Towards Extracting and Understanding the Implicit Rubrics of Transformer Based Automated Essay Scoring Models

James Fiacco
Language Technologies Institute
Carnegie Mellon University
jfiacco@cs.cmu.edu

David Adamson
Turnitin
dadamson@turnitin.com

Carolyn P. Rosé
Language Technologies Institute
Carnegie Mellon University
cprose@cs.cmu.edu

Abstract

By aligning the functional components derived from the activations of transformer models trained for AES with external knowledge such as human-understandable feature groups, the proposed method improves the interpretability of a Longformer Automated Essay Scoring (AES) system and provides tools for performing such analyses on further neural AES systems. The analysis focuses on models trained to score essays based on ORGANIZATION, MAIN IDEA, SUPPORT, and LANGUAGE. The findings provide insights into the models' decision-making processes, biases, and limitations, contributing to the development of more transparent and reliable AES systems.

1 Introduction

Since its inception over 50 years ago (Page, 1966), Automated Essay Scoring (AES) has been a valuable approach for evaluating large quantities of student essays. Recent developments in the field have sought to harness advanced natural language processing techniques to score essays on par with human raters, achieving significant progress toward that goal (Ramesh and Sanampudi, 2022; Huawei and Aryadoust, 2023; Mizumoto and Eguchi, 2023). The inability to understand the learned representations in deep learning based AES models introduces risk and validity concerns to their widespread use in educational settings (Ding et al., 2020; Kumar et al., 2020, 2023). In response to this concern, we propose a functional component-based approach to scrutinize the activations of transformer models trained for AES.

The primary goal of this study is to provide a method and tool that can provide a coherent and interpretable understanding of the functions performed by these neural models, comparing their overlaps and differences, and aligning the learned functions with human-understandable groups of features[1]. Much in the same way that human evaluators use rubrics to guide their scoring of essays, neural models learn a set of features and connections that, when combined and applied to an essay, repeatably determine the score that they will assign.Through the comparison and contrast of these components across models, we investigate how the models prioritize different aspects of writing and make stride towards unveiling that their learned rubrics are, alongside any underlying biases or limitations that they entail. Ultimately, this in-depth analysis will enhance our understanding of the neural models' decision-making processes, thereby contributing to the development of more transparent and reliable automated essay scoring systems.

Our proposed methodology involves extending the emerging domain of neural network interpretation by using abstract functional components, enabling a robust comparison between probed functional components of a network and independent feature groups. This approach specifically builds upon recent work on neural probes and derived methods, aligning a neural network's activations with external knowledge such as task metadata and implicit features (e.g., parts-of-speech, capitalization, etc.) (Conneau et al., 2018; Belinkov, 2022). We focus our interpretation in the domain of AES where each model in our investigation is trained to score essays based on distinct evaluation traits, namely ORGANIZATION, MAIN IDEA, SUPPORT, and LANGUAGE.

To probe these models, the features are drawn

[1]Code and tool available at https://github.com/jfiacco/aes_neural_functional_groups

Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), pages 232–241

from several sources that correspond to concepts of both high and low validity for essay scoring: statistical features of an essay (e.g. number of sentences, number of paragraphs, etc.) (Woods et al., 2017), tree features generated from Rhetorical Structure Theory (RST) (Mann and Thompson, 1987) parses of the essays (Jiang et al., 2019; Fiacco et al., 2022), essay prompt and genre (West-Smith et al., 2018), and a combination of algorithmically derived (Derczynski et al., 2015) and our own human defined style-based word lists. These features provide a lens that while unable to capture all of the capabilities of the models, provide insight into some of the key differences between them.

In the following sections, we provide a detailed description of the methodology used for this analysis, discuss the assumptions underpinning the method, and present potential explanations for correlated function/feature pairs through a series of experiments that validate our method's ability to reflect the internal rubric of each of the neural models.

2 Related Work

From the interpretability angle, the most closely related work to this is that of neural model probes (Shi et al., 2016; Adi et al., 2016; Conneau et al., 2018; Zhu et al., 2018; Kuncoro et al., 2018; Khandelwal et al., 2018) which have frequently being used to test whether a model has learned a set of properties (Ryskina and Knight, 2021; Belinkov, 2022). The primary gap we are working to fill in from this body of literature is that current approaches, with few exceptions (Fiacco et al., 2019; Cao et al., 2021), focus on understanding the roles of individual neurons in the greater neural network. We contend that studying the interpretability of a neural network at the individual neuron level can too easily obscure the broader picture. Our interest lies in further progress incorporating a more abstract perspective on what is learned by neural networks, complementing the work that has been done at the neuron level.

