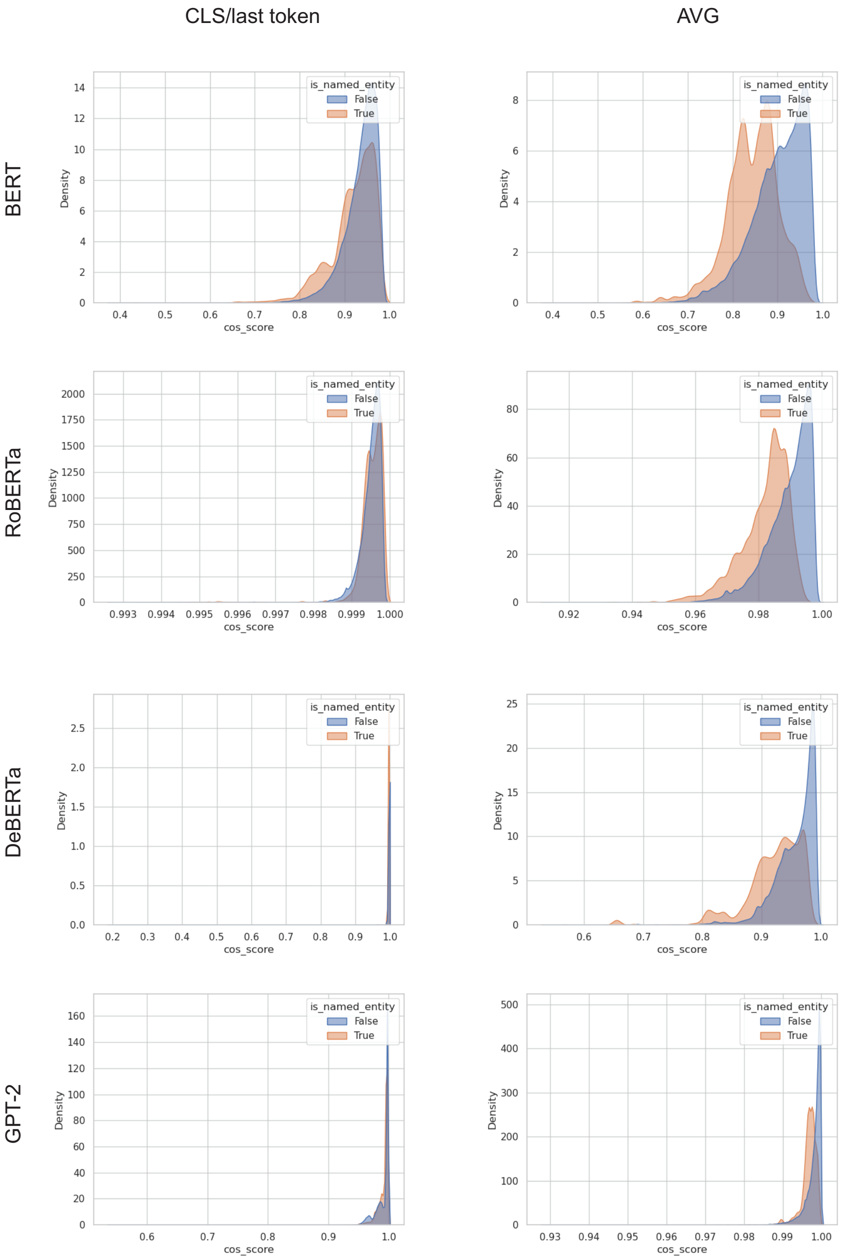

Figure 12: Distributions of compositionality score for named entities and non-named entities across model types and representation types. The AVG representation matches the intuition that named entities are usually less semantically compositional, as they point to an entity in the real world that may not relate to their name.

