Кузиной Екатерины

Обзор на статью
BERT-β: A Proactive Probabilistic Approach to Text Moderation

https://arxiv.org/pdf/2109.08805.pdf

Борьба с токсичностью текстов и комментариев очень необходима для поддержания нетоксичности медиа-платформ. Большинство работ посвящено именно тому, как определить токсичность текста, однако в данной статье авторы подходят к определению токсичности текста с другой стороны: они вводят концепцию склонности к токсичности текста с опережением, т.е. насколько вероятно, что статья будет подвержена токсичным комментариям.

Предлагается использовать вероятностный подход для определения склонности текста к токсичности. Для ранее опубликованных новостных статей с комментариями берется среднее значение показателей токсичности комментариев в качестве критерия достоверности для обучения модели.

Для оценивания токсичности комментариев и склонности к токсичности статей используется Бета-распределение, т.к. эмпирически данные оценки [токсичность комментариев и склонность к токсичности] демонстрируют асимметрию и могут плохо моделироваться распределением Гаусса (рис. 2 и 3). Кроме того, распределение оценок токсичности комментариев для отдельных статей зависит от содержания статьи, как показано на рис. 3.

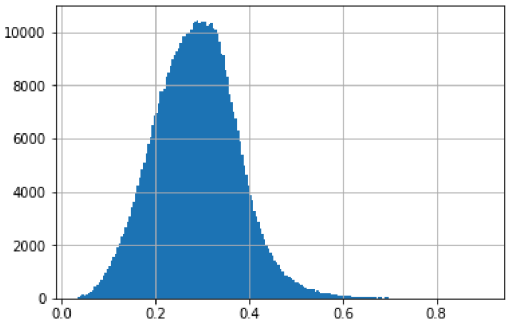

Figure 2: Toxicity propensity score (mean comment toxicity scores) distribution of news articles.

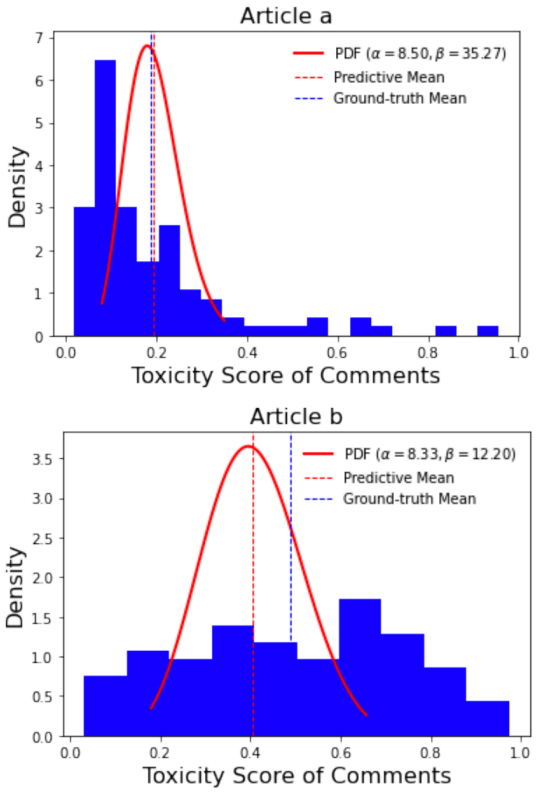

Figure 3: Toxicity score histogram density of comments for articles a (top) and b (bottom). Solid red lines represent predictive beta distribution for individual articles.

Бета-распределение очень гибкое и может моделировать довольно широкий спектр хорошо известных семейств распределений: от симметричных равномерных (α = β = 1) и колоколообразных распределений (α = β = 2) до асимметричных форм (α != β).