Compared to alternative paradigms for interpretability in machine learning models, such as LIME (Ribeiro et al., 2016) or SHAP (Lundberg and Lee, 2017), which evaluate the contribution of a given feature to the prediction of a model, the functional component based methods allow for a more granular identification of important parts of a model, independent from known features for a task. This can enable model analysts to quickly identify unexplained components and begin to propose alternative pallets of features. Furthermore, the functional components can represent intermediate steps within the neural network which would be unobservable with these alternative methods.

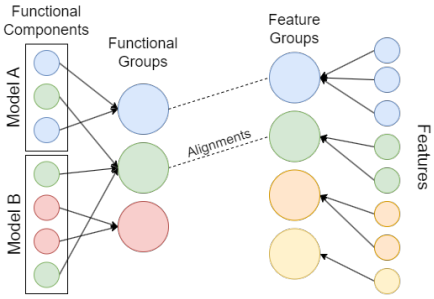

Figure 1: Diagram visualizing the structure of the methodology. Nodes of each color represent correlated values.

From the educational technologies and Automated Essay Scoring angle, our work primarily applies to the body of deep learning-based AES models such as recurrent neural network models (Jin et al., 2018; Nadeem et al., 2019), convolutional neural network models (Taghipour and Ng, 2016), and transformer models (Sethi and Singh, 2022). While our method could be applied to any type of neural model, we focus on transformers as they represent the state-of-the-art. By integrating the interpretability of neural models with the understanding of the functional components they learn, we hope to bridge the gap between human-understandable features and neural network-based essay scoring. The insights gained from our methodology can guide the development of more effective and efficient AES systems, tailored to the specific needs of educators and students. Furthermore, the lessons learned from this research may extend beyond the AES domain, providing valuable insights for the broader field of natural language processing and machine learning interpretability.

3 Methods

In this section we present our interpretation approach (Figure 1), defining the key concepts of *functional components*, *functional group*, *feature*, and *feature group*. Because the approach notably abstracts away from common terms in the neural network literature, throughout this section we draw an analogy to how one can define and describe the common features between mammals by comparing

their common and unique characteristics.

3.1 Functional Components and Groups

Functional components refer to the learned functions of a neural network, much like a particular component of a dog may be a "dog leg". In a neural AES system, these would be a group of neurons that have correlated activations when varying the input essays. The approach to extracting functional components ("neural pathways" as described by Fiacco et al. (2019)) from a neural network consists of finding the sets of coordinated neuron activations, summarized by the following steps:

1. Save the activations of neurons for each data instance in the validation dataset into an activation matrix, A of size $M \times N$, where M is the number of data instances in the validation set and N is the number of neurons being used for the analysis.

2. Perform a dimensionality reduction, such as Principal Component Analysis (PCA) (Hotelling, 1933), on A to get component activation matrix, T_{model} of size $M \times P$, where P is the number of principal components for a given model.

Functional groups are collections of similar functional components. Continuing the analogy, they would be compared to the more general concept of a "leg". We compute functional groups by concatenating the dimensionality reduced matrixes, T_{model}, of the two models that are to be compared and performing an additional dimensionality reduction over that matrix to get a matrix of group activations, T. The functional components that are highly loaded onto each functional groups are considered members of that group. An important departure from Fiacco et al. (2019), stemming from the limitation that does PCA does not guarantee independence between components, is that we use Independent Component Analysis (ICA) (Comon, 1994) instead. ICA is a dimensionality reduction technique that maximizes the independence between components, resulting in more validity in the technique's resulting alignments.

To determine if a functional group is influential in the performance of the model (designating it an *important functional group*), we can compute the Pearson's correlation coefficient between each column of the group activation matrix and the predictions of the model, the errors of the model, and the differences between the compared models.

