2020). The classification head output probabilities are not used directly in our analysis; instead, we analyze neuron activations in this pre-classification layer to understand which internal units contribute to emotion discrimination and whether these activations generalize across languages. To this purpose, we selected three typologically and morphologically distinct languages from the EXALT development and test splits: Arabic, Dutch, and Slovenian.

2.2 Neuron Activation Extraction

To investigate neuron-level behavior, we extract activations from specific layers of the fine-tuned model using forward hooks. These hooks capture the output tensors of hidden layers during the forward pass without altering model computation. Activations are collected for each sentence in the development and test sets across all three languages. For interpretability purposes, we focus on the [CLS] token representation, which typically encodes global sentence-level meaning and is used by the classifier head for prediction. While individual token embeddings may also carry emotion information, focusing on the [CLS] vector allows for a standardized comparison across languages and labels.

Each [CLS] activation is a 768-dimensional vector corresponding to the neurons in the penultimate layer. We store these vectors along with their associated emotion label and language identifier. To ensure comparability, all activations are z-score normalized within each language and then aggregated per label and per language, producing label-specific and language-specific activation profiles. The resulting activation dataset forms the basis for our subsequent quantitative and visual analyses.

2.3 Interpretability Techniques

Variance Analysis Across Labels. To identify which neurons are most sensitive to emotion distinctions, we compute the variance of activation values across the six emotion labels. High-variance neurons are hypothesized to play a critical role in label discrimination, as they exhibit strong differential activation patterns. This step highlights potentially specialized "emotion neurons", which will be the focus of the following analyses.

Cross-Lingual Neuron Consistency. To test whether the same neurons encode similar information across languages, we compute pairwise correlations between neuron activation vectors across Arabic, Dutch, and Slovenian for equivalent labels. A high correlation indicates that a neuron behaves consistently across languages. We also compute the cosine similarity of averaged activations per label and visualize these using heatmaps. This analysis provides insight into whether emotion-related neurons are language-agnostic or language-specific.

Polyglot vs. Monolingual Neurons. Based on the former analysis, neurons are classified as *polyglot* if their activation patterns for a given emotion label are consistent (i.e., above a 0.5 activation threshold) across all three languages, and as *monolingual* if they show high activation for only one language. We then measure the distribution of these polyglot and monolingual neurons per emotion label to examine which emotions exhibit stronger cross-lingual generalization.

The complete pipeline allows us to move from fine-tuned model activations to quantitative evaluation and visual interpretation of neuron-level multilingual alignment. Together, these methods form a comprehensive framework for examining the internal mechanisms by which XLM-R encodes and transfers emotion semantics across languages.

3 Results

3.1 High-Variance Neurons

Our variance analysis across emotion labels reveals that specific neurons in deeper layers of XLM-R, especially in layer 11 (the penultimate layer), exhibit particularly strong discriminative power. When computing the variance of activation values across the six emotion classes, a small subset of neurons showed substantially higher variance than the layer-wide mean. These neurons are likely to encode emotion-related semantic distinctions.

Figure 1 shows the variances of all neurons in the penultimate layer for the two labels of *Joy* and *Sadness*. These variance plots reveal distinct patterns in how XLM-R encodes affective semantics at the neuron level. For *Joy*, the variance is distributed across numerous neurons, with several moderate peaks spread throughout the 768-dimensional hidden layer. This pattern suggests a distributed representation, where multiple neurons collectively contribute to encoding the concept of joy rather than relying on a single specialized unit. This is consistent with findings from studies on distributed semantics (Rogers et al., 2020), which show that semantic categories are encoded across many units rather than localized "detectors". In contrast, the *Sadness* plot displays a sharply localized spike, indicating the presence of a highly specialized "sadness-sensitive" neuron. Such isolated peaks are rare and indicate the presence of quasi-localized emotion neurons. This contrast implies that different emotions are represented through different neural strategies: while some, like *Joy*, rely on broad, redundant networks of neurons, others, such as *Sadness*, appear to be encoded more sparsely through a small number of highly discriminative units. This aligns with affective neuroscience literature where *Joy* often co-occurs with varied linguistic contexts (metaphorical, situational, or intensifier-based), whereas *Sadness* expressions tend to be more lexically direct and contextually homogeneous (Mohammad and Turney, 2013; Kövecses, 2000). The model may thus require fewer features to capture the latter reliably.

