Appendix

A Qualitative analysis on Attention

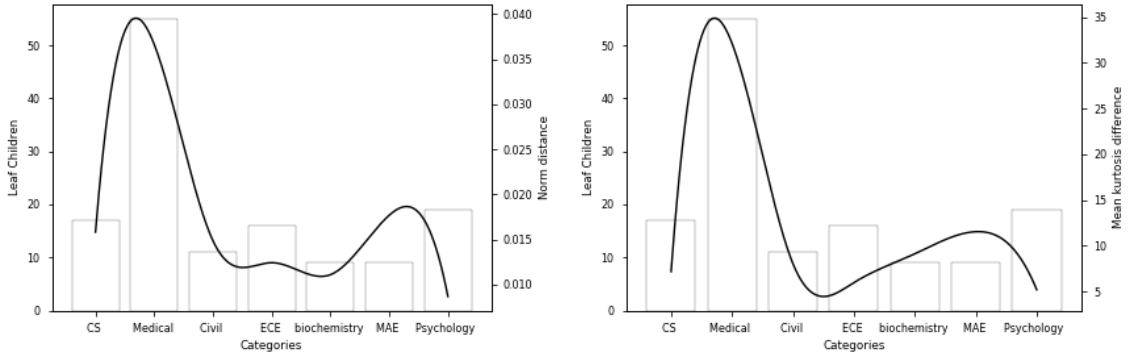

Figure 1: Difference in attentions among levels ($l_2 - l_1$) using the Euclidean distance & the Kurtosis measure. The bar chart represents the number of children within that parent category.

To analyze the focus of attention in a qualitative measure, we computed the difference in mean attentions from parent levels to child levels using the Euclidean distance (L_2 norm). We observed that the trees with more number of children in the taxonomy have higher L_2 norm differences in attention spread (Figure 1). To measure the amount of attention spread decrease from the current level to the next level, we employed the statistical metrics - Kurtosis (Mardia, 1970), which measures the tailedness of a distribution. We observed an increase in the tailedness of the attentions in level 2 (l_2) with respect to that of level 1 (l_1), quantifying the narrowing of focus throughout the dataset as we claimed in the discussion section of the main text.

B Ablation Study

In addition to the main experiments, we also perform the ablation study on our model to see the effect of various components in our proposed architecture. The results are given in Table 1. The Web of Science dataset (WOS) is used for our ablation study because it is proven to be more difficult to classify (Kowsari et al., 2017).

We tested our model with various modifications. Firstly, we experimented with a one-hot parent encoding which represents the parent class in a vector of size k, where k is the total number of parent classes. Secondly, We checked the effects of increasing and decreasing the number of attention hops. Lastly, we experimented with alternative pooling mechanisms such as max, mean or concat pooling (Collobert and Weston, 2008).

We observed only a marginal gain by using attention with respect to concat pooling on WOS. This can be attributed to the inherent mechanism of pooling. Attention mechanisms focus on certain words by either increasing or decreasing the vector representation of the words, while a pooling mechanism like *concat pooling* achieves comparable performance by identifying the discriminative *dimension* of the word representation across all words. Thus, a pooling mechanism has an advantage of using the raw representations over all words to identify discriminative signals. Although, we do acknowledge that the relatively small size of WOS can be a deciding factor of why attention mechanisms do not perform significantly better than pooling.

References

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM.

Architecture		WOS		
		l_1	l_2	Overall
Our Model		89.32	82.42	77.46
Attention	Without previous layer encoding	88.82	79.21	75.97
	Without BiLSTM encoder - pure attention	86.56	79.60	72.09
	With single final classifier	86.69	76.78	71.83
	With one-hot parent encoding	88.57	82.66	76.83
	With low attention hops - 2	89.15	78.80	74.99
	With high attention hops - 15	88.71	78.62	74.65
Pooling	Without attention - max pooling	88.37	77.39	77.39
	Without attention - mean pooling	87.69	73.59	73.59
	Without attention - concat pooling	88.63	80.92	77.28
	Without BiLSTM encoder - pure concat pooling	85.59	73.01	73.01

Table 1: The ablation study with various architecture changes on WOS

Kamran Kowsari, Donald E Brown, Mojtaba Heidarysafa, Kiana Jafari Meimandi, Matthew S Gerber, and Laura E Barnes. 2017. HDLTex: Hierarchical deep learning for text classification. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 364–371.

Kanti V Mardia. 1970. Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3):519–530.