3.2 Independent Feature Groups

Features are human understandable attributes that can be extracted from an analysis dataset. In the analogy they would represent potential descriptors of a components of a mammal, e.g. "hairy". In an AES context, these features may manifest as "no capitalization after a period". Ideally, it would be possible to create a direct mapping from each of the functional components to each of the features for which the functional component is related. However, this is non-trivial during a post-hoc analysis because, without interventions, there are limitations on what information is obtainable. Specifically, because features are not necessarily independent from each other, their correlations cannot be separated from each other, yielding imprecise interpretations. It is thus required for only independent features to be used as the unit of analysis when it comes to alignment with functional components. Unfortunately, in practice, this is a prohibitive restriction and most features that would be interesting are going to have correlations.

Fortunately, much in the same way that we can use ICA to extract independent functional components from a neural network's activations, we can use it to construct independent feature groups that can be reasonably be aligned with the functional groups of the neural networks. In the analogy, these independent feature groups can therefore, be thought of as collections of descriptive terms that can identify a characteristic of the mammal, such as "an appendage that comes in pairs and can be walked on" which would align with the "leg" functional group. In AES, an example feature group may be "uses punctuation improperly". It would be expected that this feature group would align well with a functional group in a neural AES system that corresponds with a negative essay score. Furthermore, feature groups for AES can be thought of as being roughly analogous to conditions that would be on an essay scoring rubric (as well as potentially other features that may be intuitive or obvious to human scorers but contribute to accurate scoring).

The specific process used to define these groups is to perform a dimensionality reduction on each set of feature types that may have significant correlations and collecting them into a feature matrix. We do this process for each feature type rather than

over all features at once because spurious correlations between some unrelated features may convolute the feature groups, making them far more difficult to interpret.

3.3 Alignment

Using ICA as the dimensionality reduction, the independent functional groups of the neural model can reasonably align with the independent feature groups using the following formal procedure: given a neural network, N, with activation matrix, A (as above), a independent component analysis is performed yielding a set of functional components, F. For each $f_i, f_k \in F$, $f_i \perp\!\!\!\perp f_k | X, Y$, where X is the set of inputs to the neural network and Y is the set of predictions from the neural network. With a sufficient number of components such that F contains all independent functional components in A, if there exists a common latent variable in both N and the set of independent feature groups, G, with components $g_i \in G$, then there will be some $f_i \overset{\propto}{\sim} g_j$.

4 Experiments

In this section, we delve into the specific methodology used to analyze the activations of the four transformer models for AES, as well as the steps taken to prepare the data and features for this analysis.

4.1 Datasets

Although scoring rubrics are specific to the genre and grade level of a writing task, there are commonalities between each rubric that allow their traits to be reasonably combined for modeling. All our rubrics, for example, include LANGUAGE (and style) and ORGANIZATION traits, though their expectations vary by genre and grade level. The generic MAIN IDEA trait corresponds to "Claim" and "Clarity and Focus" traits, and SUPPORT corresponds to "Support and Development" as well as "Analysis and Evidence." Rubrics and prompts were developed for validity, and essays were rigorously hand-scored by independent raters in the same manner as described in West-Smith et al. (2018).

For each generic trait, the training set was sampled down from over 50,000 available essays, responding to 95 writing prompts. Essays from 77 prompts were selected for the training set, and another 18 were held out for evaluation. Within each split, essays were sampled to minimize imbalance between essay score, genre, grade level, In the un-sampled data, longer essays tend to be strongly correlated with essay score, risking overfitting to this surface feature. Similarly, among the subset of data where school district data was available, districts with predominantly Black enrollment were under-represented among essays with a score of "4" across all traits. To counteract these potential biases, the available data was binned by length and district demographic information for each score, genre, and grade level, and essays were under-sampled from the largest bins. In addition to these balanced essays, about 800 "off topic" essays representing nonsense language or non-academic writing were included in the dataset, with a score of zero.

4.2 Models

Longformers are a transformer-based neural network architecture that have gained prominence in various NLP tasks (Beltagy et al., 2020). In the context of AES, each generic trait's model is a Longformer with a single-output regression head, fine-tuned on the trait's balanced dataset: For the remainder of this paper, the model fine-tuned on a given trait will be referred to as "the TRAIT model" (e.g. the ORGANIZATION model) for simplicity.

Although ordinal scores from 0 to 4 were used for sampling and evaluation, the training data labels were continuous, averaged from rater scores. Essays were prefixed with text representing their genre (e.g., "Historical Analysis") and prompt's grade range (e.g., "grades 10-12") before tokenization, but no other context for the writing task (e.g., the prompt's title, instructions, or source material) was included. In addition to Longformer's sliding attention window of 512 tokens, the first and last 32 tokens received global attention.