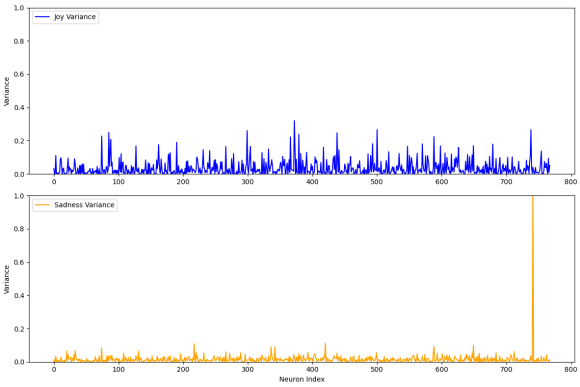

Figure 1: Variance of each neuron's activations in the penultimate layer for the labels *Joy* (top) and *Sadness* (bottom).

3.2 Cross-Lingual Consistency

To assess cross-lingual generalization at the neuron level, we compared activation distributions across the Arabic (AR), Dutch (NL), and Slovenian (SL) development sets for each emotion label. For each neuron, we computed the Spearman correlations between mean activations for the same emotion across languages. Figure 2 shows a heatmap of the results, revealing that a significant proportion of the neurons maintain moderate to high correlation ($\rho > 0.6$) across languages, indicating a shared internal encoding of emotional semantics. However, the strength of cross-lingual alignment varies by language pair. Dutch and Slovenian, which share Indo-European roots and similar syntactic patterns, exhibit higher pairwise correlations ($\rho = 0.94$) than either does with Arabic ($\rho = 0.77$ and $\rho = 0.84$ for AR-NL and AR-SL, respectively). This indicates that typological and orthographic similarity influences the ease of cross-lingual feature transfer, even at the neuron level. The results support the hypothesis that multilingual models develop both shared and language-specific subspaces, modulated by linguistic proximity.

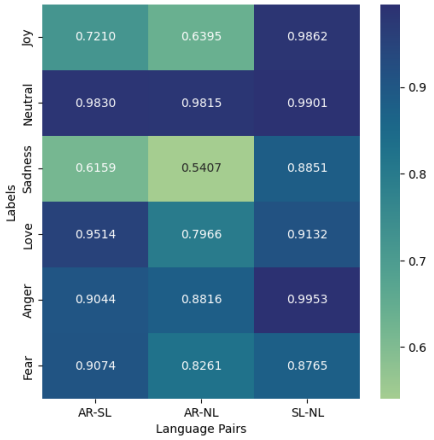

Figure 2: Heatmap of neuron activation correlations across the three language pairs for each emotion label. Darker colors represent higher correlation.

3.3 Polyglot Neurons

The cross-lingual correlation analysis enables classification of neurons into two functional groups: **polyglot neurons** (those that behave consistently across languages) and **monolingual neurons** (those that respond selectively to one language). Polyglot neurons were defined as those exhibiting correlation values above 0.7 across all language pairs for at least one emotion label.

Figure 3 illustrates the average neuron activations for the emotion label *Joy* across the three languages. The overall activation patterns for all three languages are closely aligned, with overlapping curves that indicate strong cross-lingual consistency in how *Joy* is represented within the model. This suggests that the model captures language-independent affective features rather than relying solely on language-specific cues. The highlighted points mark the neurons with the highest activation magnitudes ("top neurons") for each language. Notably, several of these top neurons overlap across languages, implying the existence of polyglot neurons that respond similarly to joyful expressions regardless of linguistic form. Minor deviations among the three lines indicate small degrees of language-specific variation, likely reflecting lexical or syntactic differences in how joy is expressed in each language.

A distinct subset of neurons emerged as polyglot, showing stable and label-specific activations across all three languages. These neurons likely encode universal affective semantics, language-independent concepts strongly associated with certain emotions. Upon closer inspection we found that these polyglot neurons were particularly abundant for emotions with clear physiological or universal expressions (e.g., *Fear*, *Joy*), and less frequent for culturally variable emotions such as *Love*. This pattern aligns with psychological evidence that basic emotions have more cross-cultural consistency than social or relational emotions (Ekman, 1992).

Considering the monolingual neurons, we found that some exhibit clear monolingual behavior, activating selectively for certain labels in only one language. This may reflect differences in the linguistic realization of emotions. For example, Arabic expressions of anger often include intensifiers and idiomatic phrasing absent in European languages (Al-Sheikh, 2014), leading to language-specific activation patterns. These findings emphasize that even in strongly aligned multilingual models, certain emotion-related representations remain grounded in linguistic form rather than universal meaning.