Предполагается, что показатель склонности к токсичности соответствует бета-распределению с функцией плотности вероятности:

$$p(y|\alpha,\beta) = \text{Beta}(\alpha,\beta) = \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha,\beta)}$$

Для каждого обучающего примера из выборки $D = \{(x_n, y_n)\}_{i=1}^{N}$ делается feature engineering или находится эмбеддинг $g(\bullet)$ для x_n, а затем с помощью построения регрессии находятся α_n и β_n как

$$\log(\alpha_n) = f_\alpha(g(\boldsymbol{x}_n))$$
$$\log(\beta_n) = f_\beta(g(\boldsymbol{x}_n))$$

где $f_\alpha(\bullet)$ и $f_\beta(\bullet)$ обучаются одновременно. Функция $g(\bullet)$ может быть как зафиксированной функцией до обучения, так и находится вместе с $f_\alpha(\bullet)$ и $f_\beta(\bullet)$. Во время обучения модели минимизируется функция потерь

$$\mathcal{L} = -\frac{1}{N}\sum_{n=1}^{N}\log\big(p(y_n|\alpha_n,\beta_n)\big)$$

В качестве точечной оценки токсичности берется $\hat{y}_n = \dfrac{\alpha_n}{\alpha_n+\beta_n}$, потому что предсказывается средняя токсичность текста.

Используемый в экспериментах датасет состоит из статей с комментариями на английском языке, опубликованных в Yahoo. У 99% проанализированных статей количество комментарий находится от 10 до 8000 с 25% квантилем 20, медианой 50 и средним значением 448. Входной текст состоит из заголовка статьи и основного текста, а оценка склонности к токсичности текста определяется как средняя оценка токсичности всех связанных комментариев. Токсичность лежит в диапазоне от 0 до 1 и оценивается с помощью Google's Perspective (это сверхточная нейронная сеть, обученная на википедии).

Авторы обучают модель, в которой $f_\alpha(\bullet)$ и $f_\beta(\bullet)$ - однослойные нейронные сети, $g(\bullet)$ - bag of words или эмбецдинги BERT'a (в зависимости от используемой функции модели называются BOW-β и BERT-β соответственно). В частности, берется последовательность одногаммных и двуграммных слов и

вычисляется соответствующие векторы частоты терминов во всех документах (tf-idf), что приводит к примерно 5,8 миллионам токенов для bag of words. Также, если вводимый текст превышает максимальную длину, авторы эмпирически выбирают первые 120 и последние 382 токена, т.к. информативные фрагменты с большей вероятностью будут располагаться в начале и конце.

Авторы хотят, чтобы тексты с более высокой средней токсичностью имели более высокий рейтинг, чем тексты с низкой склонностью, поэтому для сравнения моделей берется не только MAE, RMSE, AUC@Precision-Recall кривые, но и ранжирующие метрики, а именно коэффициент Кендалл и коэффициент Спирмена.

Table 2: Performance comparisons on test set

	Kendall		Spearman		MAE		RMSE	
	val	test	val	test	val	test	val	test
BOW-MAE	0.332	0.314	0.488	0.464	0.076	0.081	0.095	0.100
BOW-MSE	0.428	0.402	0.606	0.574	0.057	0.063	0.076	0.084
BOW-β	0.437	0.413	0.617	0.589	**0.056**	**0.061**	**0.075**	**0.081**
BERT-MAE	0.360	0.333	0.525	0.489	0.072	0.076	0.092	0.095
BERT-MSE	0.442	0.423	0.621	0.598	0.070	0.073	0.089	0.093
BERT-β	**0.462**	**0.440**	**0.642**	**0.617**	**0.056**	0.065	**0.075**	0.085

В таблице 2 видно, что модели с Бета-распределением дают более хорошее качество. Для того, чтобы убедиться в верности токсичности статей (т.к. это довольно субъективное суждение), авторы попросили людей определить склонность к токсичности по шкале Very Unlikely (VU), Unlikely (U), Neutral (N), Likely (L) and Very Likely (VL). Оказалось, что иногда люди даже не согласны друг с другом в степени токсичности.

Предложенные в статье модели могут быть полезны в модерации текста. Например, можно применять более строгие правила при модерации, если статья имеет высокую склонность к токсичности. Редакторы статей могут применять модель для нахождения фраз, которые стоит переформулировать для снижения токсичных комментариев. Также модель может быть использована в качестве дополнительной функции для последующих моделей обработки текста.

Текущая модель, по мнению авторов, не является идеальной и требует доработки, однако проведенные эксперименты показывают эффективность предложенной модели по сравнению с рядом базовых показателей в прогнозировании как среднего показателя токсичности, так и человеческого суждения о токсичности.