Scores were rounded back to integers between 0 and 4, before evaluation. On the holdout prompts, overall Quadratic Weighted Kappa (QWK) ranged from 0.784 for MAIN IDEA to 0.839 for LANGUAGE, while correlation with word count remained acceptably low: 0.441 for LANGUAGE up to 0.550 for SUPPORT.

The activations of the Longformer model were saved for each instance in the analysis set at the "classify" token to create a matrix of activations for the functional component extraction.

Model *A*	Model *B*	# Essays	Extracted Features	# Independent Feature Groups	# Aligned IFG
ORGANIZATION	MAIN IDEA	407	148	114	24
ORGANIZATION	LANGUAGE	275	118	86	39
ORGANIZATION	SUPPORT	144	90	63	37
LANGUAGE	MAIN IDEA	341	129	95	26
LANGUAGE	SUPPORT	72	67	38	23
SUPPORT	MAIN IDEA	260	127	94	27

Table 1: Comparing analysis dataset size and numbers of extracted features for each of the model comparisons, identified by the Model *A* and Model *B* columns.

4.3 Features

The features employed in this analysis encompass statistical properties of the essays, tree features generated from Rhetorical Structure Theory (RST) parse trees of the essays, essay prompt and genre, a combination of algorithmically derived and human-defined style-based word lists, and certain school-level demographic features. A description of each feature type is provided below:

Statistical Features: While statistical features such as *essay word count* are often good indicators of essay score, they are not intrinsically valuable to the different traits that our models are scoring. We thus want to see lower alignment with these features to indicate that the model is not overly relying on rudimentary shortcuts scoring an essay. We also include *average word length*, *essay paragraph count*, *essay sentence count*, *average sentence length*, and the *standard deviation of the sentence length* for completeness.

RST Tree Features: These features were integrated to capture the rhetorical structure of the text, such as the hierarchy of principal and subordinate clauses, the logical and temporal relations between propositions, and the coherence of the argument. These concepts have a high validity for scoring essays (Jiang et al., 2019), especially for ORGANIZATION, so high alignment between functional groups would be expected. To generate RST trees for each essay, we utilize a pretrained RST parser specifically fine-tuned for student writing (Fiacco et al., 2022). We include the presence of an RST relation as a feature as well as relation triplets $(\text{REL}_{parent}, \text{REL}_{child_1}, \text{REL}_{child_2})$ as tree-equivalent n-gram-like features.

Essay Prompt and Genre: Categorical representations of the essay prompt and genre were employed as features to examine if components of the AES model were preferentially activated based on the content or topic of the essay, a low validity feature.

Algorithmically Generated Word List Features: We calculate the frequency of usage of words within algorithmically derived sets of words in the essays as a group of features to probe the AES model's consideration for stylistic language. To generate these word lists, we obtain Brown clusters (Brown et al., 1992) from essays. We generate separate Brown clusters for each prompt in our dataset and subsequently derive final word lists based on the overlaps of those clusters. This approach emphasizes common stylistic features as opposed to content-based clusters.

Human Generated Word List Features: In addition to the algorithmically defined word lists, we devise our own word lists that may reflect how the AES model scores essays. We created word lists for the following categories: simple words, informal language, formal language, literary terms, transition words, and words unique to African American Vernacular English (AAVE).

Demographic Features: We used the percent to participants in the National School Lunch Program (NSLP) at a school as a weak proxy for the economic status of a student. Also as weak proxies for economic status of essay authors, we include the school level features of *number of students* and *student teacher ratio*. Furthermore, we use a school level distribution of ethnicity statistics as a weak proxy for the ethnic information of an essay's author. These features were employed to investigate the model's perception of any relationship between the writer's background and the quality, content, and style of the essay, in order to gain insight of the equity of the AES model.

4.4 Analysis Settings

To choose the number of components for ICA, a PCA was performed to determine how many components explained 95% of the variance of the activation (or 99% of the variance for the features) to be used as the number of components of the ICA.

