B. Attention Flow

Despite *attention flow*, Grad-SHAP, and Deep-SHAP all (supposedly) being valid Shapley Value explanations [11], Table 4 shows that the agreement is low.

Table 4. Mean Kendall-τ between the explanations given by *attention flow* and our chosen XAI methods for the DistilBERT model when applied to 500 instances of the test portion of each dataset. IMDb is not included among these datasets, because the long sequences made the *attention flow* computation unfeasible.

		LIME	Int-Grad	DeepLIFT	Grad-SHAP	Deep-SHAP
Attn Flow	MNLI	.1326	.1251	.2159	.1227	.2148
	Quora	.0853	.2426	.0367	.0241	.2319
	SNLI	.0844	.0753	.2178	.0571	.2149
	SST-2	.1795	.0689	.1286	.0811	.1202