Процедура оценивания NLP задачи

Для каждой пары из оригинального комментария O и его менее токсичного варианта D вычисляется скор:

$$score(D,\ O) = \begin{cases} 0 & ,\ if\ Model(D) \geqslant 0.5 \\ 1\ -\ \frac{embeddingDistance(D,\ O)}{countWords(O)} & ,\ if\ Model(D)\ <\ 0.5 \end{cases}$$

Где $Model(D)$ - это предсказание токсичности комментария с точки зрения модели.

$$embeddingDistance(D,\ O)\ =$$

$$\begin{cases} countWords(O) - OOV(O) & ,\ if\ countWords(D)\ =\ 0 \\ countWirds(D) & ,\ else\ if\ countWords(O)\ =\ 0 \\ max(OOV(D) - OOV(O),\ 0)\ + & ,\ else\ if\ OOV(D) + OOV(O)\ >\ 0 \\ \qquad embeddingDistance(dropOOV(D),\ dropOOV(O)) & \\ min_{i,j} cos(embedding(D_i),\ embedding(O_j)) + & ,\ otherwise \\ embeddingDistance(dropClosest(D, O),\ dropClosest(O, D)) & \end{cases}$$

$OOV(D)$ - Количество слов в предложении D, для которых нет предпосчитвнных эмбеддингов, $dropOOV(D)$ выкидывает такие слова из предложения.

$countWords(O)$ - считает количество слов в предложении O

$dropClosest(A, B)$ возвращает предложение A, из которого выброшено слово с ближайшим по косинусному расстоянию эмбеддингом к любым словам в B

Перед вычислением $score(D, O)$ все предложения приводятся к нижнему регистру и нормализуются: из них удаляются все символы кроме пробелов и букв, а также во всех словах остается не больше первых 5-ти BPE-юнитов.