		Functional Group Extraction			Important Functional Group Alignment			
Model *A*	**Model *B***	**# Comp. *A***	**# Comp. *B***	**# FG**	**# Aligned FG**	**# A Only**	**# B Only**	**# Mixed**
ORGANIZATION	MAIN IDEA	119	55	125	22	12	0	10
ORGANIZATION	LANGUAGE	96	66	110	29	11	0	18
ORGANIZATION	SUPPORT	66	36	68	22	9	1	12
LANGUAGE	MAIN IDEA	78	55	93	23	8	3	12
LANGUAGE	SUPPORT	34	28	38	13	2	2	9
SUPPORT	MAIN IDEA	45	49	64	25	2	2	21

Table 2: Comparing number of functional groups extracted for each model comparison and presenting the number of functional groups that were both deemed important (Section 3.1) and sufficiently aligned with at least one feature group. Also specified is the number of functional groups that are unique to a particular model and the number that are shared between the models of given a comparison pair.

To determine that a functional group was important, it needed to have an absolute value of Pearson's r value of greater than 0.2. This threshold was also used to determine if a functional group should be considered aligned with a feature group.

5 Results

In this section, we present aggregate statistics for each model comparison when it comes to computing features and independent feature groups (Table 1), extracting functional groups and aligning important functional groups (Table 2), and lastly, we provide examples taken from the model comparison between the LANGUAGE model and the MAIN IDEA model. Due to length constraints, we present detailed examples of this comparison only. Similar figures and correlation statistics can can be found on Github[2].

5.1 Independent Feature Groups

Since each trained model held out a different set of prompts from its training set, common prompts between analysis sets needed to be identified, and thus the number of features extracted and the resulting independent feature groups vary between model comparisons. Computing the independent feature groups for each model comparison (Table 1) yielded between 70% and 77% of the original extracted features for all comparisons, except LANGUAGE V SUPPORT, which only yielded 57% as many independent feature groups compared to original features. Despite high variability in the number of independent feature groups identified during the process, a much more narrow range of independent feature groups was aligned during the analysis.

[2] https://github.com/jfiacco/aes_neural_functional_groups/tree/main/supplementary_results

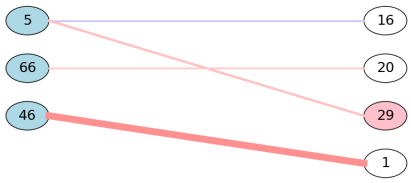

Figure 2: Alignment diagram for functional groups (*left*) that are specific to the MAIN IDEA model with their alignment to feature groups (*right*). Only functional groups and feature groups are shown if they have a positive correlation greater than 0.25 (*blue edges*) or a negative correlation less than -0.25 (*red edges*). The numbers correspond to the IDs of the functional group or feature group that the node represents (*see* Table 3).

The types of feature groups that were aligned varied considerably between different comparisons.

5.2 Functional Component Groups

The initial extraction of functional components for each model elicited numbers of functional components between 28 and 119. Table 1 and 2 show that for a given model, fewer functional components will be extracted given a fewer instances in the analysis dataset. Despite this noise, a clear pattern emerges where the ORGANIZATION model has the most functional components, followed by the LANGUAGE model. The MAIN IDEA model has fewer functional components, with the SUPPORT model having the fewest.

When performing the dimensionality reduction to compute the functional groups, there is a consistent reduction to approximately 61-71% of the combined total functional components.

5.3 Important Functional Groups

Despite the variance in the number of feature groups and functional groups extracted per comparison, there is a remarkably consistent number of

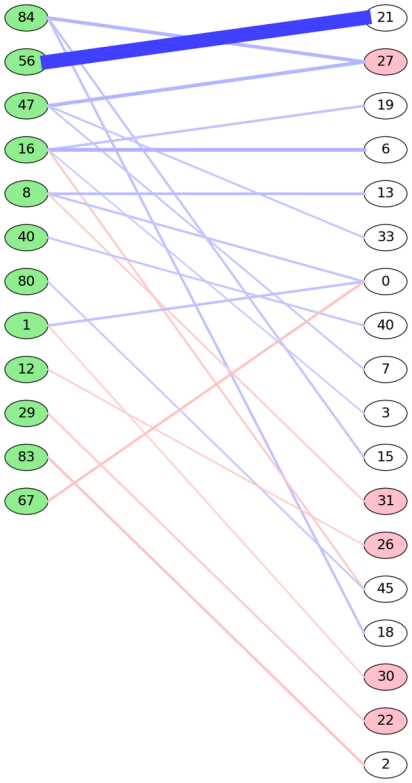

Figure 3: Alignment diagram for functional groups (*left*) that are common to both the LANGUAGE and MAIN IDEA models with their alignment to feature groups (*right*). Only functional groups and feature groups are shown if they have a positive correlation greater than 0.25 (*blue edges*) or a negative correlation less than -0.25 (*red edges*). The numbers correspond to the IDs of the functional group or feature group that the node represents (*see* Table 3).