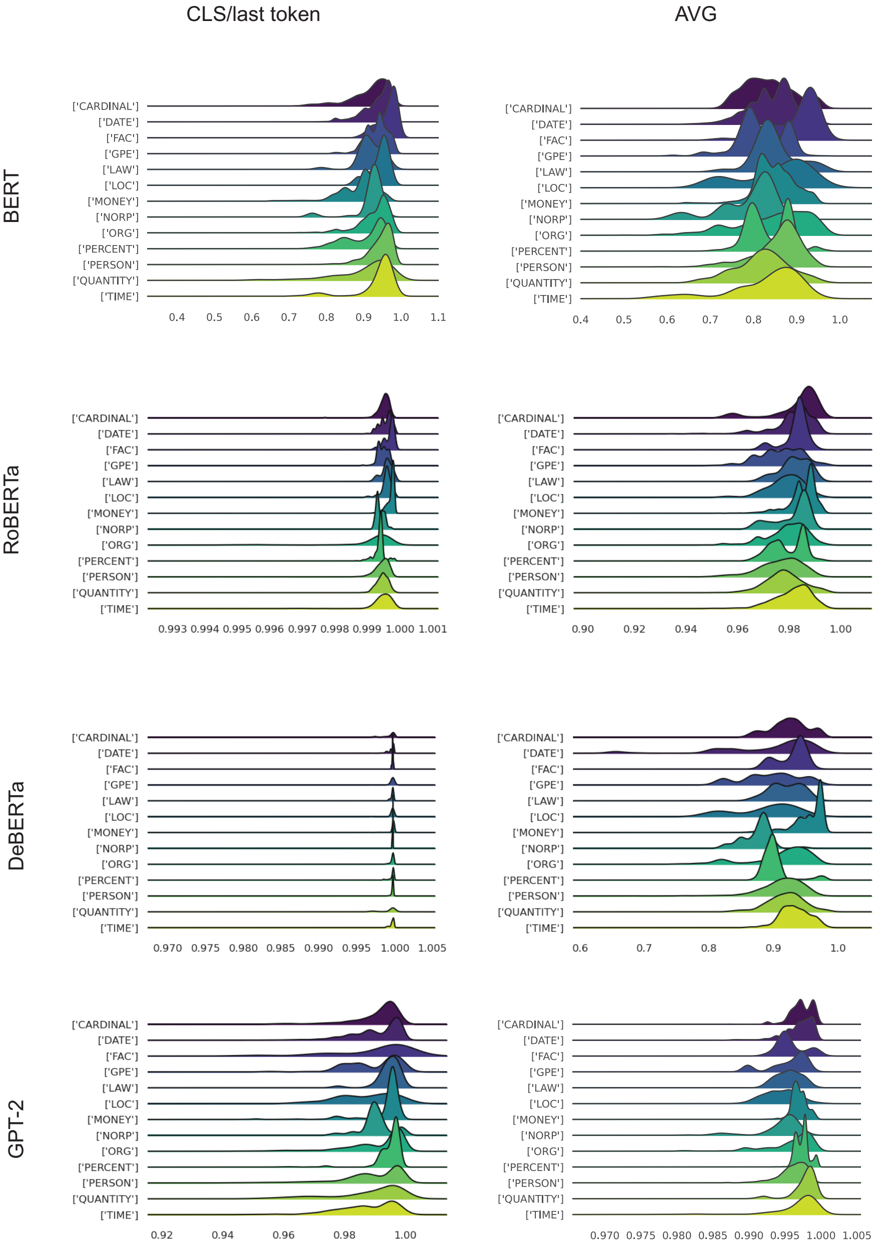

Figure 13: Visualization of distribution of compositionality scores across different types of named entities.

K Frequency and length correlations

Model and representation	Feature	Spearman ρ	p-val
BERT$_{\text{CLS}}$	Word length	0.2182	$3.055 \times 10^{-10}*$
BERT$_{\text{AVG}}$		0.007396	0.08722
RoBERTa$_{\text{CLS}}$		0.01686	0.6193
RoBERTa$_{\text{AVG}}$		0.3653	$4.773 \times 10^{-28}*$
DeBERTa$_{\text{CLS}}$		0.4087	$1.709 \times 10^{-35}*$
DeBERTa$_{\text{AVG}}$		0.4484	$1.340 \times 10^{-42}*$
GPT-2$_{\text{last}}$		0.3228	$8.481 \times 10^{-22}*$
GPT-2$_{\text{AVG}}$		0.0.3125	$1.719 \times 10^{-20}*$
Human	Word length	0.05666	0.1894
BERT$_{\text{CLS}}$	Frequency	0.2182	0.08193
BERT$_{\text{AVG}}$		-0.08582	0.07899
RoBERTa$_{\text{CLS}}$		0.02548	0.9053
RoBERTa$_{\text{AVG}}$		-0.08354	0.08193
DeBERTa$_{\text{CLS}}$		-0.1265	0.001459*
DeBERTa$_{\text{AVG}}$		-0.2185	$6.455 \times 10^{-10}*$
GPT-2$_{\text{last}}$		-0.05750	0.3595
GPT-2$_{\text{AVG}}$		0.04382	0.5891
Human	Frequency	0.008363	0.9053

Table 12: Correlations of frequency and length with human and model compositionality scores. Corrected with Holm-Bonferroni correction.