4 Conclusion

The results illustrate the dual nature of multilingual representations within XLM-R. On one hand, the existence of polyglot neurons demonstrates that the model captures *language-agnostic affective features*, allowing it to generalize emotion semantics across diverse linguistic systems. On the other hand, the persistence of monolingual neurons underscores that certain representations remain tied to language-specific patterns. From a quantitative perspective, the distribution of neuron types shows that, across the languages investigated, approximately 35% of emotion-sensitive neurons are polyglot, 45% monolingual, and the remainder are mixed or weakly active. Dutch and Slovenian share more polyglot neurons than either does with Arabic, consistent with their typological similarity and shared Latin script. From an interpretability perspective, neuron-level analysis allows us to move beyond aggregate accuracy metrics and examine how multilingual models internalize meaning. Observing shared neuron activation across languages provides direct evidence of semantic transfer, while identifying language-specific neurons reveals where transfer breaks down. Such analyses can inform model design; for example, encouraging activation alignment of polyglot neurons could enhance zero-shot transfer.

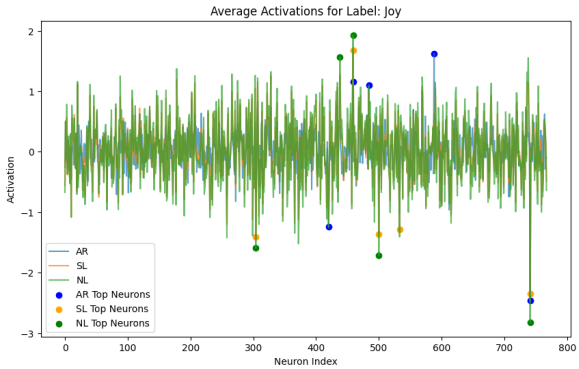

Figure 3: Average neuron activations for *Joy* (Layer-11) for Arabic, Slovene and Dutch. Top neurons are highlighted separately for each language.

Furthermore, the methodological framework developed here: combining activation extraction, variance ranking, and cross-lingual correlation, offers a scalable approach to probing multilingual representations without requiring retraining or gradient access. These findings contribute to the broader interpretability literature by showing that affective meaning leaves measurable neuron-level signatures in multilingual transformers. Future extensions could include causal interventions (e.g., targeted ablations or activation patching) to determine whether polyglot neurons are not only correlated with, but also *necessary* for, emotion prediction.

Limitations

While the present study offers novel insights into neuron-level interpretability for emotion detection in multilingual transformers, several limitations should be acknowledged. First, the analysis focuses on a limited set of languages. Extending the evaluation to a broader range of languages, particularly low-resourced and typologically distant ones, would provide a more comprehensive understanding of how multilingual models generalize across diverse linguistic systems. Second, our experiments are restricted to emotion classification. Evaluating other tasks, such as sentiment analysis or topic classification, could help to determine the extent to which the observed neuron activations are task-specific or reflect more general semantic mechanisms. Finally, the interpretability methods employed here, though effective for exploratory analysis, can be extended to more advanced techniques, such as neuron ablations or causal mediation analysis, to move from correlation towards determining causation. This would allow to better characterize neuron functions and enable stronger claims about the mechanisms underlying cross-lingual transfer.

References

Haya Al-Sheikh. 2014. A cognitive study of anger metaphors in arabic and english. *Journal of Language Studies*, 14(3):57–74.

Rochelle Choenni and Ekaterina Shutova. 2022. Investigating language relationships in multilingual sentence encoders through the lens of linguistic typology. *Computational Linguistics*, 48(3):635–672.

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.

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.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200.

Alex Foote, Neel Nanda, Esben Kran, Ioannis Konstas, Shay Cohen, and Fazl Barez. 2023. Neuron to graph: Interpreting language model neurons at scale. *Preprint*, arXiv:2305.19911.

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. 2024. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6919–6971, Mexico City, Mexico. Association for Computational Linguistics.

Zoltán Kövecses. 2000. *Metaphor and Emotion: Language, Culture, and Body in Human Feeling*. Cambridge University Press.

Miguel López-Otal, Jorge Gracia, Jordi Bernad, Carlos Bobed, Lucía Pitarch-Ballesteros, and Emma Anglés-Herrero. 2025. Linguistic interpretability of transformer-based language models: a systematic review. *Preprint*, arXiv:2504.08001.

Aaron Maladry, Pranaydeep Singh, and Els Lefever. 2024. Findings of the wassa 2024 exalt shared task on explainability for cross-lingual emotion in tweets. In *Proceedings of the 14th Workshop of on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis@ACL 2024*, Bangkok, Thailand.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Lucas Resck, Isabelle Augenstein, and Anna Korhonen. 2025. Explainability and interpretability of multilingual large language models: A survey.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. Ccnet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 4003–4012. European Language Resources Association (ELRA).

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.