Functional Group 46
Diff:LANGUAGEVSMAINIDEA $r = -0.39 (p < 0.001)$
Independent Feature Group 1 $r = -0.43 (p < 0.001)$
ModelErrors:MAINIDEA(+), ModelPairDifference(+), ModelErrors:LANGUAGE(-)
Functional Group 56
Predictions:MAINIDEA $r = -0.13 (p < 0.05)$
Independent Feature Group 21 $r = 0.75 (p < 0.001)$
EssayStats:STDDEVSENTENCELENGTH(+), EssayStats:NUMSENTENCES(+), EssayStats: MEANWORDLENGTH(+), EssayStats:NUMWORDS (-), EssayStats:NUMPARAGRAPHS(-), EssayStats: MEANSENTENCELENGTH(-)
Functional Group 92
Predictions:LANGUAGE $r = -0.13 (p < 0.05)$
Independent Feature Group 12 $r = -0.20 (p < 0.001)$
WordCluster:PRIORITIES(+), WordCluster:POPULATIONCOMPARISION(+), WordCluster:EFFICIENCY(+), WordCluster:TEENVALUES(-), WordCluster:STORYTELLING(-), WordCluster:SCHOOL (-), WordCluster:PARENTALDECISIONS(-), WordCluster:INFORMAL(-), WordCluster:HISTORICALCONFLICT(-)
Independent Feature Group 69 $r = 0.22 (p < 0.001)$
RST:NN|CONTRAST(+), RST:SN|EVALUATION(NS|ELABORATION, LEAF)(+), RST:SN|BACKGROUND(LEAF, NS|ELABORATION)(+), RST:NS|EVIDENCE(LEAF, NN|CONJUNCTION)(+), RST:NN|JOINT(NN|CONJUNCTION, NN|JOINT)(+), RST:NN|CONTRAST(LEAF, LEAF)(+), RST:NN|CONJUNCTION(NS|ELABORATION, NN|CONJUNCTION)(+), RST:SN|EVALUATION(NN|CONJUNCTION, LEAF)(-), RST:NN|CONJUNCTION(LEAF, LEAF)(-)

Table 3: Selected examples of correlated functional group/feature groups. Pearson's R values for relevant importance metric (model difference, model predictions) and feature group alignment are presented with p-values.

important functional groups that have at least one sufficient alignment to a feature group (Table 2). With the exception of the LANGUAGE V SUPPORT comparison, all other comparisons had between 21 and 29 aligned functional groups.

As a visual aid for the important functional groups, see the left sides of Figures 2 and 3. Each Figure is derived from the functional groups and feature groups of the LANGUAGE V MAIN IDEA comparison. The numbers on each node are the identifiers of a given functional group, a subset of which are represented in Table 3.

5.4 Alignment of Functional Groups

The entirety of findings from the alignments for all of the comparisons would be too numerous to present in a conference paper format. However, we will present the major trends we found in our analysis. The first main trend is that all models had functional groups that we correlated with the statistical features of the essay. Furthermore, by computing the correlations between the individual features within that type, it was determined that *number of paragraphs* is likely the most salient contributor.

The second set of trends is presented in Table 4, where the percent of the total aligned feature groups per model was computed. This revealed that the ORGANIZATION model had considerably more aligned RST-based features than the other models, while the MAIN IDEA model had the least proportion. The LANGUAGE model had the most aligned word list features, which is the combination of the algorithmically and human-created word list features. For the last percentage, we combine the prompt and demographic features and find that the SUP-

Model	%RST	%Word List	%Demo. & Prompt
ORGANIZATION	**41**	13	21
LANGUAGE	30	**26**	19
SUPPORT	36	19	13
MAIN IDEA	23	21	**23**

Table 4: % of aligned feature groups for a given model by feature type.

PORT model tended to align with fewer of these types of features. The reason for combining the demographic and prompt features is discussed in Section 6.

5.5 Qualitative Analysis

While the method that we presented can quickly advance one's understanding of a model from the black-box neural network to aligned feature groups directly, understanding what function a feature group represents can be more difficult. It is thus necessary to resolve what a feature group represents to form a strong statement on what the model is doing. For instance, we found it concerning that so many of the models were connected with feature groups that contained demographic features (colored red in Figures 2 and 3). However, a qualitative look at the datasets for which prompts were included, we found that the distribution of prompts over the different schools, when controlling for essay length, were such that certain schools (with their demographic features) were the only source of certain prompts. It, therefore, becomes likely that many of these feature groups are more topic-based rather than the potentially more problematic demographic-based. This interpretation was reinforced by many of the feature groups with demographic information also including prompts (e.g. *"Independent Feature Group 29"* from Table 3) and by examining essays that present those feature groups.

6 Discussion

The results presented in the preceding section demonstrate the efficacy of the proposed method in extracting salient feature groups and functional groups from the neural models, particularly when applied to the dataset under consideration. The true potential of this method, however, lies in its capacity to be broadly applied to any neural AES system, thereby facilitating a deeper understanding of the models and the underlying processes they employ. In the following discussion, we will delve further into the results, emphasizing the prominent trends observed in the alignment of functional groups and their correlation with essay features, as well as the implications of these findings for enhancing the interpretability and transparency of neural AES systems.

6.1 Functional Component and Feature Groups

The proposed method successfully extracted meaningful functional groups from the analyzed neural models. Notably, the LANGUAGE V SUPPORT comparison emerged as an outlier in several of our analyses. This discrepancy is likely attributable to the considerably fewer essays shared by both models' analysis sets, which may result in a noisier analysis and expose a limitation of the method. As the size of the analysis increases, one would expect the extraction of feature groups and function groups to approach their ideal independence characteristics. Despite this limitation, the method managed to condense the analysis space from thousands of activations to fewer than 125 while still accounting for over 90% of the model's variance.

Interestingly, the ORGANIZATION model exhibited the highest number of functional groups. This observation suggests that capturing the ORGANIZATION trait is a more intricate process, necessitating the learning of additional features. This notion is further corroborated by the comparisons between ORGANIZATION and other models; models which displayed very few, if any, functional groups exclusively present in the non-organization models.

6.2 Alignment of Important Functional Groups

In line with our expectations, the ORGANIZATION model demonstrated the greatest alignment with the RST tree features, while the LANGUAGE model displayed the most significant alignment with the word list features. It was postulated that ORGANIZATION would necessitate the model to possess knowledge of how ideas within essays are structured in relation to each other, a type of knowledge encoded by rhetorical structure theory. Although the RST parse trees recovered from the parser are considerably noisy (RST parsing of student essay data has been shown to be markedly more challenging than standard datasets (Fiacco et al., 2022)), the signal remained significant. Furthermore, we anticipated that the LANGUAGE model would have a

greater reliance on word choice, a concept mirrored by the word list-based feature groups.

Contrary to our expectations, the MAIN IDEA model exhibited the highest number of prompt-based feature groups. Our most plausible explanation for this observation is that certain prompts might have clearer expectations for thesis statements than others, a notion generally supported by a qualitative examination of the essays from prompts that score higher on MAIN IDEA.

7 Conclusion

The neural network interpretation technique presented in this paper demonstrates significant promise in learning the implicit rubrics of neural automated essay scoring models. By effectively mapping the intricate relationships between feature groups and the functional groups of the underlying scoring mechanism, the technique provides a step towards an understanding of the factors contributing to a transformer's evaluation of essay quality. This enhanced understanding enables researchers and educators to not only identify potential biases in scoring models, but also to refine their models to ensure a more reliable and fair assessment of student performance.

The code for this method will be released and incorporated into an analysis tool for application to neural models not limited to the ones examined in this work with the goal to pave the way for the development of more transparency in neural AES models. These advancements can contribute to the overarching goal of promoting ethical and responsible AI in education by facilitating the examination and comprehension of complex neural models.

Acknowledgements

This work was supported in part by NSF grant DRL 1949110.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *arXiv preprint arXiv:1608.04207*.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Peter F Brown, Vincent J Della Pietra, Peter V Desouza, Jennifer C Lai, and Robert L Mercer. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–480.

Steven Cao, Victor Sanh, and Alexander M Rush. 2021. Low-complexity probing via finding subnetworks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 960–966.

Pierre Comon. 1994. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $ &!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136.

Leon Derczynski, Sean Chester, and Kenneth S Bøgh. 2015. Tune your brown clustering, please. In *International Conference Recent Advances in Natural Language Processing, RANLP*, volume 2015, pages 110–117. Association for Computational Linguistics.

Yuning Ding, Brian Riordan, Andrea Horbach, Aoife Cahill, and Torsten Zesch. 2020. Don't take "nswvtnvakgxpm" for an answer–the surprising vulnerability of automatic content scoring systems to adversarial input. In *Proceedings of the 28th international conference on computational linguistics*, pages 882–892.

James Fiacco, Samridhi Choudhary, and Carolyn Rose. 2019. Deep neural model inspection and comparison via functional neuron pathways. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 5754–5764.

James Fiacco, Shiyan Jiang, David Adamson, and Carolyn Rose. 2022. Toward automatic discourse parsing of student writing motivated by neural interpretation. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 204–215.

Harold Hotelling. 1933. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417.

Shi Huawei and Vahid Aryadoust. 2023. A systematic review of automated writing evaluation systems. *Education and Information Technologies*, 28(1):771–795.

Shiyan Jiang, Kexin Yang, Chandrakumari Suvarna, Pooja Casula, Mingtong Zhang, and Carolyn Rose. 2019. Applying rhetorical structure theory to student essays for providing automated writing feedback. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 163–168.

Cancan Jin, Ben He, Kai Hui, and Le Sun. 2018. Tdnn: a two-stage deep neural network for prompt-independent automated essay scoring. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1088–1097.

Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. 2018. Sharp nearby, fuzzy far away: How neural language models use context. *arXiv preprint arXiv:1805.04623*.

Yaman Kumar, Mehar Bhatia, Anubha Kabra, Jessy Junyi Li, Di Jin, and Rajiv Ratn Shah. 2020. Calling out bluff: attacking the robustness of automatic scoring systems with simple adversarial testing. *arXiv preprint arXiv:2007.06796*.

Yaman Kumar, Swapnil Parekh, Somesh Singh, Junyi Jessy Li, Rajiv Ratn Shah, and Changyou Chen. 2023. Automatic essay scoring systems are both overstable and oversensitive: Explaining why and proposing defenses. *Dialogue & Discourse*, 14(1):1–33.

Adhiguna Kuncoro, Chris Dyer, John Hale, Dani Yogatama, Stephen Clark, and Phil Blunsom. 2018. Lstms can learn syntax-sensitive dependencies well, but modeling structure makes them better. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1426–1436.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

William C Mann and Sandra A Thompson. 1987. *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute Los Angeles.

Atsushi Mizumoto and Masaki Eguchi. 2023. Exploring the potential of using an ai language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050.

Farah Nadeem, Huy Nguyen, Yang Liu, and Mari Ostendorf. 2019. Automated essay scoring with discourse-aware neural models. In *Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications*, pages 484–493.

Ellis B Page. 1966. The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243.

Dadi Ramesh and Suresh Kumar Sanampudi. 2022. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55(3):2495–2527.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Maria Ryskina and Kevin Knight. 2021. Learning mathematical properties of integers. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 389–395.

Angad Sethi and Kavinder Singh. 2022. Natural language processing based automated essay scoring with parameter-efficient transformer approach. In *2022 6th International Conference on Computing Methodologies and Communication (ICCMC)*, pages 749–756. IEEE.

Xing Shi, Kevin Knight, and Deniz Yuret. 2016. Why neural translations are the right length. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2278–2282.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1882–1891.

Patti West-Smith, Stephanie Butler, and Elijah Mayfield. 2018. Trustworthy automated essay scoring without explicit construct validity. In *AAAI Spring Symposia*.

Bronwyn Woods, David Adamson, Shayne Miel, and Elijah Mayfield. 2017. Formative essay feedback using predictive scoring models. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2071–2080.

Xunjie Zhu, Tingfeng Li, and Gerard Melo. 2018. Exploring semantic properties of sentence embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 632–637.